Modelling the steps of early syntax acquisition

Jacqueline van Kampen
UiL OTS
Utrecht University
Utrecht, 3512 BL
jacqueline.vankampen@let.uu.nl

Remko Scha
ILLC
University of Amsterdam
Amsterdam, 1018 TV
scha@uva.nl

Abstract
This paper reports on empirical research concerning the early stages of syntactic development in first language acquisition, and investigates the prospects of modelling the acquisition process in an exemplar-based linguistic framework (Data-Oriented Parsing). We present a detailed analysis of the early stages of verbal placements in Dutch. Our analysis shows that the child invents a sequence of grammars. She starts with rigorous, but systematic simplifications. The child approaches the adult grammar in a sequence of steps. We propose that each acquisition step can be 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 discuss in some detail the how a child acquires the complexities of Dutch word order. It turns out to be extremely implausible that this process could be simulated by a model whose syntactic operations are limited to tree-substitution.

1 Introduction
This paper consists of two parts. In the first part (section 2), we discuss some complexities of Dutch word order, and some empirical findings about the stepwise process by which a Dutch child ends up introducing these complexities in the language that it is developing.

In the second part of the paper (section 3) we go into more detail about the problem of modelling this acquisition process. In particular, we investigate to what extent exemplar-based linguistic approaches such as Data-Oriented Parsing (DOP) provide suitable frameworks for this purpose.

2 Acquiring Dutch word order
Dutch clauses come in different types, each with its own word order. A Dutch three year old typically masters all the word order variants in (2)-(8). The ‘root infinitive’ type in (1), an invariant moment of the variants in (2)-(8), is as such not present in the adult input. Nevertheless, it is present in the child’s output and precedes the types in (2)-(8).

(1) ‘root infinitive’
   a. Minou muis bijten
   b. Minou mouse bite

(2) root statement
   a. Minou gaat nu in de muis bijten
      Minou will now in the mouse bite
   b. Minou bijt nu in de muis –
      Minou bites now in the mouse

(3) question
   a. gaat Minou nu in de muis bijten?
      will Minou now in the mouse bite?
   b. bijt Minou nu in de muis –?
      bites Minou now in the mouse –?

(4) imperative
   a. ga nu in de muis bijten, Minou!
      go now in the mouse bite, Minou!
In each variant, the verb *bijten* with its arguments *Minou* and *in de muis* is placed in a different grammatically marked context and the order of verb and arguments changes accordingly. The systematic similarity that binds the variants together is each time the verb *bijten* with those two arguments *Minou* (agent) and *de muis* (goal). The variant requirements of sentence-typing interact with an invariant set of lexical core elements (argument ~ theta frames). The theta frames bring in a lot of idiosyncratic restrictions that are characteristic for the specific language.

Child language clearly reconstructs the patterns (2)-(8) in a stepwise fashion. The present paper will trace the sequence of intermediate steps that lead from the one-word stage to the adult language.

Van Kampen (1997) described a few actual developments in child Dutch as an ordering of grammars $G_0, G_1, \ldots, G_n$, offering an explanation for that order. Child language clearly employs a data-reduction procedure. Initially, the child ignores all functional categories. These functional categories are successively picked up from the input in a stepwise reconstruction of the adult input. This leads to the first “syntax”, which combines two content words in a topic+comment structure (*beer slapen* ‘bear sleep’, *beer lief* ‘bear nice’) or an [illocution-operator + comment] structure (*moet slapen* ‘must sleep’, *is lief* ‘is nice’).

In the next stage, these two-word structures get nested, as in (9) and (10).

\[
\text{(9) beer slapen + moet slapen } \rightarrow \\
\text{ a. beer [moet slapen]}
\text{ (bear [must sleep])} \\
\text{ b. moet [beer slapen]}
\text{ (must [bear sleep])}
\]

\[
\text{(10) Minou bijten + muis bijten } \rightarrow \\
\text{ Minou [muis bijten]}
\text{ (Minou [mouse bite])}
\]

What one sees in child language is the insertion of new functional categories (grammatical markings) within existing frames next to rearrangements of subparts. There is an order of appearance for all innovations and an acquisition model should be able to explain that order. Some functional categories are acquired before others and by being earlier they serve as an acquisition basis for later ones.

The transition in child language from two-word structures (*beer slapen*) towards constructions that merge or rearrange preexisting phrases (*moet – beer slapen, beer – moet slapen*) enables the acquisition of grammatical categories in basic context. Besides these constructions, that suggest a basic frame of functional categories that allows lexical substitution, there seems to be a more advanced and abstract variant that suggests meaningful changes in the order of categories, rather than a mere lexical substitution in a fixed category frame.

