Multi-level Exemplar Theory

Michael Walsh, Bernd Möbius, Travis Wade, Hinrich Schütze

Institute for Natural Language Processing

University of Stuttgart, Germany

Abstract

This paper presents recent research which provides an over-arching model of exemplar theory capable of explaining phenomena across the phonetic and syntactic strata. The model represents a unique exemplar-based account of constituency interactions encompassing both linguistic domains. It yields simulation and experimental results in keeping with experimental findings in the literature on syllable duration variability and offers an exemplar-theoretic account of local grammaticality. In addition, it provides some insights into the nature of exemplar cloud formation, and demonstrates experimentally the potential gains that can be enjoyed via the use of rich exemplar representations.

Exemplar Theory was initially proposed in the domain of psychology (Nosofsky, 1986; Hintzman, 1986). However, recent years have seen a growing body of research into exemplar-based theories of perception and production and how they can account for certain linguistic phenomena. The particular attraction of exemplar-based models lies in their ability to explain phenomena which more abstractionist models find problematic. Unlike more traditional, often generative, rule-

oriented approaches, at the core of Exemplar Theory is the idea that the acquisition of language is significantly facilitated by repeated exposure to concrete language input. Central to Exemplar Theory are the notions of frequency, recency, and similarity. Extensive storage of language input exemplars takes place, categorization of input is made by comparison with extant exemplar memory traces, production is facilitated by accessing these stored exemplars, and the exemplar memory is in a constant state of flux with new inputs updating it and old unused exemplars gradually fading from memory.

Hay and Bresnan (2006) note the relatively independent development of Exemplar Theory research on phonetics and syntax and argue in favor of combining the two strands in the belief that joint predictions might emerge which neither research area alone would yield. Before briefly examining each of these lines of enquiry in turn, it is worth noting that the research presented in this paper complements Hay and Bresnan's combination of the research literatures, by presenting a single over-arching model capable of explaining phenomena from both fields, phonetics and syntax. The key innovation of the model is its explicit formalization of the relationship between exemplars on the *constituent* level and exemplars on what is referred to as the *unit* level. Constituents are segments, for example consonants and vowels, in phonetics, and words in syntax. Units are represented by syllables in phonetics and by phrases or sentences in syntax. The general hypothesis posited here is that a competition exists between the submodel at the level of constituents and the submodel at the level of units and that the unit level submodel "wins" if the unit exemplar receives sufficient activation. Although a similar relationship between constituents and units can be found in other models (Grossberg, 2003), to the authors' knowledge the model presented here is the first to explicitly model and invoke constituency interaction in Exemplar Theory to explain a number of phenomena. In particular this single model accounts for syllable frequency effects and the acquisition of local grammatical knowledge, and explores constituency interaction within the

syntax and phonetics domains.

The next two sections explore background exemplar-theoretic research from the perspective of syntax and phonetics respectively. These sections are followed by section 3 which presents a general overview of the multi-level model. This model is instantiated in section 4 for syllable duration modeling and in section 5 for local grammaticality modeling. Section 6 concludes the article with some discussion on the results achieved and opportunities for future work.

1. Exemplar-theoretic Syntax

Recent years have seen considerable debate concerning the nature and extent of children's early syntactic representations and the potential influence of distributional properties of the input upon them. One side of the debate supports the view that from a very early stage children possess abstract, generalized knowledge of the syntax of their language (or at least aspects of it), e.g. (Naigles, 1990; Gertner, Fisher, & Eisengart, 2006).

Another facet of the debate, however, centers around the proposition that abstract syntactic knowledge emerges over time and that early syntactic representations are organized around particular lexical exemplars the child has encountered (Childers & Tomasello, 2001), and a grammar is essentially an emergent property of two key processes: a) storage of exemplars, and b) inference (for categorization and production) over exemplars. The next three subsections examine exemplar phenomena in children and adults and attempt to model these effects.

Exemplar-theoretic syntax in language acquisition

The first corner-stone of exemplar-theoretic linguistic research is that during language acquisition children perceive and store concrete pieces of language which they avail of to analyze and produce new utterances. A significant body of evidence supports the idea of word and phrasal storage and sensitivity to their respective frequencies when it comes to syntactic generalizations and

productivity. In particular, much of this work has focused on investigating children's understanding of transitive, passive, and dative constructions, and their sensitivity to novel verbs and nouns, by using preferential looking experiments (based on an individual's tendency to look at a scene related to what he or she hears, instead of looking at an unrelated scene) and elicitation experiments where children are encouraged to produce sentences employing a target structure or word. In the case of the elicitation experiments the children are often primed with particular structures or lexical items beforehand. This priming technique takes advantage of an observed effect whereby individuals tend to reuse syntactic structures or lexical items which have recently been employed. Pinker et al. (1987) found that 3-and-half-year-old, and older, children presented with a novel verb in a passive sentence, e.g. *The fork is being floosed by the pencil*, were able to produce canonical transitive sentences, i.e. sentences of the form subject-verb-object (SVO), using the novel verb, e.g. *It's floosing the fork*, although they had never heard the verb used in such a construction. However, a number of studies by Tomasello and colleagues also examining children's ability to produce novel transitives, using novel verbs, would appear to indicate otherwise for younger children (Tomasello & Brooks, 1998; Olguin & Tomasello, 1993; Akhtar & Tomasello, 1997).

Akhtar (1999) report that 4-year-old children "corrected" to canonical word order with novel verbs which had only been presented in "weird word order", no matter how often they heard them. Younger children, however, were equally likely to employ SOV and VSO word orders for the verbs they heard used in that form as they were to correct to SVO order. Building on Akhtar's work, Matthews et al. (2005) employed English verbs of varying frequencies to establish if lexical frequency influences children's knowledge of word order as a grammatical marker. In an elicited production task they found that younger children (2;9) have a tendency to match weird word order, which is replaced by a partial correction to English word order, and then full transitivity as verb frequency increases. They argue that "the acquisition of word order,..., is not a binary affair but is rather

an instance of gradually strengthening, graded representations"(Matthews et al., 2005, p.132) . Interestingly, the model presented in section 3 employs graded representations as the backbone for acquiring local syntactic knowledge.

Further work by Childers and Tomasello (2001), building on research by Nelson (1977), investigated 2-and-a-half-year old children's understanding of the English transitive construction, this time with a particular focus on pronouns. They presented a large number of exemplars of a novel syntactic construction to children over a short period of time, with the result that acquisition of the construction was facilitated significantly. Huttenlocher et al. (2002) found a a significant correlation between the proportion of multi-clause sentences produced by children and the proportion produced by their parents and school teachers (used to control for genetic advantage). Indeed, Huttenlocher et al. found that the same relative frequencies of different multi-clause sentence types (coordinate clauses, relative clauses, and complement clauses) found for parents were also found for children, and that the complexity of teacher speech was significantly related to childrens' syntactic growth. From an exemplar-theoretic perspective both the Childers and Tomasello study and those of Huttenlocher et al. provide evidence for emergent exemplar-sensitive syntactic development.

Other insights into emergent exemplar-based syntactic acquisition come from Savage et al. (2003) who found that 6-year-old children, in a picture description task, could be successfully primed both lexically and structurally for active transitive and passive sentences, which they argue is an indication that these children have some knowledge of abstract structure. Children aged 3, however, showed only lexical priming indicating that their language knowledge is more item based. The performance of 4-year-old children lies in between and Tomasello (2006) argues that this kind of result is evidence for the development of stronger and more abstract representations over time, based on exposure to exemplars of particular structures.

Some of these studies which appear to indicate item-based gradual, and late, development

of syntactic abstraction are not without their critics. Recent work by Bencini and Valian (2008), which in contrast to (Savage et al., 2003) provided a familiarization phase with all nouns and verbs used in the task to reduce cognitive load due to lexical lookup, suggests that "young" 3-year-olds do indeed exhibit abstract priming (priming across sentences where content words are not shared) with passives. Fisher (2002) argues that although a child might well be in possession of an abstract understanding of what a verb is, this does not necessarily entail that the child should immediately be willing to use any new verb in any sentence construction, one reason being that some verbs can be employed both transitively and intransitively (and convey a different meaning) whereas others operate in only one of these syntactic frames. In other words, conservative verb use does not necessarily discredit arguments for early abstraction (but see Tomasello and Abbot-Smith, 2002, for a refutation.)

Further recent work indicative of early syntactic abstraction includes Gertner et al. (2006), who found that children as young as 21 months old used word order to interpret transitive sentences containing a novel verb, and demonstrated knowledge of the link between subject and agent, and object and patient (see also Fernandes et al., 2006). Conwell and Demuth (2007) demonstrated that children may well possess an abstract understanding of the dative alternation and can use it productively under certain experimental conditions.

This ongoing debate concerning the nature of children's early syntactic representations is significant in that it highlights the importance of exemplar- or item-based learning and sensitivity to frequency and distributional factors in language acquisition, despite the fact that a common consensus has not necessarily been reached with respect to rate of development and indeed whether or not abstract syntactic categories are an innate part of the language acquisition system. Furthermore, the evidence for sensitivity to input and the gradual emergence of abstraction provided in some of the papers above motivates the use of thresholding in the exemplar model presented in

section 3 and the use of representational units rich in distributional information.

However, exemplar effects are not limited to the domain of child language acquisition. Sensitivity to frequency and distributional factors is also found in adult language.

Exemplar-theoretic syntax in adults

Considerable evidence exists in the literature on adult language use in support of Exemplar Theory. For example, eye-tracking experiments provide additional evidence with native speakers attending to a word for less time if it is the final word in a frequent formulaic sequence (e.g. an idiom or well known phrase) than a non-formulaic sequence (Underwood, Schmitt, & Galpin, 2004). Further evidence is also found in reaction time experiments. In an experiment by Bod (2000) subjects had to decide, as quickly as possible, whether or not a given string was an English sentence. Bod found that high-frequency sentences received faster reactions than low-frequency sentences and concludes that frequent sentences must be stored in mind.

Psycholinguistic data from the adult priming literature offers further potential evidence for storage and use of structural exemplars in adults. Bock (1986) and Bock and Loebell (1990) found that speakers, when repeating prime sentences followed by descriptions of target pictures (semantically unrelated to the primes) exhibited an increased tendency to produce an active description having heard an active prime, a passive description after a passive prime, or a prepositional-dative description after a prepositional-dative prime. Pickering and Branigan (1998) discovered syntactic priming effects in a written sentence completion task. They report that priming can take place regardless of whether or not the verb is shared between the prime and the target, that the magnitude of priming increases with the overlap between prime and target, and that priming occurs even across different word forms of the verb between prime and target. Though one might argue that these results might be more indicative of the activation of a more abstract representation, it is important to note that Pickering and Branigan found a considerable increase in priming when

the prime and target overlapped. Hence it is possible that there might be some sort of interaction between exemplars and abstract representations at play here (this is, after all, adult data, and hence post-acquisition, so abstract representations are to be expected). Chang et al. (2000), citing evidence for persistence of priming over intervening sentences (Bock & Griffin, 2000), and over time (Saffran & Martin, 1997), argue that this persistence is evidence that structural priming might well be a form of implicit learning as the durations involved mean that neural activation cannot be the only mechanism involved and some longer-term change to the production system is taking place. Chang et al. corroborate these persistence findings using a computational model (see below). A detailed review of considerable additional research into human sensitivity to word frequency can be found in (Jurafsky, 2003).

Given the growing and significant body of evidence for exemplar storage and sensitivity to exemplar frequency in both children and adults, the next area to briefly examine is analogy-based classification over exemplars.

Formal and informal models of exemplar effects

From the perspective of modeling, to date one of the most formally elaborated approaches is the Data Oriented Parsing (DOP) model (Bod, 2006). According to the DOP approach each exemplar corresponds to the syntactic structure of a perceived utterance. On presentation of a novel utterance the DOP model employs 1) decomposition operations to split the utterance into a set of fragments, 2) composition operations for recombining fragments to produce an analysis of the utterance, and 3) a probability model which indicates how the probability of a new utterance and its meaning is arrived at on the basis of its fragments' frequencies. Interestingly, the k nearest-neighbor and radius-based approaches typically used in exemplar-based phonetics research to compare novel stimuli to extant exemplars are absent in the DOP model which captures productivity by means of a substitution process involving categories of the same type in similar phrase-structure locations.

