GraSp: Grammar learning from unlabelled speech corpora

Peter Juel Henrichsen
CMOL
Center for Computational Modelling of Language
c/o Dept. of Computational Linguistics
Copenhagen Business School
Frederiksberg, Denmark
pjuel@id.cbs.dk

Abstract

This paper presents the ongoing project Computational Models of First Language Acquisition, together with its current product, the learning algorithm *GraSp*. GraSp is designed specifically for inducing grammars from large, unlabelled corpora of spontaneous (i.e. unscripted) speech. The learning algorithm does not assume a predefined grammatical taxonomy; rather the determination of categories and their relations is considered as part of the learning task. While GraSp learning can be used for a range of practical tasks, the long-term goal of the project is to contribute to the debate of innate linguistic knowledge – under the hypothesis that there is no such.

Introduction

Most current models of grammar learning assume a set of primitive linguistic categories and constraints, the learning process being modelled as category *filling* and rule *instantiation* – rather than category *formation* and rule *creation*. Arguably, distributing linguistic data over predefined categories and templates does not qualify as grammar 'learning' in the strictest sense, but is better described as 'adjustment' or 'adaptation'. Indeed, Chomsky, the prime advocate of the hypothesis of innate linguistic principles, has claimed that "in certain fundamental respects we do not really learn language" (Chomsky 1980: 134). As Chomsky points out, the complexity of the learning task is greatly reduced given a structure of primitive linguistic constraints ("a highly restrictive schematism", ibid.). It has however been very hard to establish independently the psychological reality of such a structure, and the question of innateness is still far from settled.

While a decisive experiment may never be conceived, the issue could be addressed indirectly, e.g. by asking: Are innate principles and parameters *necessary* preconditions for grammar acquisition? Or rephrased in the spirit of constructive logic: Can a learning algorithm be devised that learns what the infant learns without incorporating specific linguistic axioms? The presentation of such an algorithm would certainly undermine arguments referring to the 'poverty of the stimulus', showing the innateness hypothesis to be dispensable.

This paper presents our first try.

1 The essential algorithm

1.1 Psycho-linguistic preconditions

Typical spontaneous speech is anything but syntactically 'well-formed' in the Chomskyan sense of the word.

> right well let's er --= let's look at the applications - erm - let me just ask initially this -- I discussed it with er Reith er but we'll = have to go into it a bit further - is it is it within our erm er = are we free er to er draw up a rather = exiguous list - of people to interview
>
> *(sample from the London-Lund corpus)*

Yet informal speech is not perceived as being disorderly (certainly not by the language learning infant), suggesting that its organizing principles differ from those of the written language. So, arguably, a speech grammar inducing algorithm should avoid referring to the usual categories of text based linguistics – 'sentence', 'determiner phrase', etc.[1]

Instead we allow a large, indefinite number of (indistinguishable) basic categories – and then leave it to the learner to shape them, fill them up, and combine them. For this task, the learner needs a built-in concept of constituency. This kind of innateness is not in conflict with our main hypothesis, we believe, since constituency *as such* is not specific to linguistic structure.

1.2 Logical preliminaries

For the reasons explained, we want the learning algorithm to be strictly data-driven. This puts special demands on our parser which must be robust enough to accept input strings with little or no hints of syntactic structure (for the early stages of a learning session), while at the same time retaining the discriminating powers of a standard context free parser (for the later stages).

Our solution is a sequent calculus, a variant of the Gentzen-Lambek categorial grammar formalism (**L**) enhanced with non-classical rules for isolating a residue of uninterpretable sequent elements. The classical part is identical to **L** (except that antecedents may be empty).

