going-to-V and *gonna-V* in child language: A quantitative approach to constructional development

KARSTEN SCHMIDTKE*

Abstract

This paper provides a corpus-linguistic, usage-based approach to the acquisition of be-going-to-V *and* be-gonna-V. *Based on longitudinal data from two American children, it is argued that the constructions develop on the basis of several low-level chunks of varying degrees of morphosyntactic complexity. I propose an empirical way of grouping these chunks according to their structural and developmental properties, which allows us to trace how constructional networks emerge, expand and change in early childhood. In addition, this method reveals insights into the way the historically transmitted layering of the constructions is accessed in language acquisition. In particular, I uncover and account for apparent 'grammaticalization effects' in child speech, and discuss the relationship between acquisition and change in the cognitive-functional paradigm.*

Keywords: going-to-V, gonna-V, *mosaic development, Configurational Frequency Analysis, constructional networks, grammaticalization, constructional grounding.*

* I would like to thank Daniel Wiechmann, Holger Diessel and three anonymous reviewers for extremely helpful comments and suggestions on the paper. Special thanks go to Katja Hetterle for checking on coding reliability, and to Stefan Th. Gries for providing and discussing several algorithms used in this paper. The usual disclaimers apply. Author's correspondence address: Institut für Anglistik/Amerikanistik, Friedrich-Schiller-Universität Jena, 07743 Jena, Germany. E-mail: karsten.schmidtke@uni-jena.de.

1. Introduction

1.1 *Aim and scope of the paper*

Over the last decades, cognitive linguistics has established itself as a powerful framework for the synchronic analysis of grammatical structure and representation. In the form of the 'usage-based model' (Langacker 1987), it has also become a major paradigm for the study of language acquisition and change. The present paper aims to contribute to the recent fruitful dialogue between cognitive-linguistic and developmental research by investigating the acquisition of the closely related *going-to-V* and *gonna-V* constructions. These two grammatical patterns are particularly well-studied in the cognitive-functional paradigm and thus provide an ideal testing ground for usage-based hypotheses of language development.

More specifically, *going-to-V* and *gonna-V* have attracted many cognitive linguists' attention because of the psychological processes involved in their historical evolution.[1] As is well-known, the pattern started out as a biclausal structure involving the matrix verb *go* and a non-finite purpose clause, which became reanalysed as a monoclausal auxiliary-verb construction with future time reference. Thus *I am going* [*to get some wine*] was reconceived as *I* [*am going to*] *get some wine*. Crucially, this constructional shift involves a host of cognitive-functional operations in actual language use.

To start with, the semantic shift was enabled by the intrinsic future-orientation of purposeful action, which easily gives rise to a pragmatic (i.e. metonymic) inference from purpose to futurity. However, as Fischer (2007: 124) points out, the reanalysis also involves a formal change, and this is made possible by the prior existence of similar [*Aux-V*] constructions with a bare infinitive. Therefore, a critical prerequisite for the rise of the grammaticalized *going-to-V* pattern is the perception of a formal analogy. Analogical perception then continues to be at work in the consolidation of *be-going-to-V* as an auxiliary-verb complex. In this process, the motion component is eventually 'hyperanalyzed out' (Croft 2000) when the construction gradually diffuses through the lexical inventory of the language. Put differently, the pattern is analogically extended to non-motion and

[1] More accurately, psychological processes underlie the crucial first step in historical change, i.e. linguistic innovations that originate with individual speakers, and the successful propagation of a change *through the linguistic system*. The spread of a new variant *through the speech community*, i.e. the observable large-scale effect of a diachronic change, is driven by social rather than cognitive-functional factors (cf. Croft 2000: 4f.).

finally also to non-intentional contexts, as in *The bridge is going to collapse*. This process has also been referred to as 'rule generalization' or 'semantic bleaching'.

Finally, the generalization of the construction ensures its highly frequent use in communication. The cognitive effects of repetition on the shape and representation of grammatical constructions have been studied intensively (cf. Bybee 2006 for an overview). They chiefly involve automation and routinization (Bybee 2003, Haiman 1994), and in this case, they led to the morphophonological coalescence and erosion of *going-to* to *gonna* (Lehmann 1995).

In sum, the grammaticalization of *be-going-to-V* is fundamentally grounded in the cognitive construal and negotiation of form-meaning relationships in language use. It remains to be seen how these are negotiated in adult-child interactions, given that "the various stages of grammaticalization of *be going (to…)* coexist in Modern English" (Hopper and Traugott 2003: 3). It is notoriously difficult to determine how exactly these layers are represented in the minds of individual speakers (i.e. one versus several distinct constructions), but it is clear that children will need to develop an understanding of the formal and functional array that this interesting pattern exhibits in present-day English.

The usage-based model makes a number of principled predictions for how such grammatical constructions are acquired. In this paper, we will explore two aspects of the developmental trajectory of *going-to-V* and *gonna-V* that relate to these predictions. In §2.1, we will investigate the hypothesis that the acquisition of morphosyntactic complexity proceeds, not in an 'all-or-nothing' rule-based fashion, but in a piecemeal, 'mosaic' pattern. It will be argued that the grammatical components of the constructions, such as the inflected form of *be* and the particle *to*, develop differentially in several constructional environments. It will be shown that even at advanced stages of development, the application of grammatical 'rules' is far from consistent but rather item-specific. In §2.2, we will attempt to extract these items based on statistically significant co-occurrences of morphosyntactic and semantic features, and to arrange them in a representation not unlike the network structures posited in usage-based Construction Grammar (henceforth CG; cf. Goldberg 2003). §3 will look at the interplay of the various historical layers of the construction in ontogenetic development. Given that the usage-based model draws on a unified set of concepts to account for both acquisition and change, we will ask for commonalities between the two types of grammatical evolution. To what extent do children

'reproduce' diachronic processes? And if so, how can we account for such effects?

Across all analyses, due attention will be paid to yet another important prediction of the usage-based model, i.e. the existence of profound inter-individual differences in language learning (cf. Richards 1990). While generative approaches still emphasize the 'striking' similarities in children's development of grammatical competence, invoking them in support of UG (cf. Dąbrowska 2004 for a critical discussion), the usage-based model would actually predict heterogeneous pathways of development, depending on individual communicative preferences and the nature of the input.

The methodological approach taken to these questions is that of quantitative corpus linguistics. Contemporary corpus-based research, with its emphasis on rigorous statistical evaluation (cf. Gries 2006), is well-suited to empirically support a theoretical approach in which frequency distributions of grammatical forms play a central role. Therefore, a major focus of the present paper will lie on the application of recent statistical techniques to developmental corpus data. The following section will provide a more detailed description of the database and its coding.

1.2 *Sampling and coding of the data*

The study is based on observational data from two monolingual American children in the CHILDES archive. I selected Adam and Sarah from the BROWN corpus because they were recorded for an exceptionally long and comparable period (Brown 1973). Four corpora, for each child and its respective input, were compiled by retrieving all instances of the *going-to-V* and *gonna-V* construction(s). Retrieval was based on a manual scanning of all transcripts, so that all occurrences could be analysed within their respective discourse context. I discarded all direct imitations and self-repetitions, but counted persistent uses of the constructions across successive discourse turns if they included novel features, i.e. a different main verb, no *to* although it was present in the previous turn etc. In collecting the data, I had to fully rely on the established transcription in the CHILDES database. I realise that it may have been problematic for the original recorders to reliably distinguish between, say, *goin(g)* and *go and*, but this an inherent risk of many corpus-based enterprises dealing with spontaneous spoken language. On the basis of all these criteria, I arrived at the database summarised in Table 1.