The model presented later in this article, however, demonstrates how local grammatical knowledge can be implicitly acquired based on distributional data and a radius-based similarity measure.

Other exemplar approaches to aspects of language acquisition have used connectionist methods. Elman (1990) trained a simple recurrent network (SRN)[1] to predict the next item in an input sequence generated by a grammar capable of generating 2 and 3 word sentences. Learning took place purely on the basis of distributional information, no semantic or category label information was provided to the network. In related work, Morris et al. (2000) employ a SRN and a small vocabulary to attempt to map sequences of words to semantic roles, e.g. agent, patient, experiencer, percept. The model was trained on a variety of sentence types and was then tested on the basis of two systematic gaps in the training data. The model failed to generalize in cases that are also very rare in parental input but generalized well in cases of synergy of syntactic constructions around the gap. Evidently, learning occurred in accordance with exemplar-theoretic predictions and on the basis of distributional information.

Other connectionist work includes Chang and colleagues (Chang, 2002; Chang, Dell, & Bock, 2006) who, building on earlier work (2000), developed a dual-path model[2]. This model comprises two pathways for influencing word production. The first pathway is a feed-forward network responsible for mapping from the message to the lexicon. The message consists of concepts and event roles (e.g. transitive agent), and the bindings between them (represented as fast-changing weights). The second pathway is an SRN responsible for sequence prediction. Both pathways converge at the model's output layer where words consistent with the intended message are produced. This

[1] An SRN is a form of artificial neural network with the addition of a set of *context units* which are employed to maintain a memory of the hidden units' previous values, allowing the network to perform sequence-prediction, i.e. predicting a word given the previous word.

[2] This dual path consists of a meaning system and a sequencing system and is not to be confused with the multi-level architecture proposed in this paper. The two paths of the multi-level architecture correspond to units at different levels within the same system. See section 2.

model was trained on message-sentence pairs (adapted to prime-target pairs) and employed in a structural priming test. The model produced significantly more target structures (actives, prepositional datives) when preceded by a prime of the same structure than with primes of the alternative structure and exhibited persistence of priming over intervening sentences.

Chang et al. argue that the mechanism of priming is the same error-based learning (i.e. incremental learning achieved by exploiting the difference between a predicted output and a target output) that is used to acquire language in the first place and that the structural representations that are primed emerge from the interaction between the learning algorithm and the model's dual-pathway architecture and role-concept bindings. Consequently, they applied the model to the debate referred to above concerning the understanding of the transitive construction in children younger than 3 years of age. In the production case the model's ability to produce accurate novel transitives emerged over time as it learned from experience with sentences with real verbs, and coincided well with results found by Tomasello (2000). In the preferential looking case, in order to mimic that paradigm, the model has to make a choice between two form-meaning options (one of which is a mismatch). The model learns to do this considerably earlier (before 4000 learning epochs, or earlier than the model's second birthday) than it can produce novel transitives, again in keeping with the literature, this time for preferential looking experiments. Hence, Chang et al. argue that the model incorporates features of both late and early-syntax approaches, the argument being that while abstract knowledge emerges over time, early representations are sufficient to make choices between two interpretations. The model presented in this paper focuses on the impact of distributional factors in acquiring syntax – as opposed to the semantic factors that Chang et al. address. It is also novel in that it directly tests judgments on grammatical and ungrammatical sentences.

From the perspective of Exemplar Theory each of the computational models described above

demonstrates the learnability of syntactic properties on the basis of distributional information, strongly motivating an exemplar-theoretic account of language acquisition. The model presented in this article seeks to do likewise but uses rich exemplar representations and the invocation of constituency interaction to do so. In addition the same model proves useful at the phonetic level too.

Formal computational models aside, Abbot-Smith and Tomasello (2006) hold the view that comprehension of an exemplar must, minimally, result in a change in its representation (even if this is a simple recording of frequency). Furthermore, they also propose that frequent summing over mutual similarities of a particular cloud of exemplars is highly likely to result in a permanent modification to the representation which is "in some way equivalent to the formation of some kind of more abstract representation" (Abbot-Smith & Tomasello, 2006, p.282). The hybrid categorization model which they propose allows for exemplar learning and retention but also offers an abstraction mechanism. It is important to note that while Abbot-Smith and Tomasello posit an interesting potential account of acquisition no formal model is provided.

This section has synopsized usage-based research from the literature on child and adult language acquisition and use and outlined computational approaches aimed at modeling some of the phenomena found. The next section undertakes a similar review of the relevant literature for phonetics.

2. Exemplar-theoretic Phonetics

Underpinning research into exemplars in speech production and perception is the general idea that encountered items (segments, words, sentences etc.) are stored, rich in phonetic detail, in memory, along with extra-linguistic information. These exemplars are categorized, on the basis of their similarity to extant stored exemplars (using a variety of metrics), into clouds of memory traces with similar traces lying close to each other while dissimilar traces are more distant.

Johnson (1997) offers an exemplar-based attention-weighted k nearest-neighbor model which successfully treats aspects of vowel perception, sex identification and speaker variability, crucially without employing the traditional notion of normalization. Other informative research includes Pierrehumbert's model of lenition, entrenchment, and neutralization in diachronic language change in the context of a perception-production loop (Pierrehumbert, 2001). At the core of her model is the idea that exemplars have a resting activation level, with exemplars encoding *frequent* and *recent* percepts having higher resting activation levels than exemplars encoding infrequent temporally distant percepts. Classification of a new exemplar is a function of its similarity to stored exemplars, where the labeling with the highest probability is computed, given the labeling of the stored exemplars in the neighborhood. Pierrehumbert's model uses a fixed size neighborhood (see section 6). The strength or activation of an exemplar is a function of its frequency and recency within the exemplar space.

Further noteworthy work includes Bybee (1999; 2006) who presents three convincing observable phenomena which augment the argument for exemplar storage at both the phonetic and syntactic levels, including: a) Reducing Effect – a higher rate of phonetic reduction for high frequency words than mid and low frequency words, the rationale being that the articulatory representation of words and word sequences is composed of neuromotor routines whose execution becomes fluent with repeated use, b) Conserving Effect – greater entrenchment of morpho-syntactic structure for high frequency sequences because the high frequency of a sequence strengthens its memory representation and facilitates its access as a single unit, c) Autonomy – high frequency complex morphological forms can lose their internal structure thus becoming in some sense independent of their etymological roots, e.g. the grammaticalization of the meaning of the sequence *be going to* into a future or intentional form. The multi-level model presented below is to some extent capable of modeling each of these effects.

Additional relevant research concerning exemplars and frequency of occurrence, important to the work presented here, concerns evidence for syllable storage in a mental repository, and dual-route production evidence.

Levelt and colleagues (1994; 1999) posit that frequently occurring syllables are stored in the form of learned motor programs in a phonetic mental syllabary. Such a syllabary facilitates ease of production by enabling speakers to produce the majority of their speech by using these motor programs which are essentially prefabricated units. The dual-route concept posits a direct and indirect route for unit (e.g. syllable) production, where the indirect route constitutes production via assembly, and the direct route constitutes retrieval from the syllabary of syllable production templates. Furthermore, the syllabary also stores coarticulatory effects since most coarticulation is syllable-internal. Indeed, high-frequency syllables, which are presumed to be stored in the syllabary, have been shown to exhibit more coarticulation than rare syllables (Whiteside & Varley, 1998), which are assembled online from the gestural specifications of smaller units, i.e. concatenated as a sequence of segment-level specifications. This online beads-on-a-string assembly occurs for rare syllables because the sequence of segments is not represented in the syllabary as a motor program, i.e. there is no gestural specification for the sequence as a whole in the phonetic syllabary that can act as a production target. Whiteside and Varley's research focuses on speech produced by patients suffering from apraxia of speech (AOS) (Varley & Whiteside, 2001). These patients produce speech which lacks many coarticulation effects typical for nonpathological speech and has been characterized as resembling allophone synthesis, that is it has a somewhat concatenative quality. Whiteside and Varley (1998) argue that features such as inconsistent articulatory movements, increased durations, and reduced gestural overlap found in speech of this kind, are indicative of disruption of stored movement gestalts. One possible reason is that AOS patients have a (gradient) loss of access to these learned motor programs for high frequency syllables. Consequently, AOS

speakers are forced to produce speech using indirect means. Thus, loss of access to a high frequency syllable gestalt would necessitate sub-syllabic assembly, whereas in nonpathological speech the syllable would be retrieved directly from the syllabary. Varley and Whiteside (2001) proposed a sub-syllabic route model, which predicts correctly that AOS speakers are disfluent without producing segment-level speech errors. Crucially, the model also predicts that error patterns would not be affected by syllable structure.

Further significant research in support of dual-route access to exemplars can be found in a neuroimaging (fMRI) study by Mayer et al. (2003). In their study they seek to establish differing brain activity patterns, as well as differing localizations of active neural clusters, depending on whether the direct or indirect routes are exploited. Subjects read bisyllabic nonsense words constructed from very high frequency (hf) and very low frequency (lf) syllables in all four possible combinations (hf-hf, hf-lf, lf-hf, lf-lf). High frequency syllables evoked activation patterns in the left temporal cortex that were absent in the case of low frequency syllable production. However, these activation patterns always co-occurred with activation patterns in the left motor and pre-motor cortices, areas known to be correlated with segmental assembly. This result might well indicate that both the direct and indirect routes operate in parallel and are in competition for the most efficient coverage of the phonological target sequence.

Building on this idea that stored exemplars act as production targets, or plans of articulation, Schweitzer and Möbius (2004) note that if this is the case then speakers should have a significant number of exemplars for high frequency syllables, which would then act as a production target region, and a small or negligible number of exemplars for low frequency syllables. Consequently they argue that low frequency syllables would have to be computed online from exemplars of their constituent segments. They predicted, and observed, greater variation in duration for frequent syllables than for infrequent syllables when looking at the relationship between syllable duration z-

scores (measure of standard deviations from the mean) and the duration z-scores of the constituent segments in a corpus of German. The prediction that frequent syllables would exhibit greater variation in duration is based on the intuition that these syllables would occur in a larger number of contexts than for infrequent syllables and that there is hence a larger opportunity for context-influenced variability. It is worth noting that while Bybee offers no formal definition of high, mid, or low frequency, the Schweitzer and Möbius experiments were performed using frequency bins derived from multivariate clustering by Müller et al. (2000) (as indeed were the fMRI experiments performed by Mayer et al., 2003). The research presented here, using the same data, employs a computational model which goes some considerable way to corroborating this effect.

Exemplar Theory has enjoyed much growth in both the phonetic and syntax domains (Pierrehumbert, 2001; Croot & Rastle, 2004; Bod, 2006; Bybee, 2006), yet little has been attempted, much less achieved, with respect to unifying research from both fields. However, Hay and Bresnan's (2006) combination of the literatures represents a noteworthy exception. In their examination of the phonetics of the common phrase *giving a hand*, they sought to establish evidence for within-word variation when a word occurs in a different syntactic or semantic location. They found that different constructions can indeed affect phonetic change, and that frequent phrases appear to be the most advanced in the sound changes examined. Most notably their research appears to indicate that *phonetically detailed* phrases may well be stored in memory. This is particularly noteworthy as it demonstrates the utility of combining research from both syntax and phonetics. Building on this idea, this paper presents a multi-level exemplar-based model of constituency interactions across both linguistic domains, which the authors believe represents a significant first step towards a unified account of exemplar-theory. Aspects of the research presented here can be found in (Walsh, Schütze, Möbius, & Schweitzer, 2007; Walsh, Schütze, Wade, & Möbius, 2007; Schütze, Walsh, Wade, & Möbius, 2007), and the present article unifies and builds upon this earlier work. The

instantiation of this model for the explanation of syllable frequency effects presented in section 4, combined with the instantiation of the model for grammaticality acquisition in section 5, provides insights into how a unified account might be achieved. These are discussed in section 6.

3. The Multi-Level Exemplar Model

Given the evidence for the mental syllabary and dual-route production pathway, combined with the syllable frequency effects discussed in section 2, the instantiation of the model presented in section 4 seeks to provide, through simulation and experimentation, corroborative evidence for these phenomena, and, in particular, offer possible explanations for why the syllable frequency effect reported in Schweitzer and Möbius (2004) might occur. Similarly, given the wealth of evidence suggesting both child, and indeed adult, sensitivity to lexical and structural frequency in language presented in section 1, the model offered below aims to both formalize the potential power of local-context distributional information in a grammaticality judgment task, and to demonstrate how rich graded exemplar representations are superior to standard categorical representations in judgments in this same task. Furthermore, to the authors' knowledge no formal exemplar model which captures both phonetic and syntactic phenomena currently exists, and the model outlined below attempts to bridge this gap.