The transition in child language from two-word combinations and merged re-arranged constructions towards abstract movement structures, can be demonstrated by considering the acquisition of the Dutch V-second pattern, see (11).

\[
\text{(11) }
\begin{align*}
\text{a. two-word ‘root infinitive’} \\
\text{beer slapen (bear sleep)}
\end{align*}
\]

\[
\begin{align*}
\text{b. merge illocution operator} <\text{aux} > \\
\text{gaat beer ook slapen (goes bear also sleep)}
\end{align*}
\]

\[
\text{c. move finite verb} \\
\text{beer slaapt ook (bear sleeps also –)}
\]

One might wonder why (11)c could not be acquired as a new construction not related to (11)a or (11)b. A more detailed consideration of the acquisition course of (11)c in the next subsection will
show that the relation with types (11)a and (11)b is clearly present.

2.1 An acquisition paradox: the instantaneous acquisition of subordinate word order

About the acquisition of the adult word order in Dutch main clauses and subordinate clauses, we may note the following paradoxical observations:

I. It is known from comparative grammar that the grammatical forms of the main sentence are more varied than those of the subordinate. Moreover, the latter are simpler and closer to the invariant properties. This is sometimes referred to as the Penthouse Principle (Ross 1973: “upstairs”, i.e. in the main sentence, there are more facilities than “downstairs”, i.e. in the subordinate). The child's input thus contains several different types of main clauses. Yet, none of the main clauses realizes the pattern required for the subordinate.

II. The stimulus for the root pattern is remarkably rich. Though 95% of the child's input sentences have a finite verb in first or second position, the finite verb is systematically left out in early child language (disregarding a few stereotypes). The items that will be verbs later on appear in the clause final position. This is the so-called “optional infinitive” stage, (Minou bijten, muis bijten). As we noted above, a second step towards the adult input is the insertion of (finite) modals, aspectuals or the copula in the first or second position. As a last step, the finite denotational verb appears in the V-second position, leaving a gap in the predicate, (Minou bijt muis –). This step takes the child some 15 weeks and roughly a million elementary experiences. It is shortly thereafter (well before the third birthday) that the finite subordinates first appear.

III. The stimulus for the subordinate pattern is remarkably poor, no more than 2% of the input. The subordinate clause pattern has the finite verb in the predicate-final position, see (8)b. This pattern is not realized in the root clauses of the preceding acquisition stages.

And here is the paradox: Although subordinates constitute less than 2% of all input sentences (Van Kampen 1997: chapter 2), the acquisition of the subordinate pattern is instantaneous. The subordinate pattern contradicts the hard-won and massively reinforced pattern of the root sentence.

The question now is: how can a model of language acquisition account for the instantaneous acquisition of the subordinate order in spite of (i) the pattern reversal, (ii) the high input frequency of the main clause, and (iii) the almost negligible input frequency of the subordinate?

The answer to this question lies in the order of the acquisition steps. The acquisition of the verb-second pattern itself was based on a previous acquisition step. All denotational verbs in Dutch have first been acquired as <-finite> forms in the predicate-final position. That previous lexical learning step must have been solidified in the lexical frame for verbs. All Dutch verbs are associated with a lexical argument-verb frame and within that lexical frame the verb takes the final position. It now follows that the subordinate pattern is immediately supported by the lexical frame of the verb, whereas the finite verb requires a change of that frame. The effect of an acquisition stimulus therefore depends on the acquisition steps that have been taken before.

We must conceptualize language acquisition as a linear array of separate acquisition steps arranged in such a way that the later steps are possible only due to the earlier steps. The questionable point of such a reconstruction effort is that the parental input offers all grammatical markings and distinctions simultaneously. Hence, the acquisition procedure must make a selection from the material of a (still) non-analyzable input. Yet, this is possible. The acquisition procedure opens de facto with a thorough but successful reduction on the input.

2.2 Input reduction and its effects

Child language and especially early child language, is reduced with respect to the parental input. A simple common sense principle for the child must be: ignore or leave out what you cannot fit in. Before any syntax is acquired, this will result in leaving out at least all grammatical words (functional categories). The residue will consist of a few, ideally two, words that have an immediate situation-bound meaning or function. The relation between such situation-bound words may be captured pragmatically. If the parental input connects these two situation-bound words with a grammatical marker, such an element can be acquired as an optional approximation of the adult paradigm. The optional addition of such a grammatical word will
become more frequent and ultimately develop into the consistent use of a functional category.