Table 1. *Overview of the database*[2]

	Age range	Token frequencies of the construction		
		going-to-V	*gonna-V*	Total
Adam (child)	2;3.04 - 5;2.12	898	628	1,526
Adam (input)		389	15	404
Sarah (child)	2;6.20 - 5;1.06	254	190	444
Sarah (input)		362	561	923
Total		1,903	1,394	3,297

The data were then coded for a number of morphosyntactic, semantic and discourse context-related variables. Grammatical features comprise the subject, the presence and – where applicable – the contraction of a form of *be*, the specific form of *go* (ranging from *go(n)* over *goin(g)* to *gonna*), the presence of *to*, the larger constructional context or sentence type (distinguishing simple declarative, interrogative and imperative as well as subordinate clauses), and the main verb. From the discourse context, I determined whether the use of the construction is self-initiated by the respective speaker (as opposed to being used in a reaction to the interlocutor's previous use of it). Additionally, I ascertained whether the construction is used in a situational context of motion, so as to distinguish literal motion-cum-purpose *going-to-V* clauses from their grammaticalized ('bleached') counterparts. The former were coded as 'ambiguous', the latter as 'auxiliaries'. It is, of course, a difficult task to decide faithfully in each case simply on the basis of text sources. In many cases, however, the 'action' tier in the CHILDES archive provided valuable clues, and the main verb also helped to clearly identify non-literal cases.

To ensure reliability of coding, a subset of the database comprising 400 randomly selected utterances was double-checked, achieving an inter-coder reliability score of 98.3 percent. We can now proceed to the qualitative and quantitative analysis of the database from a usage-based perspective.

[2] Note that the different corpus sizes are not hazardous since all analyses were performed separately for each subcorpus, and the statistical tests are designed in such a way that they take corpus sizes into account.

2. 'Mosaic' development of grammatical representations

A number of recent studies have provided detailed evidence that the formalist assumption of 'across-the-board' acquisition of grammatical rules (cf. Chomsky 1999) is problematic (e.g. Kuczaj and Maratsos 1983; Rowland 2007). An alternative account holds that children develop low-level schemas in which specific lexical and grammatical material is assembled (e.g. *Where's Mummy?*), and only gradually categorize similar form-meaning pairings into a more general constructional representation (e.g. *wh*-questions with inversion, or a syntactic category AUX). Crucially, this often involves the differential production of grammatical forms: At a given time, only a few auxiliaries may be found to occur in progressive constructions, undergo inversion, or appear in negative tag questions; the general 'rule' only emerges after a good amount of exposure to the wide range of auxiliaries in each of these constructions.

In a pilot study to this paper (Schmidtke 2007), I suggested that the grammatical components of the *going-to-V* and *gonna-V* constructions also undergo such piecemeal development. Qualitative evidence for this hypothesis comes from radically differing instantiations of the constructions at roughly the same age:

(1) *CHI: *gon fall.* (Adam 3;10.15)
(2) *CHI: *Are they going to get in there?* (Adam 3;11.01)

The question is, of course, how robust such differences turn out to be from a quantitative perspective on the database, and what mechanisms can be held responsible for their occurrence.

2.1 *Evidence and accounts of mosaic production*

One possibility is that the larger constructional context (i.e. sentence type) determines the degree of morphosyntactic complexity in such a way that simple declaratives typically come without the obligatory grammatical components, while more complex constructions, especially questions and embedded sentences, favour their overt presence. A relatively straightforward procedure to test this claim would be to assign each instance of *going-to* and *gonna* a degree of morphosyntactic complexity: 'full' instantiations of *going-to-V* phrases, for instance, show an overt form of *be* and an overt *to*, while 'semi'-reduced tokens lack one of the two features and 'reduced' ones lack both. When we cross-tabulate this degree of complexity with the variable 'sentence type', we should be able to assess their statistical association.

With regard to Adam's production (n = 1,526), a global test for full versus reduced syntax in declarative versus non-declarative clauses yields significant results, indeed ($\chi^2 = 7.629$, $df = 1$, $p < 0.01$)[3], as does a more fine-grained test that distinguishes three levels of complexity ('full, semi, reduced') and all four sentence types ($\chi^2 = 42.77$, $df = 6$, $p < 0.001$). This analysis is, however, not particularly satisfying, for two reasons. First and more importantly, it achieves only very low effect sizes ($\varphi = 0.071$ and Cramer's $V = 0.118$, respectively), suggesting that the association may have reached the significance criterion only due to the comparatively large corpus size. Second, although we find strikingly similar examples in Sarah's speech (n = 444), the tests performed on her data do not yield any significant results.

Therefore, morphosyntactic reduction may not be accurately captured as a function of sentence types at large. An alternative that ties in nicely with many versions of CG is to conceive of the [SUBJ *BE GOING TO* V] construction as itself being composed of multiple subconstructions or chunks that language learners gradually blend together in what is called 'symbolic integration' (Langacker 1987). On this 'cut-and-paste' account (Tomasello 2003: 305), a possible scenario is that the production of inflected *be* develops within the Progressive [SUBJ *BE GOING*] chunk, while the insertion of *to* may follow its own particular route of integration into this construction. Since *to*-omission is being studied intensively in a current research project (cf. Kirjavainen et al. in press), we will focus on the former aspect here.

Theakston and her co-authors (2005) proposed that the development of auxiliary forms is contingent on the subject of the construction in so far as children seem to "acquire specific subject + auxiliary combinations" (255) as fully or partially lexicalised units. Especially those combinations that are modelled frequently in the ambient language quickly become routinized units in the children's production, whereas less frequently occurring subjects and subject-auxiliary combinations in the input lead to a more differential provision of the auxiliary in the children's speech.[4] In more statistical terms, this account predicts that in the children's data,

---

[3] All statistical analyses computed for this paper were carried out in the open-source software *R*, version 2.6.1 (R Development Core Team 2007).

[4] Theakston et al. (2005: 269) acknowledge that input frequencies alone cannot fully account for the data, especially with regard to the pronominal subjects *I* and *you*. But the analyses demonstrate that they seem to play a major role in determining the development of auxiliary provision.

the provision of the auxiliary in the *going-to-V* and *gonna-V* constructions is significantly biased towards particular kinds of subjects. In order to test this claim, I categorized the subjects in the corpora into similar sets as those used by Theakston et al. (2005): pronominal *I*, *you*, *(s)he*, *it*, *we*, *they*, proper names and all lexical NP subjects (including indefinite pronouns). An additional category is formed by null-instantiated (i.e. omitted) subjects (Ø).

Starting with Adam's corpus, a Chi-squared test reveals indeed a highly significant skewing in the distribution of *be*-provision across different subjects, this time also achieving a more respectable effect size ($\chi^2 = 330$, $df = 9$, $p < 0.001$, $\varphi = 0.465$). Graphic tools such as extended mosaic plots (Friendly 1994) allow for a closer inspection of how the individual cells of our contingency table contribute to this overall skewing. Fig.1 shows, on the horizontal dimension, the proportions to which each subject type is prone to either auxiliary omission (left rectangles) or auxiliary provision (right rectangles). On the vertical dimension, the width of the rectangles reflects the total frequency of each subject type in the corpus.

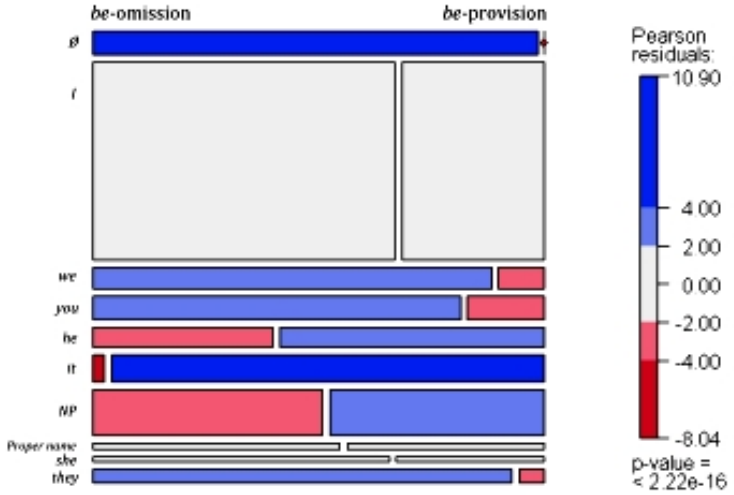

Figure 1. *Mosaic plot of Adam's differential* be*-insertion*

Importantly, the shading of the boxes indicates how dramatically the observed frequency of a particular subject-auxiliary association deviates from its expected frequency in the sample. The higher the difference (i.e. the absolute value of the Pearson residual), the darker the shading in the mosaic plot. We can see that the most significant deviations, i.e. the darkest areas, are found with null-instantiated

subjects (*be*-omission) and *it* (*be*-provision), followed by *we, you* and *they* on the omission side, and *he* and NP subjects on the insertion side.