The architecture of the model is shown in Fig. 1. The model has 4 components and two databases, and it receives input from a generation/perception interface.

- **Generation/perception interface**. This interface transmits a (possibly underspecified) input ("input" in Fig. 1) that serves as stimulus for the model. It is either instantiated by a speaker different from the one being modeled (as when grammaticality judgments are modeled) or as the part of the cognitive system that determines which words or phrases are to be generated next. For example, when applying the model to syntax the unit exemplar *xyz* is a perceived sequence of words the comprehension or parsing of which is then modeled as shown in Fig. 1. Similarly, in the case of

phonetics, *xyz* corresponds to an abstract representation of the articulatory gestural specification of the next syllable to be produced.

• **Similarity calculator**. The similarity calculator (not shown in figure) takes a stimulus and a database of exemplars as input and identifies the subset of exemplars in the database that have a minimum similarity[3] with the stimulus. It returns this subset along with the individual similarities that were calculated. It also returns the level of activation, which in the simplest case is the number of similar exemplars that were found (or activated by the retrieval).

• **Exemplar database on the unit level**. This database is the repository of unit exemplars ("Unit exemplar database" in Fig. 1).

• **Exemplar database on the constituent level**. This database is the repository of constituent exemplars ("Constituent exemplar database" in Fig. 1).

• **Parser and composer**. The parser parses a unit into its constituents (upper right arrow marked "Parse") and the composer composes a sequence of constituents into a unit (bottom arrow marked "Compose"). These two components are different for each of the instantiations of the model. For example, a stimulus sentence is parsed into individual words before testing against the model's exemplar memory; or, from the phonetics perspective, when composing segments into a syllable, the articulatory plan for the syllable is the concatenation of the articulatory plans of the segments. As a result, the duration of the syllable is equal to the sum of the durations of its constituents. Clearly the ability to compose/decompose a syllable from/to its constituent segments, or to parse a string into a sequence of words, is an acquired skill. Parsing, segmentation and composition constitute complex research areas in their own right and are therefore assumed in the current instantiation of the model.[4]

[3]This similarity is a parameter. In general, the value of this parameter will be different for different databases. See section 6 for discussion.

[4]Note that, for the syntax acquisition model, it is not assumed that the input is *syntactically* parsed. Rather, the

• **Decision component.** Let α be the activation the input receives in the unit exemplar database as calculated by the similarity calculator. If α is above a threshold θ, then perception or production will be based on the set of similar units found by the similarity calculator in the unit exemplar database. If the stimulus does not receive sufficient activation in the unit exemplar model, then perception/production is based on the sets of similar constituents (one set per constituent) found by the similarity calculator in the constituent exemplar database.

Table 1 shows how the model is instantiated in the phonetic and syntactic models. The following sections describe these instantiations in more detail.

The methodology in this paper is to model the input data in a particular linguistic scenario (articulation and language acquisition), present the model in Fig. 1 with these input data, and then compare the predictions of the model with the outcome that was observed in the linguistic scenario with a view to establishing proof-of-concept. These simulations are then followed by experiments where the model is presented with actual linguistic input data, rather than modeled data, in order to evaluate the model concept and investigate robustness.

4. Modeling syllable duration variability

In an exemplar model of speech production, exemplars serve as targets or plans of articulation. Recall that Schweitzer and Möbius (2004) posited that speakers should have a significant number of high frequency syllable exemplars acting as production target regions, and a small or negligible number of exemplars for low frequency syllables. On this basis they argue that low frequency syllables would be computed online from exemplars of their constituents. They correctly predicted

assumption is that a capability to parse a sentence into words has been acquired. This also is a difficult acquisition problem, but the assumption that words can be largely identified correctly before syntax is acquired is plausible. For example, the units the child produces in the one-word stage are correctly identified words.

greater variation in duration for frequent syllables than for infrequent syllables.[5] One of the aims of the research presented in this article is to elucidate, using an exemplar model, the underlying mechanism which accounts for the effect found by Schweitzer and Möbius. In other words, is it possible to account for the variation in syllable duration across syllable frequency categories (frequent and infrequent) using a dual-route computational process? The model is intended to represent a competition between syllables accessed as units and those that are produced as a result of accessing the exemplar clouds of their constituent segments. The assumptions and predictions of the model are as follows:

- The model assumes the ability to parse a syllable into its constituent segments. Similarly, the model assumes the ability to compose a syllable from segments.

- The model assumes, on the basis of the evidence discussed above, the possibility of a dual-route production mechanism where syllables and segments are stored and accessed separately and operate independently.

- The model predicts a thresholding effect whereby until a particular threshold is reached, through frequent exposure to co-occurring segments, the duration of a particular syllable will reflect the sum of the durations of its constituent segments. Once the threshold is passed, the constituent segments will tend to behave more as a single unit, i.e. a syllable.

General Procedure

The model is initially seeded with segment exemplar duration values using the mean duration values of the segments in the corpus (discussed below). It is important to note that these mean values are calculated from the corpus as a whole, not merely from the duration of segment tokens

[5]Note that Schweitzer and Möbius (2004) found that z-scores of frequent syllable durations were more variable than z-scores of infrequent syllable durations. This is interpreted here to mean that frequent syllables are more variable in duration than infrequent syllables.

found in the pertinent frequency bins. It is also worth noting that the initial seeding of the model, using mean duration values, will impart an initial bias to the model. However, this is cognitively plausible as an infant must, after all, have a *first perception* of each segment. The decision to use a mean value is simply to initialize each segment cloud with what could be considered a *typical* or *plausible* exemplar. Furthermore, this cognitively plausible initial bias is not equivalent to prototypicality as exemplar selection in the model is random. This is also plausible because no exemplars (segments or syllables), initially at least, should appear to have a better "fit" to a given context since an infant has limited contextual experience, i.e. there should be no preferential or biased selection, because an infant does not "know" what is a good or bad exemplar. However, over time, as a cloud acquires more exemplars the denser parts of the cloud will be more likely to be "activated" simply because of the nature of the distribution and the random selection, i.e. an exemplar from a denser part of an exemplar cloud is more likely to be selected than an exemplar lying in the outskirts under random selection.

On each iteration a syllable, either frequent or infrequent, is selected for production according to two competing production pathways known as a *composite* production pathway and a *unit* production pathway, in keeping with the dual-pathway concept discussed above. According to the composite pathway a duration for each of the syllable's constituent segments is randomly selected from their respective exemplar clouds and random noise is added, to reflect motor/articulatory perturbations. This syllable is known as a *composite* syllable. Its duration is the sum of the durations of its constituent segments. In parallel, a nominally identical *unit* syllable is also selected for production. A duration for this syllable is randomly chosen and modified by noise, commensurate with the duration of the syllable, again to reflect motor/articulatory perturbations. The unit syllable has an associated activation based on the density of its cloud. If the unit syllable reaches a certain minimum activation it is chosen for production, otherwise the composite syllable is chosen.

The winner is then produced, perceived, and stored. The next two simulations employ this general procedure using a) modeled data, and b) the same corpus data used by Schweitzer and Möbius (2004), in order to reproduce their syllable variability findings.

The addition of random noise to reflect motor/articulatory perturbations is not unreasonable, at least as a first approximation. To add noise determined by examining a corpus of adult speech would equate to modeling adult productions in infant speech. Given the inexperience and the articulatory difficulties that infants have, noise introduced by an infant is unlikely to correspond well with noise produced by a physiologically mature and vastly more experienced adult. It would, of course, be desirable to have a more faithful model of infant production noise, but at this point reasonably constrained random noise is employed.

From a different angle one might argue that random noise is inadequate because production factors that influence the duration of one segment are likely to have a similar influence on other segments within the same structural unit. Speaking rate is an obvious example of such factors. Whereas it seems plausible to assume directional invariance, i.e. all segments are shortened at higher speaking rate and lengthened at lower rates, albeit by different proportions, there is evidence for complicated interactions at the foot level. For instance, compensatory shortening of unstressed syllables is typically observed in stress-timed languages, yielding a reduction of syllable durations as a function of an increasing number of syllables in the foot (e.g., Fowler, 1977, for English; Kohler, 1983, for German; and Eriksson, 1991, for Swedish). However, compensatory shortening as well as foot-final lengthening does not seem to occur at high speaking rates in German (Wagner, 2008), indicating that the effects of speaking rate within the foot structure are not necessarily unidirectional.

What happens to segment durations as a function of changes in speaking rate is not well understood. Only very few quantitative studies have investigated the effects of speaking rate on

segmental durations within the syllable. The relative lack of success of rhythm models based on the relation between consonantal and vocalic intervals, which is unstable under changes of speaking rate (Dellwo & Wagner, 2003), may be taken as evidence for a non-uniform effect of speaking rate on different phone classes within the same structural unit. Thus, it is quite possible that segments belonging to the same syllable vary in their durations in opposite directions, whereas the syllabary concept will of course assume that segments in a syllable *unit* will co-vary in the same direction.[6]

Simulation 1 - modeled data

Stimuli. Stimuli were syllables of the form CVC where C was one of five consonants and V one of five vowels (for a total of 125 syllables). For each segment (phone) the acoustic properties are modeled as a randomly generated two-dimensional vector, and the duration value stored in a single dimension. The similarity of two segments or constituents was computed as the sum of the similarities of their acoustic vectors and their durations. For vector similarity, the cosine was employed, for duration similarity an exponential transformation of difference:

$$\mathrm{sim}(\vec{v}, \vec{w}) = \frac{\sum_i v_i w_i}{\sqrt{\sum_i v_i^2}\sqrt{\sum_i w_i^2}}$$

$$\mathrm{sim}(x, y) = e^{-\alpha(|x-y|)}$$

where x and y are durations and $\alpha = 0.05$. α was chosen to give good sensitivity for typical lengths of consonants and vowels. Durations of syllables in the seed set were chosen to be 280 ms,

[6]Nevertheless, the authors acknowledge the possibility that the composition of segments under particular production factors might vary their duration in the same direction, possibly yielding higher variability in infrequent syllables than frequent ones, however this area has been little examined in the literature, and related work (coarticulation effects) by Benner et al. (2007) found a tendency towards stronger coarticulation and greater coarticulatory variability in high-frequency syllables than in low-frequency ones, which is compatible with their hypothesis (related to ours) that high-frequency syllables are retrieved from a syllabary and are less resistant to coarticulation than syllables assembled from segmental specifications.

distributed in a ratio of 1:2:1 over the three constituents CVC. These numbers were chosen because 70 ms is a typical duration for a consonant and 140 ms is a typical duration for a vowel. The 125 syllable types were randomly assigned to either the frequent or the infrequent subclass.

Procedure

The unit exemplar database was seeded with 500 syllables. It is important to note that in all instantiations of the model, when a unit is added to the unit database, its constituents are simultaneously added to the constituent database, to reflect the fact that the segments (and words in syntax) are perceived too.

A total of 5000 iterations of a production-perception loop were performed. Each iteration consists of randomly picking one of the 125 syllable types. The probability of selecting a frequent type is 100 times that of an infrequent type. For the constituents of both infrequent and frequent syllables, acoustic vectors are generated (slightly perturbed, using uniform noise, from the canonical vector of a consonant or vowel to reflect variation in (planned) articulation). The syllable's and constituents' nearest neighbors in the unit and constituent databases respectively are retrieved, within a fixed radius. If activation in the unit database is below the threshold θ (i.e., there are fewer than θ exemplars in the cloud), then the unit cloud is discarded, and the three neighborhoods in the constituent database are employed instead. The target duration of an exemplar is inferred to be the average duration of the members of its cloud. Finally, random noise proportional to the computed duration is added. For radius parameters and for θ, values were selected that separate frequent and infrequent syllables (see section 6).

After the syllable with the inferred duration has been produced, it is added to the exemplar database. This part of the procedure models a production-perception loop, either on the individual or the community level: every *produced* exemplar becomes a *perceived* exemplar after its production.

The final phase of the procedure consists of probing the model, in an identical manner to the

initial 5000 iterations, with 10 syllables of each of the 125 syllable types. The standard deviation for the syllable type is then computed on just this sample of 10 per syllable type. In this phase, syllables and their units are deleted after each probing to make sure that infrequent syllables do not change their status to frequent in the probing phase.