Classical part

$$\frac{}{\sigma \Rightarrow \sigma}\,\mathbf{link}$$

$$\frac{\Delta_B \Rightarrow B \qquad \Delta_1\ A\ \Delta_2 \Rightarrow C}{\Delta_1\ A/B\ \Delta_B\ \Delta_2 \Rightarrow C}\,/\mathbf{L} \qquad \frac{\Delta_0\ B \Rightarrow A}{\Delta_0 \Rightarrow A/B}\,/\mathbf{R}$$

$$\frac{\Delta_B \Rightarrow B \qquad \Delta_1\ A\ \Delta_2 \Rightarrow C}{\Delta_1\ \Delta_B\ B\backslash A\ \Delta_2 \Rightarrow C}\,\backslash\mathbf{L} \qquad \frac{B\ \Delta_0 \Rightarrow A}{\Delta_0 \Rightarrow B\backslash A}\,\backslash\mathbf{R}$$

$$\frac{\Delta_1\ A\ B\ \Delta_2 \Rightarrow C}{\Delta_1\ A{*}B\ \Delta_2 \Rightarrow C}\,{*}\mathbf{L} \qquad \frac{\Delta_1 \Rightarrow A \qquad \Delta_2 \Rightarrow B}{\Delta_1\ \Delta_2 \Rightarrow A{*}B}\,{*}\mathbf{R}$$

A, B, C are categories; Δ_x are (possibly empty) strings of categories.

These seven rules capture the input parts that can be interpreted as syntactic constituents (examples below). For the remaining parts, we include two non-classical rules (σ**L** and σ**R**).[2]

Non-classical part

$$\frac{\sigma^+ \qquad \Delta_1\ \Delta_2 \Rightarrow C}{\Delta_1\ \sigma\ \Delta_2 \Rightarrow C}\,\sigma\mathbf{L} \qquad \frac{\sigma^-}{\Rightarrow \sigma}\,\sigma\mathbf{R}$$

σ is a basic category. Δ_x are (possibly empty) strings of categories. Superscripts $^+$ $^-$ denote polarity of residual elements.

By way of an example, consider the input string

right well let's er let's look at the applications

as analyzed in an early stage of a learning session. Since no lexical structure has developed yet, the input is mapped onto a sequent of basic (dummy) categories:[3]

$$c_{29}\ c_{22}\ c_{81}\ c_5\ c_{81}\ c_{215}\ c_{10}\ c_1\ c_{891} \Rightarrow c_0$$

Using σ**L** recursively, each category of the antecedent (the part to the left of $\Rightarrow$) is removed from the main sequent. As the procedure is fairly simple, we just show a fragment of the proof. Notice that proofs read most easily bottom-up.

$$\cfrac{c_5^+ \qquad \cfrac{c_{215}^+ \qquad \cfrac{\cfrac{c_{81}^+\ c_{10}^+\ c_1^+\ c_{891}^+ \qquad \cfrac{c_0^-}{\Rightarrow c_0}\,\sigma\mathbf{R}}{\cdots}\,\sigma\mathbf{L}}{c_{81}\ c_{10}\ c_1\ c_{891} \Rightarrow c_0}\,\sigma\mathbf{L}}{c_{81}\ c_{215}\ c_{10}\ c_1\ c_{891} \Rightarrow c_0}\,\sigma\mathbf{L}}{\ldots\ c_5\ c_{81}\ c_{215}\ c_{10}\ c_1\ c_{891} \Rightarrow c_0}\,\sigma\mathbf{L}$$

In this proof there are no **link**s, meaning that no grammatical structure was found. Later, when the lexicon has developed, the parser may

[1] Hoekstra (2000) and Nivre (2001) discuss the annotation of spoken corpora with traditional tags.

[2] The calculus presented here is slightly simplified. Two rules are missing, and so is the reserved category $\top$ ('noise') used e.g. for consequents (in place of c_0 of the example). Cf. Henrichsen (2000).

[3] By convention the indexing of category names reflects the frequency distribution: If word W has rank n in the training corpus, it is initialized as W:c_n.

recognize more structure in the same input:

$$\dfrac{\dfrac{}{c_1 \Rightarrow c_1}\,\mathrm{I} \qquad \dfrac{\dfrac{\dfrac{}{c_{10}\Rightarrow c_{10}}\,\mathrm{I} \quad \dfrac{}{c_{891}\Rightarrow c_{891}}\,\mathrm{I}}{c_{10}\; c_{891} \Rightarrow c_{10}{}^{*}c_{891}}\,{*}\mathbf{R} \qquad c_{81}\; c_{215} \Rightarrow c_0}{c_{81}\;\; c_{215}/(c_{10}{}^{*}c_{891})\;\; c_{10}\;\; c_{891} \Rightarrow c_0}\,/\mathbf{L}}{\ldots\; c_{81}\;\; c_{215}/(c_{10}{}^{*}c_{891})\;\; c_{10}\;\; c_1\;\; c_1\backslash c_{891} \;\Rightarrow c_0}\,\backslash\mathbf{L}$$

... let's look at the applications

This proof tree has three **link**s, meaning that the disorder of the input string (wrt. the new lexicon) has dropped by three degrees. More on *disorder* shortly.