Since auxiliary omission is ungrammatical in adult language, we would not expect to find a similarly uneven distribution in Adam's input data. Indeed, the test yields a significant result only because Adam's caregivers consistently drop the form of *be* when they also drop the subject:

(3) *MOT: *going to give it to Robin?* (Adam 3;0.11)

Such deprived productions typically occur in response to an equally reduced construction in Adam's speech. If these idiosyncratic cases are factored out, the result is far from skewed ($\chi^2 = 3.579$, $df = 8$, $p = 0.893$). Therefore, Adam's morphosyntactic production is significantly different from that in his ambient language, and the pronounced skewings in his data reflect specific subject-auxiliary associations rather than a single pattern or rule of auxiliary provision.

For Sarah, the overall association of subject type and *be*-insertion is also highly significant ($\chi^2 = 71$, $df = 9$, $p < 0.001$, $\varphi = 0.4$), with strongly positive residuals for *it* and NPs contributing firmly to the skewing. Even more pronounced is the association of subject drop (i.e. null-instantiated subjects) and *be*-omission, just as in Adam's case. This last pattern is again also mirrored in the input, but in contrast to Adam, Sarah's caregivers also show a tendency for *be*-omission with second-person subjects (e.g. *you going*), which introduces a significant skewing to the input data ($\chi^2 = 87$, $df = 8$, $p < 0.001$, $\varphi = 0.313$). The ratio of *be*-omission in *you*-contexts is actually quite similar across input and output (129:362 for caregivers, 9:20 for Sarah, $\chi^2 = 0.12$, $df = 1$, $p = 0.728$). Therefore, Sarah's morphosyntactic production, while still showing a pronounced *it*/NP + auxiliary association, more closely corresponds to the ambient language.

In addition to these significant findings across the entirety of each corpus, a convincing argument for mosaic development could be made if it were demonstrated that some of the reduced forms outlast the stage at which most of the [SUBJ *BE GOING TO/GONNA*] tokens are fully instantiated. Most recent corpus linguistic research has provided a method for tackling such questions, namely in the form of *dispersion* measures (Gries 2008a). In Adam's data, we find indeed that some of the reduced chunks are of particular longevity and not limited to specific parts of the corpus, i.e. particular ages. For example, the reduced *I going/gonna* and *you going/gonna* chunks persist all the way

up to 4;10.02 and 5;2.12, respectively, and although some stages do contribute more than others to the overall frequency of these combinations, the distribution is relatively even across the corpus and conforms to the frequencies one would expect given the respective file sizes. This is reflected in statistical measures of dispersion, e.g. for *you going/gonna* (Juilland et al.'s (1970) D_{adj} = 0.82, Gries' (2008a) *DP* = 0.36)[5]. Conversely, when the construction features *it* as its subject, it is accompanied by *is* from the earliest occurrence onwards at 2;10.30 (file 17/55), with a similarly homogeneous dispersion across the corpus (D_{adj} = 0.85, *DP* = 0.30).

In Sarah's data, we find comparable tendencies: while individual *I*-subjects appear without *am* up to 5;1.06 (file 139/139, D_{adj} = 0.87, *DP* = 0.49), *it*-subjects appear consistently with *is* from 3;5.20 (file 63/139) onwards, albeit with a heterogeneous dispersion across the relevant files (D_{adj} = 0.71, *DP* = 0.74).

Taken together, the data in this section appear to provide empirical evidence for relatively conservative learning which is not based on the consistent application of morphosyntactic rules, but rather on the persistent production of specific chunks. If this holds true at the level of a constructional subpart of *be-going-to-V* and *be-gonna-V*, it would not be surprising if the development of the construction(s) as a whole were also based on the gradual generalization over lower-level chunks. We will now proceed to a statistical method that proves able to extract these developmentally relevant chunks from the dataset.

---

[5] The dispersion measures were computed by using Gries' (2008a) `dispersions2` algorithm in *R*, which is made available on <http://www.linguistics.ucsb.edu/faculty/stgries/research/dispersion/links.html>. The calculation proceeds from all instances of the form we are interested in and the subparts (i.e. files) of the corpus in which they appear. In the present analyses, the corpus files correspond to the CHILDES files (e.g. Adam01-55), which reflect successive stages of age from 2;3-5;2. Since the corpus files are of unequal size, the adjusted version of Juilland et al.'s *D* was chosen. Its values technically range from 0 to 1, with higher values indicating a rather even distribution. (For a standard of comparison, consider, for example, that a very common and general lexical item such as *lively* scored a *D* of 0.92 in Leech et al.'s (2001) corpus study, while a much more specialized lexeme such as *HIV* ranks significantly lower (*D* = 0.56) because it is dispersed across far fewer files in the corpus.) Gries' *DP* (deviation of proportions) is somewhat more robust and based on a comparison of expected and observed relative frequencies across different corpus parts. In contrast to D_{adj}, it is lower values of *DP* that indicate a rather even (in the sense of 'expected in view of the sizes of each subpart of the corpus') distribution. For example, Gries (2008a) reports a *DP = 0.223* for the auxiliary *are* in the BNC to be an indicator of even/widespread dispersion, while the *DP = 0.989* for the highly specific lexeme *scallop* reflects its extreme underdispersion in the corpus.

2.2 *Determining statistically significant production types*

The aim of this section is to introduce a multivariate statistical application that allows us to determine in a more principled way which low-level chunks of *going-to-V/gonna-V* occur with significantly more than chance frequency in children's production and hence may be taken to be the most entrenched exemplars that are, ultimately, integrated into one or more fully schematic constructional schemas (cf. Abbot-Smith and Tomasello 2006). Since each of those low-level chunks is characterised by a specific combination of several structural features, we need to extend our previous Chi-squared analysis to a truly multivariate dataset, i.e. one that exceeds the common two-dimensional design. A well-suited method for this problem is exploratory *Configurational Frequency Analysis* (*CFA*; von Eye 1990). In this case, it is applied to all *going-to-V* and *gonna-V* instances in each child's corpus, assessing simultaneously[6] the statistical independence of the following variables of the pattern:

- subject choice (a factor distinguishing all ten subject types from above)
- presence and contraction of the auxiliary *be* (a factor with three levels: no auxiliary, contracted, full form)
- the specific form of *go* (a factor distinguishing four phonetic realisations, *go, gon, going, gonna*)
- provision of *to* (a factor with three levels: presence, absence, and non-applicability (in *gonna*-clauses))
- sentence type (a factor distinguishing declarative, interrogative, imperative, and subordinate clauses)
- constructional meaning (a factor distinguishing ambiguous/motion from grammaticalized auxiliary function).

The computation was performed by the `hcfa()` function in *R*, kindly provided by Stefan Th. Gries (cf. also Gries 2008b). Due to space limitations, we will illustrate and discuss its results for Adam's data only, but conclude with a brief outlook on Sarah's corpus.

2.2.1 *Methodological procedure* The *CFA* essentially computes all possible combinations of our factor levels and, just like a Chi-squared analysis, compares the observed and expected frequency of each of those combinations in our corpus. The difference is evaluated

[6] Since in this paper, *CFA* is applied exploratively, the null hypothesis is that of complete independence of all variables. As we are interested in extracting significant exemplars of the constructions in their entirety, no variables were excluded (as would be the case in hierarchical *CFA*s).

by a Chi-square statistic, and the associated *p*-value is adjusted to the fact that multiple comparisons are made on the same cells. Table 2 provides a brief excerpt from the output of the *CFA* containing those measures:

Table 2. *Excerpt from the* CFA *output*

Subj	Con-tract	Form of *go*	*to*	Sent. type	Sem	Freq	Exp	Cont. chisq	Obs-exp	P.adj.Holm	Dec	Q
I	n.a.	going	no	decl	aux	295	88,1091	485,805	>	1,72E-70	***	0,144
I	yes	gonna	to	decl	aux	0	14,3852	14,3852	<	0,00150643	**	0,01
they	n.a.	going	to	decl	aux	13	2,7266	38,7086	>	0,01647634	*	0,007
Ø	n.a.	going	n.a.	decl	aux	0	10,8076	10,8076	<	0,05534763	ms	0,007
we	n.a.	gonna	n.a.	quest	aux	6	0,8717	30,1703	>	0,81460591	ns	0,003
NP	no	gonna	n.a.	quest	aux	3	0,1254	65,8957	>	0,84359726	ns	0,002