In the next stage, the reduction procedure will reapply to the input with a less dramatic effect. Functional categories that have become identified will pass the reduction filter, and the acquisition procedure may shift its attention to a new functional category or marking that may be added. In principle, the input reduction allows the acquisition procedure to scan the input for the best identifiable grammatical words and markings. Grammatical systems and their use must be such that they allow such a stepwise “decoding procedure”. We may model this as in (12).

(12) Acquisition scheme
a. substitute <F?> for all unidentified grammatical markings
b. remove all input sentences containing more than one <F?> (Single Value Constraint, Berwick 1985).
c. identify the cognitive/pragmatic correspondence of all constructions marked by the same <F?> as the meaning/function of <F?> \( \rightarrow \) \( F_1 \) and reapply the procedure.

Applied to the verbal distributions of Dutch, this will lead to the following four acquisition steps. First step ((13)a). Children learn at first \(<-\text{finite}>\) structures. On the basis of input such as \( \text{gaat Minou de muis bijten?} \) (goes Minou the mouse bite?), the child constructs utterances such as \( \text{Minou muis bijten} \) (‘Minou mouse bite’). The \(<+\text{finite}>\) verbs are ignored.

Second step ((13)b). Shortly afterwards, auxiliary verbs like \( \text{gaat} \) (‘goes’) are added: \( \text{gaat Minou muis bijten} \) (‘goes Minou mouse bite’).

Third step ((13)c). Within a few months, the \(<+\text{fin}>\) denotational verb is inserted at the position of the auxiliary: \( \text{Minou bijt muis} \) (‘Minou bites mouse’).

Fourth step ((13)d). The subordinate appears with the finite verb in predicate-final position (\( \text{als Minou de muis bijt} \) (‘if Minou the mouse bites’).

(13)

a. \( [\emptyset \rightarrow [V_{\text{non-finite}}] ] \) (M.) muis bijten
b. \( [\text{Aux}_{\text{finite}} \rightarrow [V_{\text{non-finite}}] ] \) (M.) gaat muis bijten

de. \( [V_{\text{finite}} \rightarrow [\emptyset] ] \)

c. \( [V_{\text{finite}} \rightarrow [\emptyset] ] \)

The four steps are described in (14), with the input percentages added.

(14)

a. Root infinitives. Nearly all finite verbs are left out

b. Insertion of the illocution sentence type operator \(<+\text{aux}>\)

c. Filling the illocution operator gap by movement of the finite verb

d. Recognizing subordinates as non-illocutions

The four steps are as a matter of fact quantitatively present in the CHILDES files as successive stages. Consider the acquisition graph of Dutch Sarah (Evers & Van Kampen 2001, 2007) in (15).

(15) The rise of finite verbs in V-second position (Dutch Sarah)

Week 105-125 is the optional infinitive period. The graph reflects how that type of structure ebbs away. Sarah’s first subordinate clauses appear in week 135 and they offer no problem as to the verbal positions.
The instantaneous acquisition of the subordinate order may now be explained by a system-internal development that is data-driven. This process consists of five steps:

(i) The lexicon records the underlying verbal position as predicate final due to the initial reduction of all functional elements, including finite verbs.

(ii) The predicate-final position of the verb enters into the lexicon and offers the exemplar-based frame for the subcategorization of predicate heads/verbs. Fortunately, it has been shown that the linear argument structure can be identified as a subcategorizing argument frame that is associated with the predicate heads (Brent 1994). We now propose that (13)b and (13)c around week 115 (at the 60% level) can still be modeled as “alternative exemplars”. As soon as the overlap develops into a substantial percentage of the adult norm at week 122 (the 80% level), each denotational verb has some 1/3 (34%) chance to appear as <+finite> and 2/3 chance to appear as <−finite>.

(iii) The set of items that allows this variation defines the category <+V>. Each <+finite> pair shares the same argument frame.

(iv) The fact that the denotational V<+finite> corresponds with an empty place in predicate-final position, is no longer something that is noticed for each item separately. It is a categorial property of the Dutch verb.

(v) The learnability of subordinates follows now as a simplified version of the earlier pattern (13)c.

If the pattern shift (13)c is categorial rather than item-by-item, one expects the points in (16).