1.3 The algorithm in outline

Having presented the sequent parser, we now show its embedding in the learning algorithm GraSp (**Gra**mmar of **Sp**eech).

For reasons mentioned earlier, the common inventory of categories (S, NP, CN, etc) is avoided. Instead each lexeme initially inhabits its own proto-category. If a training corpus has, say, 12,345 word types the initial lexicon maps them onto as many different categories. A learning session, then, is a sequence of lexical changes, introducing, removing, and manipulating the operators /, \, and * as guided by a well-defined measure of structural disorder.

We prefer formal terms without a linguistic bias ("no innate linguistic constraints"). Suggestive linguistic interpretations are provided in square brackets.

A-F summarize the learning algorithm.

A) There are **categories**. Complex categories are built from basic categories using /, \, and *:

<u>Basic categories</u>

$c_1,\ c_2,\ c_3, \ldots,\ c_{12345}, \ldots$

<u>Complex categories</u>

$c_1\backslash c_{12345},\ c_2/c_3,\ c_4{}^{*}c_5,\ c_2/(c_3\backslash(c_4{}^{*}c_5))$

B) A **lexicon** is a mapping of lexemes [word types represented in phonetic or enriched-orthographic encoding] onto categories.

C) An input **segment** is an instance of a lexeme [an input word]. A **solo** is a string of segments [an utterance delimited by e.g. turntakes and pauses]. A **corpus** is a bag of soli [a transcript of a conversation].

D) Applying an **update** $L{:}C_1{\rightarrow}C_2$ in lexicon Lex means changing the mapping of L in Lex from C_1 to C_2. Valid changes are *minimal*, i.e. C_2 is construed from C_1 by adding or removing 1 basic category (using \, /, or *).

E) The learning process is guided by a measure of **disorder**. The disorder function Dis takes a sequent Σ [the lexical mapping of an utterance] returning the number of uninterpretable atoms in Σ, i.e. σ^+s and σ^-s in a (maximally linked) proof. $Dis(\Sigma)=0$ iff Σ is Lambek valid. Examples:

$$\begin{array}{ll} Dis(\ c_a/c_b\ \ c_b \Rightarrow c_a\ ) & = 0 \\ Dis(\ c_a/c_b\ \ c_b \Rightarrow c_c\ ) & = 2 \\ Dis(\ c_b\ \ c_a/c_b \Rightarrow c_c\ ) & = 4 \\ Dis(\ c_a/c_b\ \ c_c\ \ c_b \Rightarrow c_a\ ) & = 1 \\ Dis(\ c_a/c_c\ \ c_b\ \ c_a\backslash c_c \Rightarrow c_a\ ) & = 2 \end{array}$$

$DIS(Lex,K)$ is the total amount of disorder in training corpus K wrt. lexicon Lex, i.e. the sum of Dis-values for all soli in K as mapped by Lex.

F) A **learning session** is an iterative process. In each iteration i a suitable update U_i is applied in the lexicon Lex_{i-1} producing Lex_i. Quantifying over all possible updates, U_i is picked so as to maximize the drop in disorder ($DisDrop$):

$$DisDrop \;=\; DIS(Lex_{i-1},K) - DIS(Lex_i,K)$$

The session terminates when no suitable update remains.

It is possible to GraSp *efficiently* and yet preserve logical completeness. See Henrichsen (2000) for discussion and demonstrations.

1.4 A staged learning session

Given this tiny corpus of four soli ('utterances')

if you must you can
if you must you must and if we must we must
if you must you can and if you can you must
if we must you must and if you must you must

, GraSp produces the lexicon below.