Each row in this table represents one particular combination of the variables in columns 1-6. We can then glean the observed and expected frequencies of this combination, its Chi-squared and *p*-value and an indication of its significance level (e.g. *** $p < 0.001$). The final column of the table lists the effect size for each combination, which is called the coefficient of pronouncedness (*Q*) in *CFA*. Textbooks often remain fairly reticent as far as the interpretation of *Q* is concerned, which is particularly unfortunate given that its values are commonly far lower than comparable effect size measures in Chi-squared and related analyses. However, we will try to flesh out this measure in a more meaningful way below. The final information that the table provides us with is whether each combination is significantly more or less frequent than expected under the null hypothesis of independence. Rows marked '>' exceed their expected frequency and are called 'types'; rows with an '<' are called 'antitypes'. For this paper, we are interested only in 'types' since these are taken to reflect particularly entrenched or routinized chunks of language. In our particular case of Adam's *going-to* and *gonna* constructions, the computation yielded 28 significant types, two of which were subsequently discarded because they occurred with a negligible token frequency ($n < 5$).

Crucially, each of the remaining 26 types does not only have particular structural properties, but also developmental characteristics, i.e. a certain distributional profile across the corpus. Two types of information about this profile are particularly important for our analysis. Firstly, I identified the age range for every type, excluding earliest outliers, and transformed it into a numerical value; this piece of information basically indicates how persistent or how short-lived a particular type is. In addition, it also of course indicates the age of

emergence of each type in the data. From manual inspection of the transcripts and the corpus data, there appeared to be a qualitative breaking point at around 3;0 years of age, in the sense that quite a few *going-to* types were robustly used before 3;0, while others (notably more complex ones) only began to appear well after 3;0. To capture some of this developmental dimension, we will introduce a categorical distinction between 'early' types, i.e. those in place before 3;0, and 'emergent types' after 3;0.[7]

Secondly, I considered the overall token frequency of each type since some of the chunks occur over quite a long period of time, but with a very low frequency (and vice versa). If we combine range of occurrence and token frequency by multiplication, their mathematical product should give us an indication of how 'prominently' represented a type is in our corpus: The most prominent types are persistently frequent, whereas less prominent ones might be frequent but short-lived, or infrequent throughout. Interestingly, if we calculate this product for each and every type, we end up with values that correspond almost perfectly to the coefficient of pronouncedness *Q* in the output of the *CFA* (Pearson's $r = 0.99$, $p < 0.001$). This suggests that *Q* can be interpreted more meaningfully as reflecting the degree of entrenchment of a particular type. The theoretical premise here is, of course, that frequent persistent production reflects the facile activation of a linguistic pattern, which in turn is a function of its cognitive entrenchment (cf. also Wiechmann 2008 in relation to *CFA*).

To sum up thus far, we have extracted from Adam's corpus 26 significant production types of *going-to-V* and *gonna-V*, each of which has their own structural and semantic properties (cf. our six variables), and certain developmental characteristics, i.e. a time of emergence (cf. the two-way early/emergent distinction) and a degree of entrenchment or prominence in the corpus, expressed numerically by *Q*. We now need to present the types in an interpretable way. In many versions of CG, formally and functionally related constructions (of whatever degree of abstraction) are thought to be organized in a structured inventory or network (e.g. Croft 2001: 25). In language acquisition, then, such networks are argued to be built up from generalizations across concrete, similar exemplars of constructions. Exemplars are

[7] Needless to say, this is an arbitrary cut-off point, and it would be much more desirable to capture qualitative changes in the database with an objective bottom-up measure. The recently-established method of 'variability-based neighbour clustering' (cf. Gries and Stoll in press) is available for clustering CHILDES files, for example, on the basis of MLU scores. Its application to multivariate data has not been tested so far and would require a much larger database than the present one (Stefan Th. Gries, p.c.).

first organised in a more general way when they come to consist of a recurrent pivotal element, e.g. a lexicalised chunk such as those extracted by the *CFA*, and a more variable slot, e.g. the main verb of the *going-to*/*gonna* construction. These chunks, in turn, are grouped according to formal and functional similarities, and may form a cluster or branch in the growing constructional network. In the following, we will thus try to represent the 26 *CFA*-output types in a network-like display that reflects their structural similarities and at the same time captures developmental information. After all, some of the chunks extracted are very short-lived and can thus only be 'interim branches', as it were, of the overall network.

A useful multivariate method for grouping units according to their properties is the family of clustering methods (cf. Baayen 2008: ch.5.1). For this procedure, we need to think of our production types as vectors that are characterised along the six structural variables used in the *CFA* and along the two developmental parameters (*Q*, 'early' versus 'emergent' type), i.e. as vectors in eight-dimensional space. It then becomes possible to calculate the relative similarity of all vectors to one another by applying an appropriate (dis)similarity measure like Kendall's τ, and to display them as clusters of similar vectors by choosing an appropriate linking procedure. The so-called 'neighbour-joining algorithm' (cf. Saitou and Nei 1987) allows producing an unrooted, network-like cluster solution, which is shown in Figure 2.[8]

[8] The clustering procedure forces us to think of our vectors in spatial terms: their relative (dis)similarity is captured as a spatial distance between them. However, since we are not interested in their absolute distance, but rather in differences in their slope, i.e. the vector's curvature created by different values on each of our eight variables, a correlational measure was preferred over an actual distance measure. Furthermore, since our data are largely non-parametric, Kendall's τ is a suitable dissimilarity measure here. In order to calculate the distance matrix, factor levels were transformed into sensible numerical values (e.g. contraction: 'n.a.' = no auxiliary = 0, 'yes' = 0.5, 'no' (full form) = 1). Finally, all values v of each of the eight characterising parameters P were normalised so that the P with the wider range of values does not end up dominating the results of the clustering computation. Normalisation was achieved by calculating, for each v, $v\text{-}min(P)/max(P)\text{-}min(P)$ (cf. Cysouw 2007: 71). The neighbour-joining algorithm is implemented in the `ape` package in *R* (Paradis 2006) and produces a clustering solution comparable to single- (i.e. nearest-neighbour) linkage in the canonical `hclust` function.

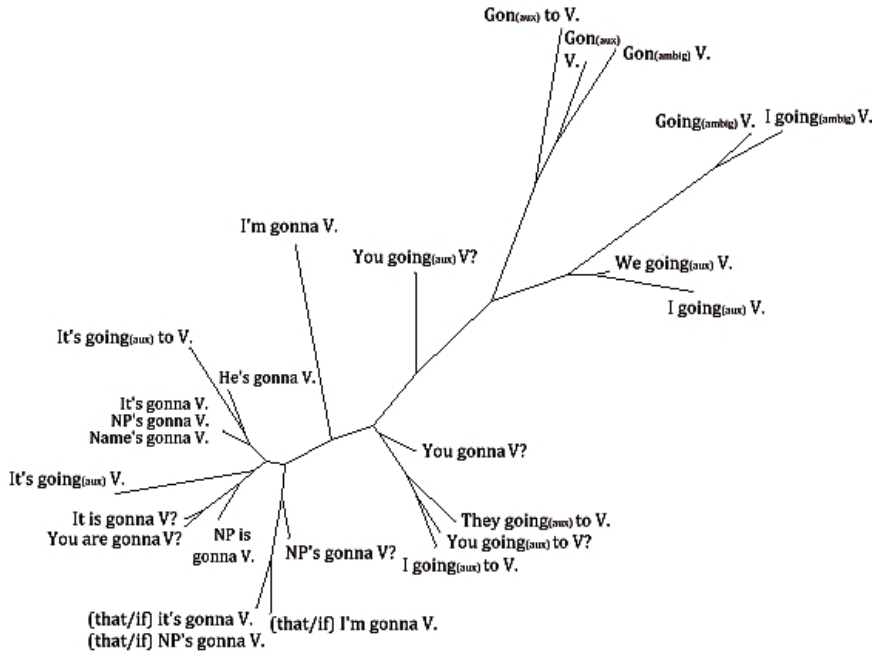

Figure 2. *Unrooted arrangement of Adam's production types (clustered according to both structural and developmental properties)*

2.2.2 *Discussion* From a bird's-eye view, this network displays significant types of the *going-to-V/gonna-V* patterns in Adam's production. It is plausible to assume that these represent the first low-level abstraction across concrete exemplars of the respective constructions and, therefore, an intermediate stage on the way to developing a schematic and fully productive pattern of each construction. Along structural lines, the cluster solution captures our previous findings: the strongest bifurcation in the tree splits reduced *Ø*- and *I*-subject chunks to the upper right from chiefly third-person chunks with full-fledged morphosyntax to the lower left. This fundamental two-way split of the clustering solution does, however, also reflect the temporal information that entered the algorithm: as we will see, it essentially severs the earliest types from the ones emerging later (i.e. around or after 3;0). We will thus start the discussion at the upper part of the tree and move downwards bit by bit in order to trace how the network might be built up and shaped over time.