Results

Fig. 2 is a cumulative histogram of 10 runs of the above experiment, corresponding to 1250 standard deviations. The model successfully simulates the finding of Schweitzer and Möbius (2004): frequent syllables are more variable in duration than infrequent syllables. This result was significant ($p < 0.001$, Welch Two Sample t-test on 1250 syllables).

The difference in variation arises from the interaction between the two submodels. Frequent syllables have enough density, so that their duration is computed in the unit model, with noise added that is proportional to the length of the syllable. Infrequent syllables are compositions of constituents that are computed in the constituent model, each with independent noise. Given that each composite syllable is composed of segments to which noise has been added, it is likely that the net effect on the syllable duration will be small as the addition of noise to each segment will to some extent result in a cancelation effect. In other words one segment might grow longer whereas another grows shorter. Over many iterations of the production-perception loop, frequent syllables become more variable in duration whereas the variability of infrequent syllables does not change much. Given the success of this simulation using artificial data the next phase is to apply the model to the data in the corpus employed by Schweitzer and Möbius.

Simulation 2 - corpus data

The corpus is a single-speaker speech database for unit selection speech synthesis, recorded by a professional male speaker of German, and contains approximately 160 minutes of speech (2601

utterances with 17,489 words, 33,800 syllables and 94,300 segments). The corpus was manually annotated on the segmental, syllabic, word and prosodic levels. In their experiments Schweitzer and Möbius extracted the 326 most frequent syllable types, each with more than 20 tokens. In total this accounted for 22,638 syllable tokens, covering approximately 67% of the corpus. These syllable types were matched against categorization criteria for frequency and infrequency based on analysis by Müller et al. (2000), the criteria being that very infrequent syllables have a probability of less than 0.00005 and very frequent syllables have a probability in excess of 0.001. This further refinement yielded 114 very frequent and 16 very infrequent syllable types. Using identical frequency bins to those employed by Schweitzer and Möbius allows for comparison between their results and the results of the experiment detailed below.

Procedure

As in the first simulation, a *composite* and *unit* syllable are selected for production in parallel. In this particular simulation however, as the main topic of interest is the duration dimension, accurate categorization of the syllable is assumed and the acoustic properties are not investigated. Given the balanced nature of the corpus the syllables modeled here are of a variety of forms, including CV, CVC and more complex structures. For the purposes of illustration the model is discussed from the perspective of CVC production. Fig. 3 and Fig. 4 provide an illustration of the *modus operandi* of the model.

Initialization. The model is initially seeded with exemplar duration values using the mean duration values of the segments in the corpus. The unit exemplar database is initially seeded with a syllable whose duration equals the sum of the values of the initial constituent segment seeds. In other words, the model is seeded with a unit syllable, and constituent segments, of plausible duration (Fig. 3 top panel).

On each iteration of the model program a syllable (frequent or infrequent) is selected for production and dual-route competition takes place as follows.

Composite selection. According to the composite pathway each iteration of the model selects a CVC syllable for production. A duration is then randomly selected for each constituent segment from their respective duration clouds (Fig. 3 middle panel). Random noise is then added to each duration value as follows (where e_{rd} corresponds to random production noise commensurate with the duration of an extant exemplar, e.g the consonant segment in onset position $C1_{ex}$, which is added to the duration of this exemplar):

$$C1_{dur} = C1_{ex} + e_{rd} \tag{1}$$

$$V_{dur} = V_{ex} + e_{rd} \tag{2}$$

$$C2_{dur} = C2_{ex} + e_{rd} \tag{3}$$

The net effect is that each segment is either lengthened or shortened (by up to 5%)[7] depending on the effect of the noise. Once again introduction of noise is intended to reflect the high degree of variability in speech production. The duration of the *composite* syllable is the sum of the segment durations.

$$CompSyll_{dur} = C1_{dur} + V_{dur} + C2_{dur} \tag{4}$$

Unit selection.

In parallel, according to the unit pathway, with each iteration of the model a *unit* syllable nominally identical to the composite syllable is also selected for production. The duration of the unit syllable is selected randomly and noise is added in a similar fashion to that described for the segments. As with the segments it is important to note that the level of additional noise is commensurate with the size of the unit (Fig. 3 middle panel).

[7]Similar results were achieved using less conservative noise estimates.

Competition, production, and storage.

Thus at this point in the execution of the model there are two competing syllable hypotheses, one *composite* and one *unit*. The determining factor in deciding between the two is the level of activation α of the unit syllable. In this simulation the model initially has one exemplar of a syllable stored as a unit. Thus the unit syllable activation is 1.

In the event that the unit syllable possesses an α value less than a unit-threshold θ, the durations of the composite syllable's constituent segments are stored in their respective duration clouds, and the composite syllable's total duration is stored in the duration cloud of its unit syllable equivalent. This represents the production, and perception, of a syllable using its constituent segments rather than accessing the syllable as a unit stored in exemplar memory. That is, the indirect segmental assembly route is chosen (Fig. 3 bottom panel, assuming a θ value greater than 1, since α=1 in this example). This indirect route is taken until the threshold is exceeded (Fig. 4, top and middle panels). On the other hand, if the activation of the unit syllable is greater than the threshold, then its duration is stored in its exemplar cloud and its duration mass is divided into the duration clouds of its constituent segments in a manner corresponding to their typical influence on syllable duration (Fig. 4, top, middle and bottom panel). This is achieved through the following normalization process:

$$Seg_{i\_weight} = \frac{Seg_{i\_mean}}{\sum_{j=1}^{n} Seg_{j\_mean}} \tag{5}$$

$$Seg_{i\_duration} = UnitDuration * Seg_{i\_weight} \tag{6}$$

where n corresponds to the number of segments in the syllable. The reason for the distribution of the unit syllable's duration mass to the segments which would normally compose it (despite the fact that it is operating as a unit) is that intuitively, from an exemplar-theoretic perspective, the production/perception of a syllable should result in greater activation of that syllable's exemplar cloud, and in greater activations of the clouds of the segments that can be perceived within the

syllable. The size of all the exemplar duration clouds increases over multiple iterations of the model.

Experiment

The algorithm described above was executed to yield n productions per syllable based on the prior syllable probabilities presented in the previous section. This was performed with a view to establishing a critical mass of durations which would facilitate inspection of the model.[8] Once these n productions per syllable were complete a further 500 inspection iterations were performed, per syllable, using the enlarged clouds provided by the pre-inspection execution. In other words the model was forced to produce, on the basis of duration results yielded by the pre-inspection execution, 500 unit syllable tokens per syllable type (across both frequent and infrequent types), and 500 composite syllable equivalents. The threshold θ was manually set to 100 (see section 6). As with the first simulation, these inspection durations did not enlarge the original pre-inspection duration clouds, that is the resulting duration clouds for the frequent and infrequent syllables were stored separately. Otherwise, if syllables produced in the inspection phase were stored with the pre-inspection syllables this could entail that an infrequent syllable could acquire enough mass to become frequent and the inspection would be unreliable. The purpose of the inspection phase is to determine what the trained model will produce, the model itself should not be interfered with.

The duration z-score of each syllable token was calculated and plotted against the mean z-scores of the involved segments in the composite equivalent, where the z-score of a token (syllable or segment) is given by:

$$z = \frac{d - \mu}{\sigma} \tag{7}$$

[8] An inspection of the model by random generation of tokens according to their prior probabilities is computationally expensive and unnecessary. For example, if the most infrequent syllable has a probability of 0.0001, then in order to produce 100 tokens to establish a critical mass, a corpus of millions of tokens would have to be generated and most of the data would not be relevant to the experiment.

where z is the z-score of the token, d is its duration, μ is the mean duration of the token's type and σ is the standard deviation of the duration of the token's type.

A linear regression model was fitted to the data, with the duration z-scores of the syllable tokens acting as the dependent variables. Aggregate results are presented for both frequent and infrequent syllables in Fig. 5.

Results and Discussion

As a function of the fact that the infrequent syllable is not produced in sufficient numbers to exceed the threshold, the z-scores of each infrequent syllable will, to a significant degree, reflect those of the constituent segments as the durations in the syllable cloud will essentially be the durations of composite syllables. Given that each composite syllable is composed of segments to which noise has been added, it is likely that the net effect on the syllable duration will be small as the additions of noise to each segment will to some extent cancel each other out. In other words one segment might grow longer whereas another grows shorter. Overall the standard deviation (on which z-scores depend) of the composite syllable will not be that large, and, since the duration cloud of an infrequent syllable relies heavily on composite syllable durations the standard deviations of infrequent syllable durations will be similarly restricted. Hence the z-score of an infrequent syllable will be quite well predicted by the z-scores of the syllable's segments (highly significant correlation between segment z-score and syllable z-score: $F = 30{,}580$, $p < 0.001$, t-test), as illustrated in the linear regression model plotted in Fig. 5 (right panel).

The syllables of the frequent category however, exhibit greater variability (no significant correlation between segment z-score and syllable z-score: $F = 2.785$, $p = 0.095$). This is due to the fact that after a number of productions they possess activations greater than θ and hence much of the durations in their clouds correspond to previous *unit* durations with added noise. The noteworthy point here is that the noise is added to the syllable unit as a whole, not to its

constituent parts. Thus there is no cancelation effect, and the syllable unit is more likely to vary more significantly from the mean.

This result is in keeping with the results of the first simulation and the findings of Schweitzer and Möbius (2004). Their results are illustrated in Fig. 6. What is particularly striking is that the results for the frequent case in the current experiment (Fig. 5 left panel) appear to be noisier than those established by Schweitzer and Möbius (Fig. 6 left panel). This is likely due to the fact that, apart from the initial segment seeds, all the exemplars in the current model are the result of self-productions and self-perceptions, i.e. the corpus is only employed for the purposes of initialization and the exemplar clouds become denser on the basis of the model's subsequent productions. Of course this is somewhat limited as an infant will continue to perceive exemplars other than its own. However, estimating how often an infant might perceive external exemplars of a particular segment relative to perceptions of self-produced exemplars of the same segment is problematic, and establishing reasonable estimates is currently being investigated. It is anticipated that if satisfactory estimates can be arrived at, one would expect to see results for the more frequent syllable case corresponding more closely to those found by Schweitzer and Möbius, as the exemplar clouds would then represent a mixture of externally perceived and self-produced exemplars. Nevertheless, it is important to note that the overall prediction of the model is borne out.

Similarly, the results are in line with the dual-pathway theory posited by Levelt et al. (1999), Whiteside and Varley (1998) and Mayer et al. (2003) among others. Employing the frequency bins specified by Schweitzer and Möbius (2004) allows for evaluation of the current results in the context of their findings. However, such frequency bins are not required for the model to work. For example, a syllable with mid-range frequency should exhibit less variability than a high frequency syllable as it would not have so many post-threshold exemplars in its syllabary. It would nevertheless exhibit

more variability than an infrequent syllable which barely gets above threshold level activation.

In both simulations the same general competitive interaction model, operating across both the unit and constituent levels, was employed. The next section demonstrates how the same model can account for the development of local grammatical knowledge.

5. Modelling the acquisition of local grammatical knowledge

One of the basic tasks children master when acquiring a language is to distinguish between grammatical and ungrammatical sentences. Rote learning is of limited help in judging grammaticality because of the productivity of language. This section demonstrates that emergent acquisition of local grammatical knowledge can be modeled in the multi-level exemplar model via local grammaticality judgments which are formalized as activation of a sentence as a unit. When applying the model in Fig. 1 to syntax acquisition, the input is a sentence. The sentence is processed as a unit (left half of the figure) by searching the unit exemplar database (i.e., a database of sentence exemplars) for similar sentences. In parallel, the constituent exemplar database on the right is used to retrieve clouds of similar exemplars for the words (or constituents) of the sentence. Words are shown as x, y, and z. If the sentence receives activation α in the unit exemplar database that exceeds a threshold θ, then the syntactic structure of the input sentence is constructed to be analogous to the syntactic structure of the set of similar sentences retrieved. For example, if for the stimulus *Peter loves coffee*, the most similar retrieved sentences are *Mary likes tea* and *John craves chocolate*, then the syntactic relationship between *loves* and *Peter* is construed to be similar to that between *likes* and *Mary* and to that between *craves* and *John*. Analogously, the syntactic relationship between *loves* and *coffee* is construed to be similar to that between *likes* and *tea* and that between *craves* and *chocolate*.