(16)

a. absence of exceptions, verb-like items that lack a +finite> or <−finite> variant.

b. sudden rise of verbs that appear as <+finite>.

c. the instantaneous acquisition of the subordinate pattern (13)d.

An acquisition model needs something like rearrangements to represent the succession of the exemplars (13)b,c,d as a system-internal development.

This observation about the acquisition process fits an observation about the adult language. As we saw in (2)-(8) above, there is a variety of verb-argument patterns in Dutch; the choice depends on the type of clause involved. A grammatical model that does not provide a verb movement operation cannot describe the systematic relations between these patterns. It would not be able to derive a main-clause pattern from a subordinate-clause pattern (or vice versa). It would therefore be forced to learn the patterns one by one for each combination of verb and clause-type. Such an approach might ultimately be able to reproduce the distributional facts, but it would fail to draw any advantage from the exception-less regularity of these patterns: in Dutch main and subordinate order are systematically different without any lexical exception. It would especially fail to predict the instantaneous character of the actual acquisition step.

The variation between root and subordinate exemplars can be captured if they are related by derivational rules. Interestingly, the acquisition procedure itself betrays that, due to an initial reduction of the input strings, exemplars displaying the S-O-V<−finite> structure are stored before the derived exemplars displaying the S-V<+finite>-O order appear.

2.3 An acquisition procedure

An empirically adequate acquisition algorithm should simulate the acquisition of V-second along the lines above. It should select the data without a priori guidance and predict the actual order of acquisition steps. It would, for every new input string (i) perform the input reduction.

(ii) select a frame for a new functional category, such that:

a. the syntactic/morphological form and the semantic/pragmatic form of the category is identified

b. the match between actual input and reduced intake is improved

c. a result can be stored in the acquisitional memory base of the model

By processing a large enough corpus of child-directed utterances, the program should be able to derive all the grammatical markings, and at least the V-second and subordinate patterns. The model may support the following perspective:

a. The actual order of acquisition steps reveals a hierarchy between less variant and more variant properties of the structures of the language.
b. Clause structure is a matter of transformationally deriving the sentence typing and variant parts of the structure from the lexical invariants.
c. Functional categories that trigger movement transformations allow a simpler lexicon and a short-cut to the acquisition of the variety in grammatical patterns.

3 Language acquisition and Data-Oriented Parsing

3.1 DOP

Scha (1990, 1992) argued that human language processing crucially employs a memory for previously processed analyses and their occurrence frequencies: a new input string gets the analysis that displays the simplest analogies with the largest number of stored structures. So far, computational models based on this approach usually employ some form of Stochastic Tree Substitution Grammar (STSG) (cf. Bod 1995; Bod, Scha & Sima'an 2003).

Data-Oriented Parsing differs from conventional STSG in its choice of the set of elementary trees of the grammar. This is a very large and redundant set: the sub-trees of all utterance-analyses in an annotated corpus that is assumed to represent the "past linguistic experiences" of the language user that is being modelled.

In a language model of this sort, a person's "knowledge of language" is not summarized in a concise grammar. It is distributed across a large, redundant data set that allows competing sets of rules and exceptions. The model is "exemplar-based" in that all successful analyses are stored and may have an influence on future processing. But it has the same kind of recursive productivity as a conventional grammar, because every exemplar implies all its abstractions.

For instance, if a child-language corpus includes the exemplars

S:[topic: α, pred: slapen ]
S:[topic:beertje, pred: β ]
S:[topic: α, pred: β ]
S:[topic: α, pred:[patiens: χ, action: bijten ]]
S:[topic: α, pred:[patiens: χ, action: δ ]]

Note that these abstractions from the exemplars include the subcategorization frames of the verbs (which in a conventional grammar would be stored in a "lexicon"), as well as completely abstract structures that in a conventional grammar would be called (context-free) rules – and indeed, they function as such. The important difference with a conventional frame or a conventional CFG-rule lies in the fact that in a DOP-model the abstract structures must always compete with more completely lexicalized tree-fragments.

What may be substituted for α, β, χ, δ depends on the intersubstitutability-classes ("categories") that have been developed at this stage.

The redundancy of the DOP approach has an advantage when we model the process of language acquisition: new constructions can emerge gradually and supersede existing ones. Language acquisition itself may be viewed as an evolutionary process. Certain patterns may become dominant and reanalyze the domain of their competitors. How this idea pans out in practice is not yet known. We need both a computational model that implements it and an assessment of its compatibility with the empirical facts of language acquisition. Below we explore the prospects of such an enterprise.