The network is initiated by the top branch on the upper right. The earliest type to emerge is *Gon$_{(aux)}$ to V.* (2;3)[9], which opens up a small cluster of very reduced forms in either literal or auxiliary function. Interestingly, although the three types in this cluster are all morphosyntactically deprived, they persist in Adam's production well into his fifth year of life (average age of disappearance = 4;3.21), albeit with low frequency. This branch, then, seems to be quite robust throughout development. To its right, the next cluster emerges, consisting of the first *going* forms, which itself comprises two rather distant subclusters. The *going* forms in the long upper branch are both semantically ambiguous (i.e. are likely to reflect actual motion), and while the subjectless one drops out after only three months, its counterpart with *I* persists in the corpus till 4;2.17. Importantly, these literal chunks precede in their emergence the second subcluster, *We going V.* and *I going V.* in auxiliary function; we will return to this apparent 'grammaticalization' of child speech below. At this point, it needs to be stressed that both types are very robust in the corpus and *I going$_{(aux)}$ V.* is additionally the most highly entrenched pattern of all (n = 295, Q = 0.144), i.e. this single branch is likely to be sustained as a highly accessible and hence distinct node of the network.

We now cross the 3;0-borderline and move on to 'emergent' types. It is worth observing that from now on, the established *going*-patterns become more complex, either gaining an inflectional form of *be* or *to* or appearing in a non-declarative construction. In fact, the *You going V?* distinctly branches off next at 3;0 and constitutes a noticeable deviation from the previously prevailing first-person subject choice, as well as a new constructional environment (i.e. sentence type). Similarly, the new branch at the lower middle now shows *to*-insertion with *I, we* and *they* subjects. These three types emerge and disappear at roughly the same time (3;0-4;1), thus constituting temporary branches in the network. Finally, and also at around 3;0, the *going* construction is extended to inanimate subjects (*It's going V.* and *It's going to V.*). The two chunks compete in production until the latter wins out and prevails from 4;2 onwards. Notice, however, that these two chunks feature the auxiliary *'s* and thus closely resemble almost all *gonna*-types in form and distributional properties (especially range

[9] All *specific* temporal information given in the subsequent analyses cannot, of course, be read off the graph, but was gleaned from (i) examining in the corpus the precise age range in which a given chunk occurs, as outlined previously, and (ii) from investigating the specific frequency distribution of each chunk across this age range, i.e. whether, for example, a chunk has a slow start but becomes considerably more frequent in later files, or vice versa.

of occurrence). Therefore, they are to be found in the final major, *gonna*-dominated cluster at the left side of the network.

This large cluster reflects a decisive qualitative change in Adam's production (*go* > *gonna*). *Gonna* makes its way into the network at around 3;0 in the dense cluster of abstract third-person subjects (*it, NP*), with which it is attested all the way up to the final recording. Interestingly, *he* and *name* join this branch only much later (3;9), and it also takes more than nine months for Adam to extend *gonna* to *I*. Thus although *I'm gonna V.* is the second-most entrenched pattern overall ($n = 197$, $Q = 0.114$), securing it a distinct major branch in the middle of the network, its distribution across the corpus is peculiar because it only gathers momentum well after a 'critical mass' of other *gonna*-models is established. Ruhland and his colleagues (1995: 116) describe this scenario as a 'precursor relation' in the acquisition of formally similar constructions. Perhaps surprisingly for contructivists, the comparatively late emergence of *I'm gonna V.* is *not* due to the fact that the *gonna*-precursors are more frequent in the ambient language. At least in the maternal input recorded here, *gonna* does not surface before 3;7 at all. An alternative explanation could be that the existence of the well-entrenched *I going (to) V.* preempts the consolidation of a synonymous construction type. In other words, constructional competition may be at work (cf. Bates and MacWhinney 1989).

The minor branch of *gonna*-types with fully articulate *be*-forms (*It is gonna V?*, *You are gonna V?*, *NP is gonna V.*) testifies once more to the fact that the auxiliary is more likely to be spelled out completely in third-person or non-declarative contexts. Finally, the latest types to emerge involve both the overall shift to *gonna* and an increased level of grammatical complexity: these are the subordinate clauses at the very bottom of the cluster, with an average age emergence at 4;3.

In sum, a multivariate statistical approach to our data has uncovered entrenched lower-level grammatical schemas and suggested an objective, bottom-up account of how an early constructional network may grow and differentiate. Of course, we are still left in the dark with respect to the time at which Adam proceeds to a yet more schematic representation of the two constructions, or even conflates the *going-to* and *gonna* patterns into a single, multilayered unit. Before we look at the development of such layering more closely, let me briefly point out that Sarah's *CFA* types and their corresponding network are considerably less substantial due to the more limited size of her corpus. The *CFA* yielded eight significant types, only four of which come with a respectable overall frequency

(*I'm gonna V.*, *I going$_{(ambig)}$ V.*, *I'm going$_{(aux)}$ V.* and *I'm going$_{(aux)}$ to V.*). Again we find that the ambiguous *going-to*-phrases precede the grammaticalized ones in the corpus, but at the same time, we witness a much more rapid development of *gonna* when compared to Adam. It is this interplay of constructions that will concern us in the next section.

3. 'Grammaticalization' of child speech?

In the growth of the constructional network, we already encountered a potential 'grammaticalization' of child speech in so far as that the children's earlier uses of the *going-to-V* construction resemble its diachronic source in being used literally as a motion-cum-purpose clause, while later uses of this pattern are significantly less often ambiguous between this literal and the grammaticalized immanent-intentional future reading. Indeed, we find examples such as the following accumulated in the children's early production of the construction:

(4) *CHI: *going cut* [/] *cut a (to)mato juice.*
%act: <aft> *went to trash can and cut into it*
(Adam 2;7.01)

(5) %act: *goes in kitchen to wash towel*
*CHI: *going wash a hands.* (Adam 2;8.01)

(6) *CHI: *I want go wash dis kids do it.* […]
*CHI: *I goin(g) wash em .* (Sarah 3;0.18)

In these and similar cases, the construction is used with agentive, intentional subjects, notably the (implicit) first person singular pronoun, and in a conversational context that involves directed motion, as far as this could be retrieved from the transcripts. In other words, the children themselves or other agents in the scene are literally moving somewhere in order to achieve their purpose in mind. At the same time, however, such contexts invite a reinterpretation of *going to* as a marker of immanent-intentional futurity. This ambiguous constellation is characteristic of the source construction in diachronic change, and it is only when the pattern is extended to non-motion contexts and, ultimately, to third-person inanimate subjects that we can actually detect a reanalysis of the construction. This 'bleaching' or functional shift also appears to be operative in child language. Thus the children soon come to use *going-to-V* in metaphorical, i.e. grammaticalized contexts:

(7) *CHI: *This going be a dog?* (Sarah 3;4.09)
(8) *CHI: *It's going to fall.* (Sarah 3;5.20)

It is important, however, to submit the 'grammaticalization hypothesis' to more rigorous statistical analysis.