On the other hand, if the sentence does not receive sufficient activation, then there are no sentences that can be used to analogize the input stimulus to. In that case, the sentence is

perceived as a concatenation of syntactically unconnected words, that is, as a sentence unspecified with respect to grammaticality. The composition function merely joins the words into a sequence and does not add any information about how the words relate to each other (e.g., that one word is the object of another) since that information is not available due to insufficient unit activation.

This section illustrates how the model can handle grammaticality via local syntactic coherence and, in a comparative experiment, demonstrates the benefits of employing exemplar representations over categorical ones. To begin, the challenge posed by the acquisition of local syntactic patterns is discussed, followed by a motivation of the representation of words in the multi-level exemplar model.

Difficulty of acquiring local syntactic patterns. What does it mean to possess a native-like mastery of the grammar of a language? The key test of successful acquisition is productivity. Memorization and successful retrieval from memory by themselves do not demonstrate that any nontrivial learning has taken place.

There are several different aspects of syntactic productivity. At the most basic level, productivity means the production of novel sentences, i.e., the ability to produce a sentence that was never part of the input. Generative grammarians have traditionally focused on the most complex form of syntactic productivity, the production of a (usually recursive) pattern of constituents that has never been experienced; for example, the production of four levels of embedded relative clauses if only up to three have been experienced. The acquisition of morphology is also closely tied to the acquisition of syntax, but this paper will only be concerned with syntax. Finally, the ability to infer the correct syntactic properties of new words is also a task that the learner of the syntax of a language has to solve. For example, a native speaker of English knows that the word *foobar* in the utterance *the foobar in the bag* is most likely a noun.

In this paper, the most basic form of syntactic productivity is addressed, the ability to distin-

guish syntactic environments in which a word can occur from those that are not grammatical. In the simplest case, this amounts to learning the part-of-speech patterns of simple English sentences, e.g., of two-word sentences in the early stages of language acquisition.

All of these part-of-speech patterns are very frequent in the child's input. If this input was presented to a child in grammatically annotated form (indicating, e.g., that *butter* in *give me the butter* is a noun), then the task of learning legal part-of-speech patterns would be a trivial task of memorization. But since the child perceives words without labels, the acquisition of the correct part-of-speech patterns of English sentences is a difficult task. Note also that simple heuristics like *If* w_1 *and* w_2 *are encountered in the same context* c_1*, then that means that they can be substituted for each other in any other context* c_2*.* are not viable. Simple heuristics fail because of ambiguity. An example where the heuristic just given fails is w_1 = *fun*, w_2 = *trouble*, c_1 = *I've never had so much* _, and c_2 = *That's a really* _ *game.* The sentences $c_1(w_1)$ and $c_1(w_2)$ are both grammatical, but only $c_2(w_1)$ is good whereas $c_2(w_2)$ is not.

One of the major obstacles facing an exemplar theory of syntactic development is to account for how a child begins by hearing syntactically unlabeled utterances, yet at some point becomes able to infer what the syntactic properties of a particular word are and to then produce this word in novel contexts, e.g., to use *nutella* in *give me the nutella* even though *nutella* was never experienced in this particular context. When acquisition is completed, the child represents sentences in a partially abstract representation that enables him or her to comprehend and produce novel combinations. It is this process which this section of the paper attempts to explain.

A number of researchers have attempted, with considerable success, to model aspects of language acquisition such as category learning (Elman, 1990; Redington, Chater, & Finch, 1998; Mintz, 2003), the emergence of abstract syntactic structures and how they combine (Morris et al., 2000), structural priming (Chang et al., 2006) etc. The syntactic instantiation of the model

presented here differs in that it focuses both on mathematically and computationally formalizing the idea of local grammaticality and on establishing the benefits of graded representations over categorical ones.

Representation of words. The similar syntactic behavior of two nouns like *coin* and *hen* is not directly apparent from their pronunciation or semantics. But an exemplar-theoretic account of syntactic behavior requires a similarity measure where *coin* and *hen* are similar. Building on the ideas described in (Schütze, 1995) (see also Schütze, 1992, and Schütze, 1993), the left-context and right-context components of the representation of a given focus word are defined, where the left (right) context consists of a probability distribution over all words that occur to the left (right) of the focus word and the dimensionality of the distributional vector for each word is dependent on the number of distinct words (or word types). For example, having experienced *take coin* twice and *drop coin* once, the left context distribution of *coin* is $P(\text{take}) = 2/3, P(\text{drop}) = 1/3$. That is, each word can be treated as comprising two *half-words*, a left half-word that characterizes the word's behavior to the left and a right half-word that characterizes the word's behavior to the right. Thus, the phrase *the red hen* would be represented as the six half-words: the_l, the_r, red_l, red_r, hen_l, and hen_r. Table 2 presents right context distributions for three half-words. Thus, for example the probability that *like* is followed by *people* (according to some arbitrary corpus) is 0.0054. The similarity between two half-word distributions is arrived at by calculating their cosine, thus, the similarity measure used in the first phonetics simulation is employed here too.[9]

Before presenting formal definitions of exemplar-theoretic and category-based local syntactic coherence and a comparative experiment, the following *proof-of-concept* simulation illustrates how

[9]The context distributions used here are *first order* in the sense that only directly observed neighbors are used. For example, if noun w_1 has only been experienced after the definite article and noun w_2 has only been experienced after the indefinite article, then the similarity of the left contexts of w_1 and w_2 may not be recognized in a first-order model. See (Schütze & Walsh, 2008) for an extension of the model that takes second order effects into account.

the exemplar model employs the half-word representations described above to make local grammaticality judgments.

Simulation 3 - modeled data

Stimuli. Using 10 different verbs, 10 different nouns and 10 words with a verb-noun ambiguity, all grammatical sentence types of the form *I verb noun* (e.g, *I love coffee*) were generated. There are a total of 400 types because there are 20 nouns (including the 10 ambiguous words) and 20 verbs (again, including the 10 ambiguous words).[10] A random subset of 100 of these sentence types was selected and assigned to the subclass *unattested*. The remaining 300 were assigned to the subclass *attested*. In addition, 100 ungrammatical types of the form *I coffee love* were also generated, but only using unambiguous words, i.e., the 10 unambiguous words were not used for generating ungrammatical sentences. Finally a training set of 100,000 training sentences was randomly generated from the attested subclass. The generation was biased towards unambiguous sentences in a ratio of 3:1, that is there were 3 times as many sentences that did not contain an ambiguous word than sentences that did.[11]

Procedure

The training set was stored as the exemplar database of the model. The left and right half-word distributional vectors were calculated for all words in the training corpus. An instance of each of the 100 unattested and of the 100 ungrammatical sentences was then presented to the model as a probe and their activations were calculated using a similarity measure. For example, in order to ascertain the activation of the probe sentence *I love coffee*, it is compared against all the sentences stored in the model's exemplar database. For a given exemplar sentence *You like tea* the

[10] The words are unanalyzed symbols for the model, any similar set will produce analogous results

[11] Since a random generator was used, the actual numbers deviate from a ratio of 3:1: 77,495 without an ambiguous word, 22,505 sentences with an unambiguous word.

comparison operates as follows:

1. The similarity of the *I* and *You* left half-words is calculated using the cosine.

2. In the same way the similarity of their right half-words is calculated.

3. These two results are summed to form a similarity value for these two constituents.

4. The same process is applied to the pairs *love* and *like*, and *coffee* and *tea*.

5. A boundary symbol ("_b_") was introduced and represented as a word in order to provide a representation for the beginning of a sentence and for the end of a sentence. In terms of the representation of a sentence as a sequence of half-words, this means that each sentence begins with the right half-word $\_b\__r$ and ends with the left half-word $\_b\__l$. The boundary symbol captures the intuition that knowledge about whether words can occur at the beginning or end of sentences is also part of the learned representation of a word. For example, *the* cannot occur at the end of a sentence.

6. The similarity scores for the three pairs and the two boundary half-words are then summed to produce an overall similarity score for the probe sentence and the stored exemplar. If the similarity is such that it can be classed as activated with respect to a radius threshold value then it contributes to the level of activation associated with the probe.

How realistic is it to compare a stimulus to all exemplars in memory? This is a general problem in exemplar theory. Clearly, going sequentially through all stored exemplars and comparing each to the stimulus is not feasible if realistic processing times are to be achieved. The brain is a highly parallelized processor, but even taking parallelism into account, it is unlikely that each stored exemplar is reexamined every time a new stimulus is encountered. The general mechanisms by which this computation could be made more efficient are not addressed here but clearly this is a problem that exemplar theory will have to account for (Grossberg, 2003; Hintzman, 1986). The authors' view is that whatever general mechanism is capable of performing the required computations efficiently

can also be employed for computing local syntactic coherence.

Fig. 7 shows a histogram of similarities for the 200 test sentences. For each of the 100 ungrammatical and the 100 unattested sentences its similarity to the closest sentence in the training corpus was computed. The histogram shows the distribution of these 200 similarities. Ungrammatical sentences (similarities < 7.0) and unattested sentences (similarities ≥ 7.0) are perfectly separated.[12] The simulation successfully models the acquisition of local grammaticality because ungrammatical sentences are dissimilar to grammatical sentences due to different left and right contexts. An example for the latter is that when comparing *I love coffee* with *I tea drink*, the left context of *love* (containing the subject *I*) is very different from the left context of *tea* (consisting of verbs like *love*, *drink* and *make*). Given that grammaticality judgments are based on at least implicit knowledge of syntactic behavior it can be concluded that the model has to some extent, in an emergent fashion, acquired local grammatical knowledge. Although the learning taking place here is with respect to a small subset of English, generalizing to richer left and rights contexts is not problematic as will be shown in the next section. In addition, it is important to note that, as with the previous phonetics experiments, the same model of unit and constituent interaction is employed here. That is, frequently co-occurring segments/words give rise to more autonomous higher level units. Furthermore, as with the phonetics experiments, given the success of the simulation using hand-crafted data the next step is to apply the model to data that are representative of the kind of language input a child receives and, to place this model in the context of previous research, to

[12]The ambiguous words in the experiment were introduced to make the histogram in Fig. 7 non-trivial. If only unambiguous verbs and nouns occur then all grammatical sentences receive a score of $> 8.0 - \epsilon_1$ and all ungrammatical sentences receive a score of $< 4.0 + \epsilon_2$, where ϵ_1 and ϵ_2 are small numbers. The score 8.0 is the sum of 8 half-word cosine scores of close to 1.0, the six half-words of the sentence and the left and right boundary half-words. The score 4.0 is the sum of 4 half-word scores close to 1.0 (the half-words of the left and right boundaries and the two preposition half-words) and 4 half-word scores close to 0.0 (the two half-words of the verb and the two half-words of the noun).

compare it in the same task against a category-based approach. First, however, a more rigorous description of local syntactic coherence is presented for the multi-level exemplar model, and for a category-based model.

Local exemplar-theoretic coherence. A sequence of half-words $h_1, \ldots, h_n$ exhibits local exemplar-theoretic coherence if, and only if, there exists a sequence of half-words $g_1, \ldots, g_n$ in memory such that

$$\sum_{i=1}^{n} \text{sim}(h_i, g_i) > \rho$$

where ρ is a parameter acting as a radius.

$\text{sim}(h_i, g_i)$ is the similarity between the relevant distributions of half-words h_i and g_i. For example, given a stimulus sequence *by the blue lake* and an exemplar in memory *beside a grey mountain*, one would expect that the overall similarity between the corresponding half-words, i.e.

$$\text{sim}(by_l, beside_l) + \text{sim}(by_r, beside_r) + \text{sim}(the_l, a_l) + \text{sim}(the_r, a_r) + ... + \text{sim}(lake_r, mountain_r)$$

might well be higher than the similarity threshold ρ as each word (and half-word) would have reasonably similar distributions. Hence the stimulus sequence *by the blue lake* could be described as locally coherent. In order to determine the similarity between distributions in the simulations below, the cosine was calculated. It is important to note that the radius parameter ρ gives rise to a precision-recall trade-off. A large ρ will impose stringent requirements on which sequences in memory match, resulting in false negative decisions for local grammaticality. A small ρ will incorrectly judge many locally incoherent sequences to be grammatical.[13]

Note that coherence as defined here is discrete. A sentence is either coherent or incoherent.