Lexeme	Initial Category	Final Category[4]	Textbook Category
must	c_1	$c_2\backslash c_1$	NP\S
you	c_2	c_2	NP
if	c_3	$(c_3/c_1)/c_1$	(S/S)/S
and	c_4	$(c_3\backslash c_4)/c_3$	(S\S)/S
can	c_5	$c_2\backslash c_1$	NP\S
we	c_6	c_2	NP

As shown, training corpora can be manufactured so as to produce lexical structure fairly similar to what is found in CG textbooks. Such close similarity is however not typical of 'naturalistic' learning sessions – as will be clear in section 2.

1.5 Why categorial grammar?

In CG, *all* structural information is located in the lexicon. Grammar rules (e.g. $VP \rightarrow V_t\ N$) and parts of speech (e.g. 'transitive verb', 'common noun') are treated as variants of the same formal kind. This reduces the dimensionality of the logical learning space, since a CG-based learner needs to induce just a single kind of structure.

Besides its formal elegance, the CG basis accomodates a particular kind of cognitive models, viz. those that reject the idea of separate mental modules for lexical and grammatical processing (e.g. Bates 1997). As we see it, our formal approach allows us the luxury of not taking sides in the heated debate of modularity.[5]

2 Learning from spoken language

The current GraSp implementation completes a learning session in about one hour when fed with our main corpus.[6] Such a session spans 2500-4000 iterations and delivers a lexicon rich in microparadigms and microstructure. Lexical structure develops mainly around content words while most function words retain their initial category. The structure grown is almost fractal in character with lots of inter-connected categories, while the traditional large open classes – nouns, verbs, prepositions, etc. – are absent as such. The following sections present some samples from the main corpus session (Henrichsen 2000 has a detailed description).

2.1 Microparadigms

{ "Den Franske", "Nyboder",
"Sølvgades", "Krebses" }

These four lexemes – or rather lexeme clusters – chose to co-categorize. The collection does not resemble a traditional syntactic paradigm, yet the connection is quite clear: all four items appeared in the training corpus as *names of primary schools*.

Lexeme	Initial Category	Final Category
Den	c_{882}	c_{882}
Franske	c_{1588}	$((c_{882}\backslash c_{97})/c_{1588})^*c_{1588}$
Nyboder	c_{97}	c_{97}
Sølvgades	c_{5351}	$(c_{97}/c_{5351})^*c_{5351}$
Krebses	c_{3865}	$(c_{3865}/c_{288})^*c_{97}$
Skole	c_{288}	$c_{97}\backslash c_{288}$

The final categories are superficially different, but are easily seen to be functionally equivalent.

The same session delivered several other microparadigms: a collection of family members (in English translation: *brother*, *grandfather*, *younger-brother*, *stepfather*, *sister-in-law*, etc.), a class of negative polarity items, a class of mass terms, a class of disjunctive operators, etc. (Henrichsen 2000 6.4.2).

GraSp-paradigms are usually small and almost always intuitively 'natural' (not unlike the small categories of L1 learners reported by e.g. Lucariello 1985).

2.2 Microgrammars

GraSp'ed grammar rules are generally not of the kind studied within traditional phrase structure grammar. Still PSG-like 'islands' do occur, in the form of isolated networks of connected lexemes.

---

[4] For perspicuity, two of the GraSped categories – viz. 'can':$(c_2\backslash c_5)^*(c_5\backslash c_1)$ and 'we':$(c_2/c_6)^*c_6$ – are replaced in the table by functional equivalents.

[5] A caveat: Even if we do share some tools with other CG-based NL learning programmes, our goals are distinct, and our results do not compare easily with e.g. Kanazawa (1994), Watkinson (2000). In terms of philosophy, GraSp seems closer to connectionist approaches to NLL.