3.1 *Data and Analysis*

In this section, we will examine the subcorpus of the data that contains all (and only) *going-to-V* instances. One reviewer of the pilot study pointed out that the grammaticalization hypothesis would only be compelling if it were shown that the construction is more likely to be ambiguous in child speech than in the input, or more likely in younger children's speech than in older children's speech. In order to pursue this argument, let us first take a global perspective. It can easily be demonstrated that in Adam's production, the ratio of ambiguous and grammaticalized uses (149/758) is significantly higher than that of his input (39/351), albeit with a low effect size ($\chi^2 = 8.58$, $df = 1$, $p < 0.01$, $\varphi = 0.084$). The same is true for Sarah, but in a somewhat more pronounced way ($\chi^2 = 19.28$, $df = 1$, $p < 0.01$, $\varphi = 0.179$). In both cases, the residuals point out more ambiguous cases than expected for the children, and at the same time less of these cases than would be statistically expected in the caregivers' speech. Overall, then, the construction is more likely to be ambiguous in child speech than in the input, but that by itself is not sufficient to support the grammaticalization hypothesis. The significantly higher amount of literal *going-to* constructions may simply reflect the children's mobility during the recordings, e.g. their frequent change of locations in order to bring something to the scene. Therefore, we have to probe into developmental patterns again.

On this more fine-grained level, our hypothesis suggests that the frequency distribution of ambiguous cases is biased towards early stages of development, while in later periods the grammaticalized uses of the construction take over and dominate the children's production; for the input, no particular distribution is expected *a priori*. This intuition can be tested by examining the rate of ambiguous and grammaticalized uses over time. Fig. 3 plots the corresponding frequency distributions for Sarah.

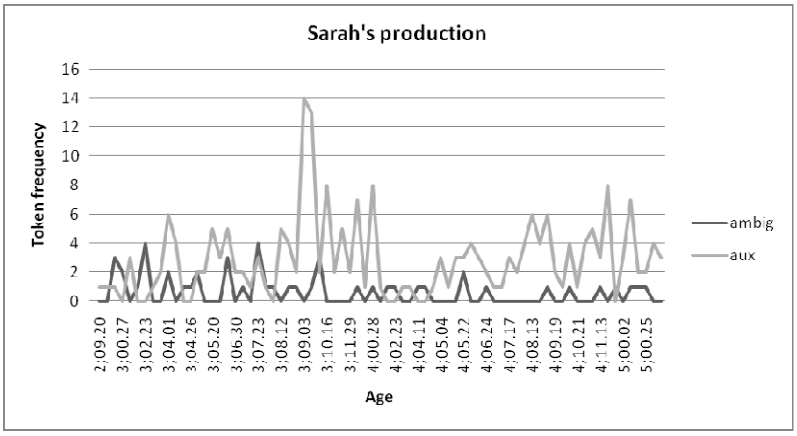

Figure 3. *Frequencies of ambiguous and grammaticalized* going-to-V *constructions in Sarah's production over time*

We can see that the ambiguous cases outnumber the grammaticalized ones only during the first months in which the construction is used. The construction then relatively quickly becomes more versatile, being readily applied to metaphorical contexts. A very similar picture arises from Adam's data. Needless to say, a much denser corpus, such as the daily recordings of Leo used by Abbot-Smith and Behrens (2006), would presumably bring out the pattern more clearly, but even in our more coarse-grained data, there seems to be some empirical support for the grammaticalization hypothesis.

A further apparent parallel between historical and ontogenetic processes can be found in Adam's long-term development of the *going-to-V* and *gonna-V* constructions. We saw above that in diachronic change, the form *gonna* arose via routinization and automation of the frequently used *going-to* chunk. In Adam's speech, too, *gonna* develops only after the emergence and consolidation of *going-to*, and there is a notable shift in preference of the two constructions, as displayed by Fig. 4. More specifically, from 3;11.01 (file 40) onwards, the scores for *gonna* are significantly higher than those for *going-to*, which is indicated by a non-parametric comparison of the two patterns (two-sample *U*-test for independent observations, $W = 14.5$, $p < 0.001$). Crucially, this development is independent of the input in the recordings since *gonna* has a very low token frequency throughout in the caregiver's speech (cf. Fig. 5).

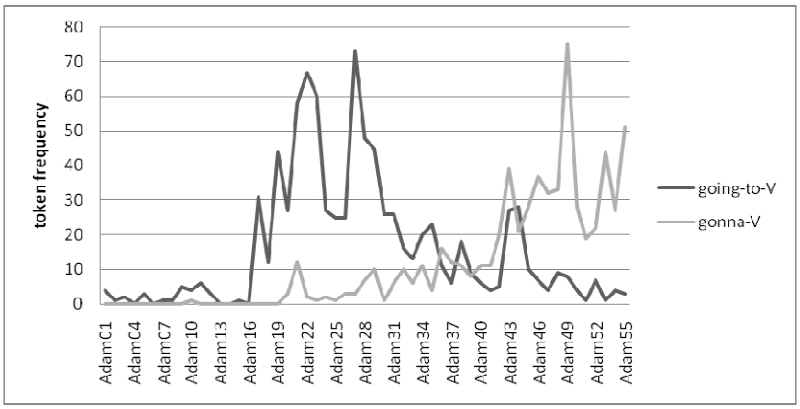

Figure 4. *Frequency distribution of* going-to *and* gonna *in Adam's production over time*

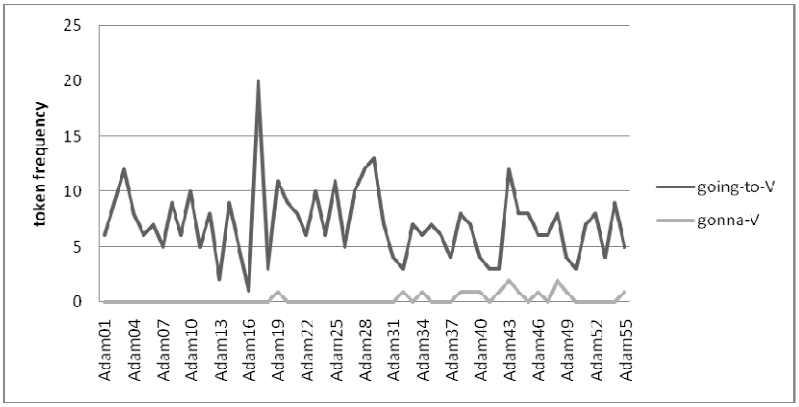

Figure 5. *Frequency distribution of* going-to *and* gonna *in Adam's input over time*

In fact, Adam is frequently corrected when he uses *gonna*:

(9) *CHI: *it's gonna break.*
*MOT: *it's going to break.* (Adam 3;0.11)

Moreover, 8 of the 15 *gonna* tokens in the input occur as direct reactions to Adam's use of the construction, only 7 are initiated by the caregivers. Therefore, Adam's predilection for *gonna* at later stages of the recordings is due either to input that is not recorded (e.g. by other family members) or to a genuine shift in constructional preference, which may in turn be caused by routinization of the *going-to-V* pattern, not unlike in historical grammaticalization.

Sarah's data, however, clearly show that children do not reproduce grammaticalization pathways perfectly, but instead draw heavily on the available input. In her data, uses of *gonna* are attested quite early, along with and sometimes even superseding *going-to* (Fig. 6).