[13] ρ is formally different from θ because it defines the radius of an exemplar cloud using a similarity value. In this section, θ is always 1. If there is at least one exemplar in the exemplar cloud defined by ρ, that is, if $\alpha \geq 1$ in Fig. 1, then $\alpha \geq \theta$ and there is sufficient activation for perception based on the unit pathway. λ is the counterpart to ρ for the category-based model.

It would be desirable to generalize this to a gradient notion of coherence to account for cases of intermediate grammaticality. See Section 6 for discussion.

Category-based representations. Several previous papers have worked with variants of left and right context vectors in a categorical framework. For example Mintz (2002), in an artificial language learning study, provides psycholinguistic evidence that adults avail of a word's immediately adjacent neighbors to help categorize it. Subsequent work (Mintz, 2003), employing corpora of child-directed speech, demonstrates that robust word categorization results can be achieved by grouping words which lie within the same frequently occurring frame, where a frame is any two words that co-occur in a corpus with a single word intervening between them. Monaghan et al. (2007) in a number of experiments across English, Dutch, French and Japanese found that both the preceding and succeeding words around a target word were strong indicators of the target word's class, e.g., for English, three preceding words, *he*, *we* and *to*, were better indicators of verbs than nouns.

The work by Redington et al. (1998) is perhaps the best known representative and the experiments below are based on their method. It is important to note that although the use of half-word distributional vectors of context-based information is reminiscent of Schütze (1995) and Redington et al. (1998), the approach here differs from theirs (and others, e.g. Elman, 1990; Morris et al., 2000) in that it is situated in an exemplar-theoretic framework where grammaticality in a particular syntactic context is the object of interest rather than clustering of words within syntactic category types. In addition, at the core of Exemplar Theory is the idea that linguistic generalizations are achieved on the basis of similarity between a novel stimulus and exemplars residing in memory, where context is a crucial component. Redington and colleagues do not speculate as to how the word clusters they induce could be used in syntactic processing and essentially view words as abstract objects divorced from their contexts. The task they evaluate on is a clustering task:

Words are clustered into groups based on left and right context vectors.[14] The clustering method is agglomerative hierarchical clustering, which generates a hierarchy of clusters with a large cluster consisting of all words at the top and with leaves consisting of individual words at the bottom. This hierarchy is then transformed into a set of disjoint clusters and compared with linguistic judgments. For example, if the articles *the* and *a* are not put in the same category, then this is judged an error. For judging local coherence in a categorical way, the following definition is adopted.

Local categorical coherence. A sequence of half-words $h_1, \ldots, h_n$ exhibits local categorical coherence if, and only if, there exists a sequence of half-words $g_1, \ldots, g_n$ in memory such that

$$\sum_{i=1}^{n} \text{common\_path}(w(h_i), w(g_i)) > \lambda$$

where $w(h)$ is the word that h is one half of and common_path(w_1, w_2) is the length of the longest common path that starts at the top of the hierarchy and ends at a node that is a parent of both w_1 and w_2.

The greater the similarity between the distributions of two words, the later the words are separated into two clusters, that is, when moving from the top of the hierarchy to the leaves, they end up in different clusters towards the bottom of the hierarchy, corresponding to a long common path.

For example, common_path$(w, w) = n - 1$ (where n is the number of words) for all words w; common_path$(w, v) = 18$ if the lowest cluster c in the hierarchy that contains both w and v is at level 18 of the tree, that is, of the two subclusters of c at level 19 of the tree one contains w and the other v. To elucidate further, given the half-word stimulus sequence "by_r, the_l", and an extant

[14] (Redington et al., 1998) also looked at a more complex representation than that employed here, one that is based on words that are not directly adjacent to the target word. However, they did not find an appreciable difference in categorization accuracy between the two variants.

exemplar sequence "$beside_r, a_l$", then

$$\text{common\_path}(by_r, beside_r) + \text{common\_path}(the_l, a_l)$$

will have a value of $26 = 12 + 14$, if for example the words *by* and *beside* were separated when moving from level 12 to level 13 of the tree and *the* and *a* were separated when moving from level 14 to level 15 of the tree. If the sum total is more than a threshold value λ then the stimulus sequence "by_r, the_l" has local categorical coherence.

According to this categorical definition, a sequence is viewed as being locally grammatical if for a particular clustering, a sequence with the same categories can be found in memory where "same" is defined with respect to the agglomerative clustering hierarchy and the threshold λ. It is obvious that with very few clusters, almost all sequences will be locally grammatical (including sequences that are clearly not grammatically well-formed) while with a large number of clusters, it becomes very hard to find identical sequences – even for a sequence that is perfectly well-formed. In addition, it is particularly important to note that half-word representations are only employed during the clustering process (essentially the training phase). In the simulation presented below the comparison between a stimulus sequence and a previously stored sequence is carried out using common path values. In other words, the key difference between local exemplar-theoretic coherence and local category-based coherence is that the latter is determined using a more abstract representation, the membership of words in clusters as defined by an agglomerative hierarchy.

Given the definitions for both local exemplar-theoretic and local category-based syntactic coherence for *sequences*, the following definition at the level of the *sentence* is required.

Local syntactic coherence of a sentence. A sentence can be said to possess local syntactic coherence if all of its subsequences are locally coherent. Local syntactic coherence is a function of the definition of sequence coherence used (exemplar-theoretic or category-based) and parameters n (the subsequence length) and ρ (exemplar-based similarity threshold) or λ (category-based similarity

threshold).

The distinction between sequences and sentences is important. In the experiment below, each time a stimulus sentence is tested against a model's memory (exemplar-theoretic or category-based) it is split into sequences of length n (a variety of lengths are investigated) and each sequence is compared to all sequences of the same length in memory. For each sequence, the model then identifies the sequence in memory with the largest similarity. To be conservative, the smallest of these largest similarities (i.e. the most ungrammatical) is treated as the result. For example, (assuming direct word comparison rather than half-word representations, purely for the purposes of illustration), given the phrase *I love coffee* and the stored memory *You like the big tree*, and sequence length of $n = 2$, then the stimulus subsequence *I love* will be compared against *You like*, *like the*, *the big*, and *big tree* and the largest similarity will be retained. The stimulus subsequence *love coffee* is then compared against *You like*, *like the*, *the big*, and *big tree* and the largest similarity is retained here too. If the smallest of these two retained similarities is larger than the similarity threshold (ρ for the exemplar-based model or λ for the category-based model), then the phrase *I love coffee* has sentence level local syntactic coherence as all of its subsequences are locally coherent.

Defining the local grammaticality of a sentence as the grammaticality of its least well-formed subsequence puts long and complex sentences at a potential disadvantage since they contain more subsequences of a given length that can be locally incoherent. However, the notion of local grammaticality is intended to capture the fact that neighboring words are syntactically compatible and that no rare or unusual combinations like *Peter bring* (singular noun + plural verb) occur. Of course, such combinations do occur in language due to long-distance dependencies, e.g., when a subject relative clause is followed by the main verb. The present model is limited to local grammaticality and can therefore not account for this type of long-distance effect. It is plausible that the ability to judge local grammaticality is acquired first and that the acquisition of more complex

syntax then builds on this capability. When only considering local grammaticality it is appropriate to insist that all subsequences be locally coherent – which justifies taking the smallest (least grammatical) value in the definition of local coherence.

Armed with definitions for both models and a formal definition for grammatical coherence at the sentence level the following comparative experiment was performed.

Simulation 4 - corpus data

The goal of this experiment is to establish the extent to which the multi-level exemplar model is capable of distinguishing between ungrammatical and real-world grammatical sentences and, motivated by the exemplar-theoretic idea that detailed context information should facilitate linguistic inference, to ascertain whether the exemplar-based model offers benefits which enable it to outperform a category-based model.

To test these hypotheses, the well-known CHILDES database (MacWhinney, 2000), a repository of transcripts and media documenting conversations between young children and their playmates, siblings, and caretakers, was employed. In order to avoid mixing varieties of English (e.g., British English vs. American English), the largest homogeneous subcorpus of CHILDES, the Manchester corpus, was selected. It contains almost 350,000 sentences consisting of more than 1.5 million words. This is a conservative estimate of the amount of child-directed speech a child would receive annually (Redington et al., 1998). All names in the corpus (i.e., all capitalized words) were replaced with a special word "_n_". Again, a boundary symbol "_b_" was introduced to begin and end sentences. The representation of the corpus is then a concatenation of all its sentences. The vocabulary consists of 8601 words.

Stimuli - Construction of the evaluation set. In order to test the ability of the two models to distinguish locally coherent vs. incoherent sentences, a total of 100 "unattested" sentences were

selected from the corpus and were not used to train the model. Only unattested sentences that were not a substring of a sentence in the training corpus were selected, since, presumably, any substring of a sentence in the training corpus is locally coherent. A further constraint was that the unattested sentence was not allowed to contain a word that did not occur in the training corpus. This constraint is motivated by the desire to address the problem of local syntactic coherence for known words only, since unknown words present special challenges.[15] Finally, each unattested sentence contained a word that occurred in only one sentence type in the training corpus. Early experiments indicated that local grammatical inference for frequent words is easy to determine as there is redundant evidence available that characterizes legal syntactic environments for frequent words – a frequent word's local syntactic context in the test set often occurs verbatim in the training set, so that an evaluation of local syntactic judgments for frequent words is mostly an evaluation of memorization (as opposed to true productivity). In contrast, rare words are a real challenge in language acquisition. For this reason, only those sentences that contained at least one rare word were selected as possible unattested sentences.

A set of 100 ungrammatical sentences was then generated by randomly selecting words from

[15] While the authors think that the acquisition of local syntactic coherence for known words should be addressed before turning to unknown words, they nevertheless wanted to verify that the model can handle unknown words in principle. To ascertain this, the words *the*, *in* and *bag* were represented in the probe sentence "the *unknown_word* in the bag" using half-words. This probe sentence was then compared with all 5-word sequences in the corpus, represented as $w_{1l}w_{1r}w_{2l}w_{2r}w_{3l}w_{3r}w_{4l}w_{4r}w_{5l}w_{5r}$, but, in contrast to the work presented in the main body of this article, w_{2l} and w_{2r} were not used to compute the similarity. The n-word sequences (n=5) that were most similar to the probe sentence were then determined. The words w_2 in these n sequences were: *baby*, *bag*, *bricks*, *camera*, *cards*, *food*, *lady*, *ones*, *orange*, *pieces*, *shopping*, *things*, and *top*. These are all words that could plausibly have the same part of speech as the unknown word appearing in the context "the *unknown_word* in the bag". This suggests that the model can infer the syntactic properties for unknown words based on sequences in memory, although working out a complete computational model of this process is nontrivial.

the vocabulary and concatenating them. This was accomplished as follows. For each unattested sentence, its length l was determined. Random sentences of the same length were then repeatedly generated until one was produced that met the conditions of (i) not being a substring of a sentence in the training corpus (and, in particular, not being identical to a sentence in the training corpus); and (ii) containing at least one rare word.

The experiment just described was repeated 20 times. Each run generated a training set of about 350,000 training sentences, a disjoint set of 100 unattested grammatical sentences and a set of 100 ungrammatical sentences.[16]

Unlike the proof-of-concept simulation (Simulation 3), the task of discriminating the 100 unattested from the 100 ungrammatical sentences cannot be solved perfectly as CHILDES contains ungrammatical, incomplete and misspelled sentences, a few of which were randomly selected as unattested sentences (e.g., *higgledy piggledy my*). Similarly, a number of the automatically generated ungrammatical sentences were actually grammatical (see example given in Table 4).[17] However, the evaluation set is appropriate for a comparative evaluation.

Procedure. In order to train the model, left and right half-word representations were computed on the corpus as described in Table 2, i.e., using relative frequencies that correspond to the maximum likelihood estimate for each probability. The left half-word and right half-word distributions were then available for all 8601 words. For the category-based approach, following Redington

[16]The number of training sentences varies slightly from trial to trial. This is because an unattested sentence that occurs 20 times in the Manchester corpus will "remove" more instances from the training set than an unattested sentence that occurs only once.