[6] The Danish corpus BySoc (person interviews). *Size*: 1.0 mio. words. *Duration*: 100 hours. *Style*: Labovian interviews. *Transcription*: Enriched orthography. *Tagging*: none. *Ref.*: http://www.cphling.dk/BySoc

Lexeme	Initial Category	Final Category	Con-nection
Sankt	c_{620}	c_{620}	c_{620}^{+}
Sct.	c_{4713}	$(c_{620}/c_{4713})^{*}c_{4713}$	
Skt.	c_{3301}	$(c_{620}/c_{3301})^{*}c_{3301}$	
Annæ	c_{3074}	$c_{620}\backslash(c_{22}\backslash c_{3074})$	c_{620}^{-}
Josef	c_{2921}	$c_{620}\backslash c_{2921}$	
Joseph	c_{3564}	$c_{620}\backslash c_{3564}$	
Knuds	c_{6122}	$c_{620}\backslash c_{6122}$	
Pauls	c_{1218}	$c_{620}\backslash c_{1218}$	
Paulsgade	c_{2927}	$c_{620}\backslash c_{2927}$	
Pouls	c_{2180}	$c_{620}\backslash c_{2180}$	
Poulsgade	c_{4707}	$c_{620}\backslash c_{4707}$	
Pauls	c_{1218}	$c_{620}\backslash c_{1218}$	c_{1218}^{+}
Gade	c_{3849}	$c_{1218}\backslash(c_{9}\backslash c_{3849})$	c_{1218}^{-}
Plads	c_{1263}	$c_{1218}\backslash(c_{22}\backslash c_{1263})$	

Centred around lexeme 'Pauls', a microgrammar (of street names) has evolved almost directly translatable into rewrite rules:[7]

PP $\rightarrow$ 'i' N_1 'Gade'
PP $\rightarrow$ 'på' N_1 'Plads'
PP $\rightarrow$ 'på' N_2
N_1 $\rightarrow$ X 'Pauls'
N_2 $\rightarrow$ X 'Annæ'
N_x $\rightarrow$ X Y
X $\rightarrow$ 'Sankt' | 'Skt.' | 'Sct.'
Y $\rightarrow$ 'Pauls' | 'Josef' | 'Joseph' | 'Knuds' | ...

2.3 Idioms and locutions

Consider the five utterances of the main corpus containing the word 'rafle' (*cast-dice*$_{\text{INF}}$):[8]

det gør den der er ikke noget at <u>rafle</u> om der
der er ikke så meget at <u>rafle</u> om
der er ikke noget og <u>rafle</u> om
sætte sig ned og <u>rafle</u> lidt med fyrene der
at <u>rafle</u> om der

On most of its occurrences, 'rafle' takes part in the idiom "der er ikke noget/meget og/at rafle om", often followed by a resumptive 'der' (literally: *there is not anything/much and/to cast-dice*$_{\text{INF}}$ *about* (*there*), meaning: this is not a subject of negotiations). Lexeme 'ikke' (category c_8) occurs in the left context of 'rafle' more often than not, and this fact is reflected in the final category of 'rafle':

rafle: $((c_{12}\backslash(\mathbf{c_8}\backslash(c_5\backslash(c_7\backslash c_{5808}))))/c_7)/c_{42}$

Similarly for the lexemes 'der' (c_7), 'er' (c_5), 'at' (c_{12}), and 'om' (c_{42}) which are also present in the argument structure of the category, while the top functor is the initial 'rafle' category (c_{5808}).

The minimal context motivating the full *rafle* category is:

... der ... er ... ikke ... at ... **rafle** ... om ... der ...

("..." means that any amount and kind of material may intervene). This template is a quite accurate description of an acknowledged Danish idiom.

Such idioms have a specific categorial signature in the GraSped lexicon: a rich, but flat argument structure (i.e. analyzed solely by $\sigma\mathbf{R}$) centered around a single low-frequency functor (analyzed by $\sigma\mathbf{L}$). Further examples with the same signature:

... det ... kan ... man ... ikke ... **fortænke** ... i ...

... det ... vil ... **blæse** ... på ...

... ikke ... en ... **kinamands** ... chance ...

– all well-known Danish locutions.[9]

There are of course plenty of simpler and faster algorithms available for extracting idioms. Most such algorithms however include specific knowledge about idioms (topological and morphological patterns, concepts of mutual information, heuristic and statistical rules, etc.). Our algorithm has no such inclination: it does not *search* for idioms, but merely *finds* them.

Observe also that GraSp may induce idiom templates like the ones shown even from corpora without a single verbatim occurrence.

---

[7] Lexemes 'Sankt', 'Sct.', and 'Skt.' have in effect cocategorized, since it holds that $(x/y)^{*}y \Rightarrow x$. This cocategorization is quite neat considering that GraSp is blind to the interior of lexemes. c_9 and c_{22} are the categories of 'i' (*in*) and 'på' (*on*).