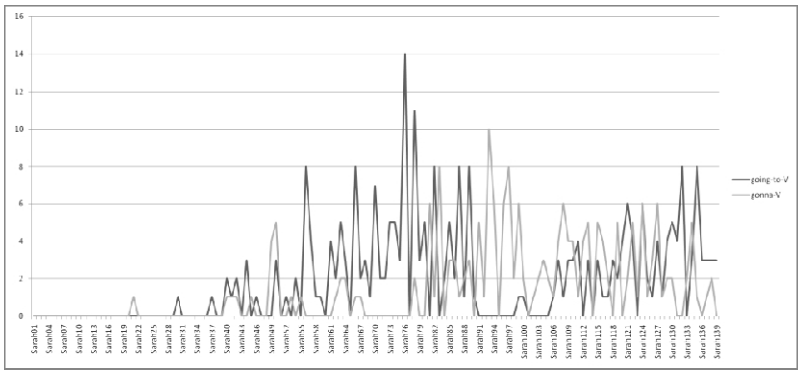

Fig. 6. *Frequency distribution of* going-to *and* gonna *in Sarah's production over time*

The comparatively early use of *gonna* is likely to be due to her input, which is replete with *gonna*-phrases, especially in the initial period of the recordings. It seems, however, that Sarah, too, shifts preferences from *going-to* to *gonna* from 4;1.11 (file 91) onwards, so that her development parallels that of Adam. But since *going-to* catches up again at the latest stages of the recordings, the difference between the scores for the two constructions from 4;1.11 onwards is not significant at the 95-percent level (two-sample *U*-test for independent observations, $W = 965.5$, $p = 0.09$). What is more, Sarah's caregivers also shift around that time from the prevalent use of *going-to* to the almost consistent application of *gonna*. For these reasons, it can be excluded that Sarah replicates pathways of historical change.[10]

3.2 *Discussion*

So, are the parallels between ontogenetic and diachronic development "illusory", as Slobin (1994) suggests? I would like to submit that what we called the 'grammaticalization of child speech' as a working hypothesis should be recast as 'grammaticalization effects', which are neither due to the fact that children replicate historical processes, nor

[10] The opposite perspective, i.e. that in historical grammaticalization, adult members of a speech community "are 'recapitulating' developmental processes from early childhood", was already forcefully refuted by Slobin (1994: 128).

to the causal role that is sometimes attributed to children in language change (cf. Croft 2000: ch.3.2 for discussion). There are actually profound differences between the two types of development. As Slobin (2002) points out, the synchronic layering effect of grammaticalization implies that children are presented with the whole range of available variants of a particular grammatical construction at once, whereas in historical grammaticalization, this range only emerges step by step. More precisely, grammaticalization processes in adult speech communities are typically due to pragmatic inferences drawn from particular contexts in which a construction is used (cf. Hopper and Traugott 2003: 81ff.). Contemporary children, by contrast, need to access the current formal and functional array of a grammatical construction, and I would argue that it is two factors that give rise to characteristic developmental patterns.

On the one hand, the most highly grammaticalized variant of a construction tends to have the most general applicability and hence typically occurs most frequently in the ambient language, provided that it is a socially accepted form (i.e. not stigmatized as a marker of a particular style, register, social group). *Gonna*, on this account, would be expected to be acquired relatively early if and only if it is a 'highly available cue', in Bates and MacWhinney's (1989) parlance, in the maternal input. This is, in fact, what we observe in Sarah's input data, hence her comparatively early use of the construction (cf. Fig. 6 again). The absence of *gonna* in Adam's input may precisely be due to its sociolinguistically marked nature: *gonna* is characterised in the *OED* as the "colloquial or vulgar" variant of *going to*, and the caregivers might have deliberately avoided the use of *gonna* because they considered it inappropriate – either its use in child-directed speech more generally or specifically during the recordings for a scientific investigation.[11] At any rate, the input frequency of a variant of a grammatical pattern seems to be a major determinant of the acquisition process.

On the other hand, if one construction is simultaneously available to the child in two different functions, as in the case of the polysemous *going-to-V* pattern, then it may turn out that one of the two is cognitively more easily accessible than the other (cf. Slobin 1994: 129). This is, in fact, the idea behind Johnson's (1999) notion of 'constructional grounding', the hypothesis that more concrete source constructions, whose "interpretations are more easily demonstrated by and inferred from non-verbal cues" (Johnson 1999: 1), are acquired

[11] I am grateful to one reviewer of this paper for bringing this possibility to my attention.

earlier by children than their metaphorical counterparts. Johnson provides empirical evidence for this kind of development for deictic and existential *there*-clauses and the *What's X doing Y?* construction in English. From this perspective, it makes sense that Adam's data testify to a developmental path from literal to metaphorical *going-to-V* usages because the former are grounded in directly observable motion in space, whereas the latter apply to any kind of immanent future context, no matter how abstract it is. In fact, there are communicative situations in the earliest recordings in which the children appear to have difficulties with the interpretation of metaphorical *going-to-V*:

(10) *RIC: *is that money?*
*RIC: *what are you going to buy with that?*
*CHI: *I # simply # going* [/] *going somewhere.*
(Adam 2;10.02)

What is particularly interesting is that in one of the initial sessions, Adam's mother seems to suggest to the child that combinations with *go* are reserved for literal motion:

(11) *CHI: *record go work.*
%mor: n|record v|go n|work.
*MOT: +" *record going to work?*
*MOT: *yes # it is working # but what is it going?*
(Adam 2;3.18)

Conversely, when, in the earliest files, the children are keen on expressing intentional future actions, they do not use any form of *go*, but the simple present tense. Since this pattern is ungrammatical in contemporary English, it is usually corrected to *going-to-V* or *gonna-V* by the respective caregivers (cf. (12)). This strategy provides the child with crucial evidence for the grammaticalized function of the two constructions and may trigger their usage.

(12) *CHI: *I do my dance.*
*MOT: *oh # you gonna do your dancin(g) lesson?*
(Sarah 2;10.24)

Finally, it is worth noting that even at later stages, when the toddlers are presumably aware of the full potential of both constructions, *going-to-V* and *gonna-V* still tend to parcel out their work ecologically: Motion contexts typically trigger a full *going-to* construction, while mere intention or immanent future are expressed by *gonna*. In (13) below this happens in one and the same discourse turn. In fact, of Adam's 628 *gonna* tokens, only 1 clearly involves directed motion. Conversely, we also find examples in which a

metaphorical use of *going-to* is reacted upon by the child's use of *gonna* (14). This latter phenomenon shows not only that Adam has come to learn that the two constructions can potentially be used interchangeably, but also that he may conceive of *gonna* as the more appropriate variant in metaphorical contexts.

(13) *CHI: *Mommy # he's going to dump dis one off.*
*CHI: *he's gonna kill it.*
(Adam 4;6.24)

(14) *URS: *that's going to be a big job.*
*CHI: *that's gonna be a big job eye?*
(Adam 4;7.01)

In sum, constructional grounding, which "can be considered a special case of a more general process of conceptual development" (Johnson 1999: 1), provides an explanatory tool for the earlier appearance and mastery of literal *going-to-V* usages. The shift in preference from *going-to* to *gonna*, which at least in Adam's case is significant and deviant from the recorded input, may be related to the growing recognition that *gonna-V* is the most widely applicable marker of immanent futurity and that *going-to-V* is best reserved for contexts which do still involve literal motion. Therefore, the seeming parallels between ontogeny and diachrony reduce to a 'pseudo-grammaticalization' of child language.

Of course, this does not deny the more overarching similarities between acquisition and change in the emergentist paradigm, especially in the cognitive mechanisms that drive the analogical extension of novel patterns and their consolidation as productive grammatical constructions, and in the ways in which usage frequencies affect these diachronic patterns in both acquisition and change (cf. Diessel in press for a systematic overview).

4. Conclusion

This paper has provided a quantitative analysis of the development of *going-to-V* and *gonna-V* in child language from a decidedly cognitive-functional vantage point. On this view, the acquisition of morphosyntax is not accomplished by linking input data to maximally abstract and fully productive categories of a prespecified grammatical representation. Children are rather considered to be conservative learners, closely attending to input distributions and their communicative contexts, from which they can extract form-meaning pairings and progressively generalise across related units. I have suggested that children gradually develop a network of specific low-

level chunks of a schematic adult construction, and that rather different structural realisations of what appears to be one and the same construction coexist in child language production.

We have also investigated how constructions as richly layered as the variants of *be-going-to-V* are structured in children's language use: it was suggested that, especially in Adam's case, physically 'grounded' layers of the polysemous *going-to-V* pattern develop prior to the more abstract, metaphorical variants, and that there can be large-scale shifts in preferences for the more highly grammaticalized construction. These patterns of use were recast as 'grammaticalization effects' that are analogous to historical grammaticalization only to the extent that they involve similar psychological mechanisms of categorization, analogical perception and ecological systematization of constructions. What is more, if the most highly grammaticalized variant of a construction (here: *gonna-V*) is frequently available in the ambient language (corresponding to what grammaticalization theory would predict), then the replication of historical pathways in child speech becomes even more imperfect: Frequently modelled form-function pairings are picked up early, and further develop in parallel to the other variants (or layers) of the construction. This is what we observed in Sarah's case. Taken together, the study has also demonstrated how differences in input frequencies, cognitive accessibility and competition of constructional variants can lead to different developmental patterns across children rather than uniform syntactic acquisition.