[17]In order to make the experiment replicable, manual intervention in the production of the test set was avoided. Grammaticality is a gradient notion and different experimenters would create qualitatively different test sets based on the notion of grammaticality they apply. The manual correction of all of CHILDES is a significant effort that would be beyond the scope of this project. Hence, the ungrammatical sentences (automatically generated or CHILDES), and the misspellings were not corrected.

et al. (1998), the most frequent 1000 words were clustered (using single-link clustering, Manning & Schütze, 1999) by combining items which are closest together with respect to the similarity measure. Combined items, or "clusters", can then be further combined with nearby items or clusters. For each remaining word w, the closest neighbor w' in the 1000 most frequent words was determined and w was then assigned to the cluster of w'. Once trained, each model was then presented with the unattested and ungrammatical test sentences. As described above, the local syntactic coherence score of each test sentence was computed as the largest similarity between the test sentence and any training set sentence. Based on this largest similarity, a grammaticality judgment was made as detailed below.

Analysis of accuracy and window length. Fig. 8 shows the performance of category-based and exemplar-based discrimination for different subsequence sizes n. To compute the accuracy for each n, the ρ or λ with optimal discrimination performance was chosen. Results shown are aggregated for all 20 runs because all runs were consistent with the main results of the paper. For example, for $n = 4$, the smallest accuracy of the exemplar-theoretic model was 98% and the largest 100%; the smallest accuracy of the category-based model was 86.5% and the largest accuracy was 94%; the smallest difference in accuracy between the two models was 5% and the largest difference was 12.5%.

For a subsequence of size $n = 1$, the performance is 0.5 in both cases since the 200-sentence test set does not contain unknown words. So for every half-word, there is a sequence of one half-word in the training corpus with maximum similarity (which is 1.0 in the exemplar-theoretic model and $n - 1$ in the category-based model). Thus, all sentences get the same local syntactic coherence scores, both in exemplar-based and in category-based discrimination.

This is not the case for $n = 2$ since a sentence, as defined above, is considered to be locally coherent if all of its subsequences are coherent. While subsequences of 2 half-words that are part

of the *same* word have local syntactic coherence score 2.0 ($= 1.0 + 1.0$) in the exemplar-theoretic model, this is not true of subsequences of 2 half-words that are part of *different* words, e.g., the subsequence $<black_r, dog_l>$ in *black dog*. If *black dog* does not occur in the training set, then its local syntactic coherence score will in general be smaller than 2.0.

The experimental results demonstrate that, though imperfect for the reasons outlined above, exemplar-based discrimination of locally coherent vs. incoherent sentences is largely accurate for $n > 1$ and superior to the category-based discrimination investigated. Indeed, except for $n = 1$ the differences between category-based discrimination and exemplar-based discrimination were significant (Pearson's χ^2 test applying Yates' continuity correction, $p < 0.1$ for $2 \leq n \leq 10$ and $p < 0.05$ for $2 \leq n \leq 10, n \neq 9$).

Analysis of errors made by the models. In order to perform a qualitative analysis, the data corresponding to one of the window sizes ($n = 4$) was selected. This data set consists of 2000 ungrammatical and 2000 unattested sentences as summarized in Table 3. Table 3 shows the number of errors (both false positives and false negatives) and correct decisions (both true positives and true negatives) of the two models for the 4000 sentences. The category-based model makes about 10 times as many errors as the exemplar-based model, in keeping with the foregoing quantitative analysis.

A sentence-by-sentence analysis of the errors of the exemplar theoretic model was then performed. Table 4 gives examples of the types of errors found when analyzing the 37 ($27 + 10$) errors made by the model. Each line of the table gives a test sentence, which type of error was made, a false positive (FP) or a false negative (FN), the subsequence s_{test} of the test sentence whose greatest similarity to any subsequence in the training set was smallest, the training sentence subsequence s_{train} that s_{test} was most similar to, and a short description of the reason the error was made.

Sentence 1 (*higgledy piggledy my*) is an example of a sentence in CHILDES that is not grammatical in the ordinary sense of "grammatical sentence". It is the truncated title of a nursery rhyme. It is not surprising that this sentence is categorized as ungrammatical.

Sentence 2 (*little bird cannot speak she's got a worm in her beak*) exemplifies a problem that could also be alleviated by a better annotation of the CHILDES corpus. In this case, the sentence boundary (or, at a minimum, the intonational break) between *speak* and *she* is not annotated. This confuses the model and causes the incorrect false negative judgment.

Sentence 3 (*you've got a stuffed nose haven't you*) is an example of the difficulties that rare words pose. The word *stuffed* only occurred once in the training set of the relevant run (run 9 of the 20 runs) and its left and right half-word distributions were so dissimilar from the distributions of all other half-words that the most similar sequence of half-words in the training set was the only occurrence of *stuffed* in the training corpus. Due to the ubiquity of rare events in linguistic statistics, it is unavoidable that some errors due to unreliable rare event statistics are made.

Sentence 4 (*chocolate tummy broke*) is an example of a false positive that is caused by a violation of the assumption that the randomly generated sentences are all ungrammatical. While the sentence may be anomalous semantically, the three word sentence structure (consisting of a two-noun compound and an inflected verb) is perfectly grammatical.

The reason that sentence 5 was incorrectly judged to be similar to the sequence *one bottle there* is probably that more complex long-distance relationships are not modeled in the model. The hallmark of locally coherent strings is that they are perceived as well-formed independent of context. This is not true for noun+adverb combinations. Bigrams like *(noun) there* or *(noun) yesterday* require particular contexts to be well-formed. E.g., this type of bigram is grammatical in contexts like *I'm not in the kitchen right now, so the bottle there is not visible to me* and *Gonzales' performance yesterday will be remembered for many years to come*, but are of questionable

grammaticality in contexts like *the tv yesterday was green* or *his mother there is sick*. Since the model is a model of local coherence only, it does not represent sequences correctly that depend on a larger context. The premise is that more complex syntactic structures are learned later and the current goal is to model the acquisition of local syntactic coherence only.

Table 5 shows the number of decisions that the two models agree and disagree on as well as the type of agreement or disagreement, broken down according to which model was correct and which was wrong and according to type of sentence affected (grammatical or ungrammatical). It is apparent from the table that most errors committed by the exemplar-theoretic model are also committed by the category-based model: the two models share 19 false negative errors and 4 false positive errors. The exemplar-theoretic model committed only 8 false negative errors and 6 false positive errors that were not also committed by the category-based model. Several of the errors were the types of spurious errors in Table 4 that were discussed earlier. In particular, both sentences 1 and 4 were among those that in this evaluation were misclassified by the exemplar-based model, but "correctly" classified by the category-based model. In a few of the remaining cases, the category-based model categorized words into related classes whereas the exemplar-based model did not recognize their similarity. For example, the category-based model recognized that *diving* can be used prenominally and was therefore able to correctly classify *is that his diving board* as grammatical. In contrast, the exemplar-theoretic model analogized the word *diving* to the interjection *um* due to sparse data problems – *diving* occurred only once in a prenominal position in the relevant training set (run 6 of the 20 runs). Thus, there were a few cases where category-based inference was more robust. However, the experiment clearly shows that inference based on rich context, as available in the exemplar-theoretic model, is more successful overall.

This conclusion was confirmed by an analysis of a sample of the 384 sentences that the exemplar-theoretic model handled better than the category-based model (284 grammatical sen-

tences and 100 ungrammatical sentences). Note that the number of false positives in this set is much larger than the set of false negatives because the category-based model has a tendency of putting words in the same class even if they have different grammatical behavior. Examples of pairs of words that were analogized by the category-based model even though they are syntactically different are *museum/along*, *soapy/horse* and *ton/no*. Recall that the optimal number of clusters is chosen for the category-based model in each run – if such bad decisions of putting words in the same cluster are undone, then the model also separates words that should be in the same cluster and overall performance decreases. Apparently, it is difficult to represent all required syntactic context information with a fixed number of classes.

Summary. In the authors' view, the experiment demonstrates that the ability of exemplar theory to incorporate rich context information into discrimination decisions results in a more complete acquisition of local syntactic knowledge than a categorical model that has to operate on a representation from which most fine contextual detail has been removed.

The challenge for a categorical model is that it must provide an initial underspecificied representation that has not been modified by context yet – or one would assume the result of the successful disambiguation process that is the subject of investigation. The exemplar model, due to its ability to capture rich context in the representation of an individual word, does not have this problem.

This argument is of particular importance for ambiguous words. Underspecified representations for ambiguous words do not reflect intuitive syntactic categories like verb and noun. More importantly, they obfuscate the similarity of, for example, *increase* (which has a verb / noun ambiguity) to *bring* on the one hand (an unambiguous verb) and to *tree* (an unambiguous noun) on the other. In other words, category-based representations of ambiguous words are problematic because they are either too similar or not similar enough to the two alternatives. As a result, if a word with

a verb/noun ambiguity is represented as one of the alternatives, e.g. as a verb, then subsequences containing its noun use will no longer be similar to other subsequences with nouns. If a special conflation category noun/verb is introduced, then the same problem is encountered: subsequences containing the noun/verb category are not similar to subsequences containing either unambiguous verbs or unambiguous nouns.

Hence it seems hard to conceive of a categorical model that can reconcile the competing demands of (i) initial context-independent representation and (ii) final disambiguated representation. Of course, only one particular categorical model has been investigated here, and found lacking. However, Redington et al.'s (1998) model is the best known and respected computational model of the acquisition of syntactic categories. Consequently, the authors view the experiments reported here as potential evidence that the limitations of category-based models discovered here might in fact be general limitations of all category-based models, but this requires further investigation.

6. General Discussion

This article has presented an exemplar-theoretic model that makes correct predictions using both hand-crafted and real corpus data for two specific linguistic phenomena, and, in the case of local syntactic coherence judgments, illustrates the benefits of rich exemplar representations over their categorical counterparts. Interestingly, the phenomena examined hail from different linguistic domains, yet the same model is capable of accounting for their behavior. It is also noteworthy that the model achieves this without prototypes or any explicit abstraction mechanism. At least for the phenomena investigated here, a simple exemplar model without prototypes seems to be sufficient. Note, in particular, that Abbot-Smith and Tomasello (2006) express doubts that a pure exemplar-theoretic model can account for grammaticality judgments. Their hybrid categorization model, outlined in section 1, allows for exemplar learning and retention but also offers an abstraction mechanism where a more abstract schema is somehow encoded in the summed similarities.

However, while the comprehension of an exemplar might strengthen the activation of an exemplar cloud as a whole, this does not necessarily entail that the exemplar representations themselves have to change. Indeed, the model presented here illustrates accurate local grammaticality acquisition without the need for any modification of stored exemplars nor any form of more abstract representation.[18] Similarly, the syllable and segment exemplar representations in section 3 are not modified; their clouds, in the frequent cases, simply become more dense over time and hence produce greater activation. From the viewpoint of categorization, in both fields, novel stimuli are correctly categorized through direct comparison with exemplars in memory. Thus, for the disparate phenomena examined above, exemplar theory appears to provide an adequate account. It is important to note that this is not necessarily a rejection of the development of abstract structures overall but, with respect to local grammaticality judgments, an indication that at this phase of language acquisition the largely accurate grammaticality judgments, which are a function of the emergence of syntactic awareness, can be accounted for in a purely exemplar-theoretic account such as described above. While it could perhaps be argued that some form of abstraction is implicitly encoded in the summed similarities in this model, there is certainly no explicit abstraction component.

In fact, the syntactic representations in the multi-level model could be the basis for a more explicit abstraction component in further language development. There is a lack of explicit computational models that show how the transition is made from an initial state in which syntactic categories are not recognized to a state where they are used in comprehension. In later stages of development, the exemplar clouds formed in the multi-level model could be linked either to innate syntactic categories or to a more abstract non-innate layer of representation. This interpretation of the multi-level model is not incompatible with a hybrid model like the one presented by Abbot-

[18] One can view half-words as representations that change over time. For example, each time a left neighbor w_l is observed for word w, the activation of w_l in the left half word representation of w is strengthened. While activations in the model presented here change, there is no more fundamental change to the representations.

Smith and Tomasello (2006). The difference is that an explicit mathematical and computational formalization was provided here.

From the phonetics perspective the model's ability to model the syllable duration variability observed by Schweitzer and Möbius (2004) in high frequency syllables (albeit with some underlying assumptions) is not dissimilar to the Reducing Effect discussed by Bybee. Similarly, her Conserving Effect is essentially captured by the multi-level model's competition between units and constituents in that greater frequency strengthens the memory representations of sequences enabling their access as whole units. Bybee's third effect concerns autonomy and she offers the example of the grammaticalization of *be going to*, i.e. its diachronic evolution into an intentional phrase. The multi-level model offered here is also capable of modeling this language evolution phenomenon (see Schütze et al., 2007).