[8] In writing, only two out of five would probably qualify as syntactically well-formed sentences.

[9] For entry **rafle**, Danish-Danish dictionary Politiken has this paradigmatic example: "Der er ikke noget at rafle om". Also **fortænke**, **blæse**, **kinamands** have examples near-identical with the learned templates.

3 Learning from exotic corpora

In order to test GraSp as a general purpose learner we have used the algorithm on a range of non-verbal data. We have had GraSp study melodic patterns in musical scores and prosodic patterns in spontaneous speech (and even dna-structure of the banana fly). Results are not yet conclusive, but encouraging (Henrichsen 2002).

When fed with HTML-formatted text, GraSp delivers a lexical patchwork of linguistic structure and HTML-structure. GraSp's uncritical appetite for *context-free structure* makes it a candidate for intelligent web-crawling. We are preparing an experiment with a large number of cloned learners to be let loose in the internet, reporting back on the structure of the documents they see. Since GraSp produces formatting definitions as output (rather than requiring it as input), the algorithm could save the www-programmer the troubles of preparing his web-crawler for this-and-that format.

Of course such experiments are side-issues. However, as discussed in the next section, learning from non-verbal sources may serve as an inspiration in the L1 learning domain also.

4 Towards a model of L1 acquisition

4.1 Artificial language learning

Training infants in language tasks within artificial (i.e. semantically empty) languages is an established psycho-linguistic method. Infants have been shown able to extract structural information – e.g. rules of *phonemic segmentation, prosodic contour*, and even *abstract grammar* (Cutler 1994, Gomez 1999, Ellefson 2000) – from streams of carefully designed nonsense. Such results are an important source of inspiration for us, since the experimental conditions are relatively easy to simulate. We are conducting a series of 'retakes' with the GraSp learner in the subject's role. Below we present an example.

In an often-quoted experiment, psychologist Jenny Saffran and her team had eight-months-old infants listening to continuous streams of nonsense syllables: *ti*, *do*, *pa*, *bu*, *la*, *go*, etc. Some streams were organized in three-syllable 'words' like *padoti* and *golabu* (repeated in random order) while others consisted of the same syllables in random order. After just two minutes of listening, the subjects were able to distinguish the two kinds of streams. Conclusion: Infants can learn to identify compound words on the basis of structural clues alone, in a semantic vacuum.

Presented with similar streams of syllables, the GraSp learner too discovers word-hood.

Lexeme	Initial Category	Final Category[10]
pa	c_2	c_2
do	c_1	$(c_2\backslash c_1)/c_3$
ti	c_3	c_3
go	c_5	c_5
la	c_6	c_6
bu	c_4	$c_6\backslash(c_5\backslash c_4)$
...	...	...

It may be objected that such streams of presegmented syllables do not represent the experimental conditions faithfully, leaping over the difficult task of *segmentation*. While we do not yet have a definitive answer to this objection, we observe that replacing "pa do ti go la bu (..)" by "p a d o t i g o l a b u (..)" has the GraSp learner discover syllable-hood *and* word-hood on a par.[11]

4.2 Naturalistic language learning

Even if human learners can demonstrably learn structural rules without access to semantic and pragmatic cues, this is certainly not the typical L1 acquisition scenario. Our current learning model fails to reflect the natural conditions in a number of ways, being a purely syntactic calculus working on symbolic input organized in well-delimited strings. Natural learning, in contrast, draws on far richer input sources:

- continuous (unsegmented) input streams
- suprasegmental (prosodic) information
- sensory data
- background knowledge

[10] As seen, *padoti* has selected *do* for its functional head, and *golabu*, *bu*. These choices are arbitrary.

[11] The very influential Eimas (1971) showed one-month-old infants to be able to distinguish /p/ and /b/. Many follow-ups have established that phonemic segmentation develops very early and may be innate.

Any model of first language acquisition must be prepared to integrate such information sources. Among these, the extra-linguistic sources are perhaps the most challenging, since they introduce a syntactic-semantic *interface* in the model. As it seems, the formal simplicity of one-dimensional learning (cf. sect. 1.5) is at stake.