One of the most interesting issues for future research into *going-to-V* and *gonna-V* in acquisition would be their external constructional relationships (rather than the internal ones described in this paper). In the recent developmental literature (e.g. Abbot-Smith and Behrens 2006), due attention has been paid to how formally related constructions may support or interfere with the acquisition of a particular target pattern. Along the same lines, a closer inspection of our data and the transcripts does, in fact, suggest that the serial *go-V* construction (*I go get it.*, cf. Wulff 2006), grammatically simple and well-grounded in physical experience, may serve as an important precursor or model construction for *going-to-V*, but their precise interaction would need to be spelled out in more detail. Similarly, given that the more grammaticalized variants of *going-to-V* participate in a rich inventory of future constructions in English, their relationship to these other, potentially competing forms such as *will*, needs further exploration (though see Klecha et al. 2007 for a study in progress).

References

Abbot-Smith, Kirsten and Heike Behrens
2006 How known constructions influence the acquisition of other constructions: The German passive and future constructions. *Cognitive Science* 30(6), 995-1026.

Abbot-Smith, Kirsten and Michael Tomasello
2006 Exemplar-based learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review* 23, 275-290.

Baayen, Harald
2008 *Analyzing Linguistic Data.* Cambridge: Cambridge University Press.

Bates, Elizabeth and Brian MacWhinney
1989 Functionalism and the Competition Model. In Bates, Elizabeth and Brian MacWhinney (eds.), *The Cross-Linguistic Study of Language Processing.* New York: Cambridge University Press, 3-73.

Bybee, Joan
2003 Cognitive processes in grammaticalization. In Tomasello, Michael (ed.), *The New Psychology of Language. Vol. II.* Mahwah, NJ: Erlbaum, 145-167.
2006 *Frequency of Use and the Organization of Language.* Oxford: Oxford University Press.

Brown, Roger
1973 *A First Language: The Early Stages.* Cambridge, MA: Harvard University Press.

Chomsky, Noam
1999 On the nature, use, and acquisition of language. In Ritchie, William C. and Tej K. Bhatia (eds.), *Handbook of Child Language Acquisition.* San Diego: Academic Press, 33-54.

Croft, William
2000 *Explaining Language Change: An Evolutionary Approach.* London: Longman.
2001 *Radical Construction Grammar: Syntactic Theory in Typological Perspective.* Oxford: Oxford University Press.

Cysouw, Michael
2007 New approaches to cluster analysis of typological indices. In Köhler, Reinhard and Peter Grzbek (eds.), *Festschrift für Gabriel Altmann.* Berlin: Mouton, 61-75.

Dąbrowska, Ewa
2004 *Language, Mind and Brain. Some Psychological and Neurological Constraints on Theories of Grammar.* Washington, DC: Georgetown University Press.

Diessel, Holger
in press Language change and language acquisition. In Bergs, Alexander and Laurel Brinton (eds.), *Historical Linguistics of English: An International Handbook.* Berlin: Mouton de Gruyter.

Fischer, Olga
2007 *Morphosyntactic Change: Formal and Functional Perspectives*. Oxford: Oxford University Press.

Friendly, Michael
1994 Mosaic displays for multi-way contingency tables. *Journal of the American Statistical Association* 89, 190-200.

Goldberg, Adele
2003 Constructions: A new theoretical approach to language. *Trends in Cognitive Sciences* 7(5), 219-224.
Gries, Stefan Th.
2006 Some proposals towards more rigorous corpus linguistics. *Zeitschrift für Anglistik und Amerikanistik* 54(2), 191-202.
2008a Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4), 403-437.
2008b *Statistik für Sprachwissenschaftler.* Goettingen: Vandenhoeck und Ruprecht.
Gries, Stefan Th. and Sabine Stoll
in press Finding developmental groups in acquisition data: Variability-based neighbor clustering. *Journal of Quantitative Linguistics* 16(3).
Haiman, John
1994 Ritualization and the development of language. In Pagliuca, William (ed.), *Perspectives on Grammaticalization.* Amsterdam: John Benjamins, 3-28.
Hopper, Paul J. and Elizabeth Closs Traugott
2003 *Grammaticalization*. 2nd ed. Cambridge: Cambridge University Press.
Johnson, Christopher R.
1999 *Constructional Grounding: The Role of Interpretational Overlap in Lexical and Constructional Acquisition*. Ph.D. dissertation, University of California, Berkeley.
Juilland, Alphonse G., Dorothy R. Brodin and Catherine Davidovitch
1970 *Frequency Dictionary of French Words.* The Hague: Mouton.
Kirjavainen, Minna, Anna L. Theakston, Elena V.M. Lieven and Michael Tomasello
in press *I want hold Postman Pat*: An input-driven explanation for children's infinitival-*to* omission errors. *First Language.*
Klecha, Peter, Joseph Jalbert, Alan Munn and Cristina Schmitt
2007 Explaining why *gonna* precedes *will* in acquisition. Paper presented at the 32nd *Boston University Conference on Language Devleopment*, November 2007.
Kuczaj, Stan A. and Michael P. Maratsos
1983 Initial verbs in *yes-no* questions: a different kind of general grammatical category? *Developmental Psychology* 19, 440-444.
Langacker, Ronald W.
1987 *Foundations of Cognitive Grammar. Vol. I: Theoretical Prerequisites.* Stanford: Stanford University Press.
Lehmann, Christian
1995 *Thoughts on Grammaticalization*. München: Lincom Europa.
Leech, Geoffrey N., Paul Rayson and Andrew Wilson
2001 *Word Frequencies in Spoken and Written English: Based on the British National Corpus*. London: Longman.
Paradis, Emmanuel
2006 *Analysis of Phylogenetics and Evolution with R*. New York, Springer.
Richards, Brian J.
1990 *Language Development and Individual Differences: A Study of Auxiliary Verb Learning.* Cambridge: Cambridge University Press.
Rowland, Caroline F.
2007 Explaining errors in children's questions. *Cognition* 104, 106-134.

Ruhland, Rick, Frank Wijnen and Paul van Geert
1995 An exploration into the application of dynamic systems modeling to language acquisition. In Verrips, Maaike and Frank Wijnen (eds.), *Amsterdam Series in Child Language Development: Vol.4. Approaches to Parameter Setting.* Amsterdam: University of Amsterdam, 107-134.
Saitou, Naruya and Masatoshi Nei
1987 The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution* 4, 406-425.
Schmidtke, Karsten
2007 The acquisition of English purpose clauses. Paper presented at the theme session on language acquisition, *10th International Cognitive Linguistics Conference,* Krakow.
Slobin, Dan I.
1994 Talking perfectly: Discourse origins of the Present Perfect. In Pagliuca, William (ed.), *Perspectives on Grammaticalization.* Amsterdam: John Benjamins, 119-133.
2002 Language evolution, acquisition and diachrony: Probing the parallels. In Givón, Talmy and Bertram F. Malle (eds.), *The Evolution of Language out of Pre-Language.* Amsterdam: John Benjamins, 375-392.
Theakston, Anna L., Elena V.M. Lieven, Julian M. Pine and Caroline F. Rowland
2005 The acquisition of auxiliary syntax: BE and HAVE. *Cognitive Linguistics* 16(1), 247-277.
Tomasello, Michael
2003 *Constructing a Language: A Usage-Based Theory of Language Acquisition.* Cambridge, MA: Harvard University Press.
von Eye, Alexander
1990 *Introduction to Configurational Frequency Analysis*. Cambridge: Cambridge University Press.
Wiechmann, Daniel
2008 Looking for the right type: Towards a principled entrenched pattern recognition. Paper presented at the conference on *Language, Communication and Cognition*, Brighton, UK.
Wulff, Stefanie
2006 *Go*-V vs. *go-and*-V in English: A case of constructional synonymy? In Gries, Stefan Th. and Anatol Stefanowitsch (eds.), *Corpora in Cognitive Linguistics. Corpus-based Approaches to Syntax and Lexis*. Berlin, Heidelberg, New York: Mouton de Gruyter, 101-125.