3.1 Modelling the acquisition of Dutch verbal placements

The earliest stage of a person's linguistic development is sometimes viewed as particularly problematic for the data-oriented approach, since at that point there is as yet no corpus of previously analyzed input. How does language acquisition begin? This question highlights a semantic/pragmatic component of language processing which is to some extent independent of syntax. A language user always tries to match the semantic/pragmatic context of an utterance with its possible interpretations. In the early stages of language use, this component must be the dominant one. Later, the
influence of past language experience becomes increasingly prominent (cf. Scha 1990).

Consider, for instance, the stereotype topic-predicate meanings of the child's first word combinations. Constructions like [beer slapen] (‘bear sleep’) [pappa lief] (‘daddy nice’) are built up by connecting a name with a word for a situation type, borrowing saliently stressed words from the adult utterances. A computational model demonstrating this process is presented by Chang & Gurevich (2004).

In the present paper we focused on a later stage in the acquisition of Dutch. After two-year olds have built up a corpus of one-word and two-word sentences they go through an interesting syntactic development. They insert modals and copulas between the topic and the predicate, thus moving from stage (13)a to (13)b. The new structure (13)b is extended later on by the introduction of finite denotational verbs.

In the adult speech that the child encounters, the surface word order patterns (17)a,b,c occur, whereas the patterns (17)d,e do not occur.

(17)

a. S Aux OV<finite> Minou gaat muis bijten
b. S V<finite> O Minou bijt muis
c. V<finite>SO nu bijt Minou muis
d. *S Aux VO

e. *Aux S VO

The sentence types (17)a,b,c are all in the input, and all allow the type assignments for subject, object and (denotational) verb, but due to the large number of auxiliaries (>66%, see (14)) the pattern [S (Aux) OV] (17)a is dominant whereas the others are marginal. The primary frame associated with the proto-verb bijten is OV. The same holds for all other denotational verbs. One may interpret the appearance of the later construction (13)c as evidence that new exemplars with SVO structure are stored now, next to the analyses with pattern (13)b, which maintains its more than 2/3 dominance. If, by contrast, we assume that the earlier exemplars displaying the OV pattern ((13)a, (13)b) exert an overriding influence, the model needs a transformational move, modulo V<+finite>, to derive (13)c from (13)a and (13)b.

How to decide which of these conjectures is the right one? The answer comes from the next step in the acquisition process: (13d). In Dutch subordinate clauses (als Minou de muis bijt), the verb appears at the end. The subordinate clause is thus markedly different from the main clause, while it closely resembles the sentence of early child Dutch. Above we observed that Dutch children acquire the subordinate clause word order extremely fast, on the basis of very sparse input.

An elegant explanation in terms of a DOP-model would be, that the exemplars displaying the early child language word order are still accessible in the corpus. The underlying structures which are assumed in transformational grammar, would then be available due to earlier acquisition steps. The early exemplars thus play the same role as fixed theta frames (Brent 1994) in conventional grammars.

However, this account should also be applicable to verbs that the child encounters in main-clause exemplars, after it has acquired the verb-second word order. For that to be the case, there must be a cognitive relation between verb-second main clauses and their verb-final infinitival counterparts. This can be done in several ways.

Scha (1990) suggested the use of transformations in the style of Chomskyan grammars. For the example in (17)b this would mean that the model can construct the syntactic tree for Minou bijt de muis by applying the verb-second transformation to Minou de muis bijten. To achieve this, the generative machinery of the Stochastic Tree Substitution Grammar would have to be enriched with movement transformations.

An alternative possibility is suggested by the work on LFG-DOP, where the generative model consistently works with pairs of (sub-)trees: a surface (sub-)tree aligned with a tree that represents the "functional structure" of the sentence (cf. Bod & Kaplan 2003). In that approach, one should probably assume that in early child language surface structure and functional structure coincide. For the Dutch child, the acquisition of verb-second word order marks the moment that surface structure and functional structure split apart: the surface tree of a verb-second utterance is paired with the tree of the corresponding verb-final utterance to represent its functional structure.
4 Conclusion

There are developmental arguments for the thesis that Dutch is an SOV-language: SOV is the order of early Dutch child language, and the steps of the language acquisition process can only be understood if we assume that the early SOV-frames remain available in the later stages.

Exemplar-based language models may be able to explain this process, if they allow early exemplars to remain accessible long enough. Models that are exclusively based on tree-substitution, however, do not seem very plausible, since they cannot connect Dutch main clause word order to the underlying SOV pattern.

References