If, however, semantic information (such as sensory data) could be 'syntactified' and included in the lexical structure in a principled way, single stratum learning could be regained. We are currently working on a formal upgrading of the calculus using a framework of constructive type theory (Coquant 1988, Ranta 1994). In CTT, the radical lexicalism of categorial grammar is taken even a step further, representing *semantic* information in the same data structure as *grammatical* and *lexical* information. This formal upgrading takes a substantial refinement of the Dis function (cf. sect. 1.3 E) as the determination of 'structural disorder' must now include contextual reasoning (cf. Henrichsen 1998). We are pursuing a design with σ^+ and σ^- as instructions to respectively *insert* and *search for* information in a CTT-style context.

These formal considerations are reflections of our cognitive hypotheses. Our aim is to study learning as a radically data-driven process drawing on linguistic and extra-linguistic information sources on a par – and we should like our formal system to fit like a glove.

5 Concluding remarks

As far as we know, GraSp is the first published algorithm for extracting grammatical taxonomy out of untagged corpora of spoken language.[12] This in an uneasy situation, since if our findings are not comparable to those of other approaches to grammar learning, how could our results be judged – or falsified? Important issues wide open to discussion are: validation of results, psycho-linguistic relevance of the experimental setup, principled ways of surpassing the context-free limitations of Lambek grammar (inherited in GraSp), just to mention a few.

On the other hand, already the spin-offs of our project (the collection of non-linguistic learners) do inspire confidence in our tenets, we think – even if the big issue of psychological realism has so far only just been touched.

[12] The learning experiment sketched in Moortgat (2001) shares some of GraSp's features.

The GraSp implementation referred to in this paper is available for test runs at

`http://www.id.cbs.dk/~pjuel/GraSp`

References

Bates, E.; J.C. Goodman (1997) *On the Inseparability of Grammar and the Lexicon: Evidence From Acquisition, Aphasia, and Real-time Processing*; Language and Cognitive Processes 12, 507-584

Chomsky, N. (1980) *Rules and Representations*; Columbia Univ. Press

Coquant, T.; G. Huet (1988) *The Calculus of Constructions*; Info. & Computation 76, 95-120

Cutler, A. (1994) *Segmentation Problems, Rhythmic Solutions*; Lingua 92, 81-104

Eimas, P.D.; E.D. Siqueland; P.W. Jusczyk (1971) *Speech Perception in Infants*; Science 171 303-306

Ellefson, M.R.; M.H.Christiansen (2000) *Subjacency Constraints Without Universal Grammar: Evidence from Artificial Language Learning and Connectionist Modelling*; 22nd Ann. Conference of the Cognitive Science Society, Erlbaum, 645-650

Gomez, R.L.; L.A. Gerken (1999) *Artificial Grammar Learning by 1-year-olds Leads to Specific and Abstract Knowledge*; Cognition 70 109-135

Henrichsen, P.J. (1998) *Does the Sentence Exist? Do We Need It?*; in K. Korta et al. (eds) *Discourse, Interaction, and Communication*; Kluwer Acad.

Henrichsen, P.J. (2000) *Learning Within Grasp – Interactive Investigations into the Grammar of Speech*; Ph.D., http://www.id.cbs.dk/~pjuel/GraSp

Henrichsen, P.J. (2002) *GraSp: Grammar Learning With a Healthy Appetite* (in prep.)

Hoekstra, H. et al. (2000) *Syntactic Annotation for the Spoken Dutch Corpus Project*; CLIN2000

Kanazawa (1994) *Learnable Classes of CG*; Ph.D.

Moortgat, M. (2001) *Structural Equations in Language Learning*; 4th LACL2001 1-16

Nivre, J.; L. Grönqvist (2001) *Tagging a Corpus of Spoken Swedish*: Int. Jn. of Corpus Ling. 6:1 47-78

Ranta, A. (1994) *Type-Theoretical Grammar*; Oxford

Saffran, J.R. et al. (1996) *Statistical Learning By 8-Months-Old Infants*; Science 274 1926-1928

Watkinson S.; S. Manandhar (2000) *Unsupervised Lexical Learning with CG*; in Cussens J. et al. (eds) *Learning Language in Logic*; Springer