A further important aspect of the research presented in this article concerns the use of radius-based and nearest-neighbor approaches to exemplar cloud formation. The experiments presented above offer compelling evidence that neighborhoods in exemplar theory must be radius-based as opposed to relying on a nearest-neighbor formation method. θ, ρ and λ are all radius parameters. While other researchers have used nearest neighbor approaches in their exemplar-theoretic research, in the case of grammaticality, even ungrammatical sentences have nearest neighbors (albeit neighbors that are far away). It is therefore not clear how grammaticality judgments could be modeled with nearest-neighbor clouds. Previous arguments in favor of nearest-neighbor clouds were based on difficulties found in implementing fixed-radius models (Pierrehumbert, 2001) and not on any fundamental reasons.

One challenge for exemplar theory is to explain how exemplars of constituents interact with exemplars of compositions of constituents into larger units. Segments and words on the one hand, and syllables and phrases on the other hand, each give rise to exemplar clouds at different levels.

One of the key properties of language is the interaction of such units at different levels. This multi-level approach provides the first exemplar-theoretic model that explicitly incorporates constituency, both at the level of phonetics and syntax. Furthermore, this research represents a first step towards placing syntactic exemplar theory on a more formal footing with explicit statements of the assumptions of the model and the ability to test them against data. Up to now, some of the more noteworthy examples of exemplar-theoretic work on syntax have lacked a significant formal component (e.g., Abbot-Smith & Tomasello, 2006; Bybee, 2006). However, Bod (2006) has recently argued that data-oriented parsing (DOP) is an exemplar model. While no formal exemplar model can provide such a full exemplar-based account of grammatical productivity as DOP provides, it is perhaps noteworthy that DOP offers no notion of similarity between a stimulus and an exemplar cloud member. This notion of similarity tends to be a critical factor in most "traditional" (and perhaps therefore non-syntax oriented) exemplar models. Moreover, it is unclear to what extent DOP might be successfully applied to account for linguistic phenomena other than grammatical productivity, e.g. the syllable duration variability discussed above.

Challenges for the model

Before concluding it is important to note that although the model presented above has not yet been applied to phrase or sentence production, research currently underway is investigating how this can be best achieved. In addition, although the model implicitly acquires local grammatical knowledge it is important to note that it does not model the developmental timecourse of acquisition. It simply demonstrates the learnability of local syntax using representations rich in distributional information. One possible way to address this would be to provide the model, in increments, with data representative in quality and quantity to data received by children by particular developmental milestones, and to then test the model to see if the corresponding developmental targets have been acquired. This is clearly non-trivial and beyond the scope of this

article. A further area of future work to be considered concerns the fact that the radius parameters of the exemplar neighborhoods, and the activation thresholds below which the constituent level is chosen, were manually selected. Obviously, the performance of the model depends on the values of these parameters. If the radius in the grammaticality model is too large, then even ungrammatical sentences will be judged grammatical. However, it was not the goal of these experiments to automaticlly find θ or ρ, rather the aim was to determine if thresholding and a radius-based approach would work at all, and a variety of empirically selected values model the data well. However, the authors believe that these parameters can be estimated from the distribution of exemplars. For example, the similarities of ungrammatical sentences to the nearest neighbor are much smaller than those of grammatical sentences. One possibility for selecting θ is to divide the input in two, for training and testing, compute the distribution of distance values obtained when finding the nearest neighbor for each test sequence among the training sequences, and then select a value close to the maximum (e.g. the first percentile). The assumption would be that almost all of the input sequences are grammatical and that for a large enough test set, the range of distance values would be represented well by this sample. Taking a value close to the maximum will then ensure that all grammatical test sequences will have a training sequence with distance below θ. Density estimation is another possible solution to this problem. It might also be worth generalizing the multi-level model to account for gradience. For example, instead of making a discrete grammaticality decision based on the radius ρ, the degree of grammaticality could be computed by a function like the sigmoid that assigns discrete values approximating 1 or 0, for sentences that are a considerable distance above or below the threshold respectively, but intermediate values to sentences that are close to the threshold. This would approximate thresholding behavior while also providing continuous gradient grammaticality judgments in keeping with those found in the literature (see Theakston, 2004; Ambridge, Pine, Rowland, & Young, 2008).

To conclude, the exemplar model presented above is successful in accounting for phenomena from both phonetics and syntax. It does so by employing a novel architecture which facilitates the explicit modeling and invocation of constituency interaction. The model offers insights into the nature of exemplar cloud formation and illustrates the benefits of rich exemplar representations when compared with abstract category-based representations. These results are in keeping with Hay and Bresnan's (2006) prediction that combining insights from across a variety of linguistic domains might yield more universal and cognitively plausible models of language acquisition and use.

Acknowledgements. The authors are grateful to the anonymous reviewers for their constructive and extensive comments on an earlier draft of this paper.

References

Abbot-Smith, K., & Tomasello, M. (2006). Exemplar-learning and schematization in a usage-based account of syntactic acquisition. *The Linguistic Review*, *23*, 275–290.

Akhtar, N. (1999). Acquiring basic word order: Evidence for data-driven learning of syntactic structure. *Journal of Child Language*, *26*, 339–356.

Akhtar, N., & Tomasello, M. (1997). Young children's productivity with word order and verb morphology. *Developmental Psychology*, *33*(6), 952–965.

Ambridge, B., Pine, J., Rowland, C., & Young, C. (2008). The effect of verb semantic class and verb frequency (entrenchment) on children's and adult's graded judgements of argument-structure overgeneralization errors. *Cognition*, *106*, 87–129.

Bencini, G., & Valian, V. (2008). Abstract sentence representations in 3-year-olds: Evidence from language production and comprehension. *Journal of Memory and Language*, *59*, 97–113.

Benner, U., Flechsig, I., Dogil, G., & Möbius, B. (2007). Coarticulatory resistance in a mental syllabary. In *Proceedings of the international congress of phonetic sciences (saarbrücken)* (pp. 485–488).

Bock, J. (1986). Syntactic persistence in language production. *Cognitive Psychology*, *18*, 355–387.

Bock, J., & Griffin, Z. (2000). The persistence of structural priming: Transient activiation or implicit learning? *Journal of Experimental Psychology: General*, *129*, 177–192.

Bock, J., & Loebell, H. (1990). Framing sentences. *Cognition*, *35*, 1–39.

Bod, R. (2000). The storage and computation of frequent sentences. In *Architectures and mechanisms for language processing conference (amlap-2000).* Leiden, The Netherlands.

Bod, R. (2006). Exemplar-based syntax: How to get productivity from examples. *The Linguistic Review*, *23*, 291–320.

Bybee, J. (2006). From usage to grammar: the minds response to repetition. *Language*, *84*, 529–551.

Bybee, J., & Scheibman, J. (1999). The effect of usage on the degrees of constituency: the reduction of don't in english. *Linguistics*, *37*, 575–596.

Chang, F. (2002). Symbolically speaking: A connectionist model of sentence production. *Cognitive Science*, *26*, 609–651.

Chang, F., Dell, G., & Bock, K. (2006). Becoming syntactic. *Psychological Review*, *113*(2), 234–272.

Chang, F., Dell, G., Bock, K., & Griffin, Z. (2000). Structural priming as implicit learning: A comparison of models of sentence production. *Journal of Psycholinguistic Research*, *29*(2), 217–229.

Childers, J., & Tomasello, M. (2001). The role of pronouns in young children's acquisition of the english transitive construction. *Developmental Psychology*, *37*(6), 739–748.

Conwell, E., & Demuth, K. (2007). Early syntactic productivity: evidence from dative shift. *Cognition*, *103*, 163–179.

Croot, K., & Rastle, K. (2004). Is there a syllabary containing stored articulatory plans for speech production in english? In *Proceedings of the 10th Australian International Conference on Speech Science and Technology* (pp. 376–381). Sydney, Australia.

Dellwo, V., & Wagner, P. (2003). Relationships between speech rate and rhythm. In *Proceedings of the 15th International Congress of Phonetic Sciences (barcelona)* (pp. 471–474).

Elman, J. (1990). Finding structure in time. *Cognitive Science*, *14*, 179–211.

Eriksson, A. (1991). *Aspects of Swedish speech rhythm.* Unpublished doctoral dissertation, University of Göteborg, Sweden.

Fernandes, K., Marcus, G., Nubila, J. D., & Vouloumanos, A. (2006). From semantics to syntax and back again: Argument structure in the third year of life. *Cognition*, *100*, B10–B20.

Fisher, C. (2002). The role of abstract syntactic knowledge in language acquisition: a reply to Tomasello. *Cognition*, *82*, 259–278.

Fowler, C. (1977). *Timing control in speech production*. Bloomington, IN: Indiana University Linguistics Club.

Gertner, Y., Fisher, C., & Eisengart, J. (2006). Learning words and rules: Abstract knowledge of word order in early sentence comprehension. *Psychological Science*, *17*(8), 684–691.

Grossberg, S. (2003). Resonant neural dynamics of speech perception. *Journal of Phonetics*, *31*, 423–445.

Hay, J., & Bresnan, J. (2006). Spoken syntax: The phonetics of "giving a hand" in new zealand english. *The Linguistic Review*, *23*.

Hintzman, D. L. (1986). 'Schema Abstraction' in a Multiple-Trace Memory Model. *Psychological Review*, *93*, 328–338.

Huttenlocher, J., Vasilyeva, M., Cymerman, E., & Levine, S. (2002). Language input and child syntax. *Cognitive Psychology*, *45*, 337–374.

Johnson, K. (1997). Speech perception without speaker normalization: An exemplar model. In K. Johnson & J. W. Mullennix (Eds.), *Talker variability in speech processing* (pp. 145–165). San Diego: Academic Press.

Jurafsky, D. (2003). Probabilistic modeling in psycholinguistics: Linguistic comprehension and production. In R. Bod, J. Hay, & S. Jannedy (Eds.), *Probabilistic linguistics* (pp. 39–95). Mass.: MIT Press.

Kohler, K. (1983). Stress-timing and speech rate in German—a production model. *Arbeitsberichte des Instituts für Phonetik und digitale Sprachverarbeitung (Univ. Kiel), AIPUK*, *20*, 5–53.

Levelt, W., Roelofs, A., & Meyer, A. (1999). A theory of lexical access in speech production. *Behavioral and Brain Sciences*, *22*, 1–75.

Levelt, W., & Wheeldon, L. (1994). Do speakers have access to a mental syllabary. *Cognition*, *50*, 239–269.

MacWhinney, B. (Ed.). (2000). *The CHILDES project: Tools for analyzing talk.* Mahwah, NJ: Lawrence Erlbaum Associates. (3rd edition)

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. Boston, MA: MIT Press.

Matthews, D., Lieven, E., Theakston, A., & Tomasello, M. (2005). The role of frequency in the acquisition of english word order. *Cognitive Development*, *20*, 121–136.

Mayer, J., Ackermann, H., Dogil, G., Erb, M., & Grodd, W. (2003). Syllable retrieval vs. online assembly: fMRI examination of the syllabary. In *Proceedings of the 15th International Congress of Phonetic Sciences* (pp. 2541–2544). Barcelona, Spain.

Mintz, T. (2002). Category induction from distributional cues in an artificial language. *Memory & Cognition*, *30*, 678–686.

Mintz, T. (2003). Frequent frames as a cue for grammatical categories in child directed speech. *Cognition*, *90*, 91–117.

Monaghan, P., Christiansen, M., & Chater, N. (2007). The phonological-distributional coherence hypothesis: Cross-linguistic evidence in language acquisition. *Cognitive Psychology*, *55*, 259–305.

Morris, W., Cottrell, G., & Elman, J. (2000). A connectionist simulation of the empirical acquisition of grammatical relations. In S. Wermter & R. Sun (Eds.), *Hybrid neural symbolic integration* (pp. 175–193). Springer Verlag.

Müller, K., Möbius, B., & Prescher, D. (2000). Inducing probabilistic syllable classes using multivariate clustering. In *Proceedings of the 38th meeting of the Association of Computational Linguistics* (pp. 225–232). Hong Kong.

Naigles, L. (1990). Children use syntax to learn verb meanings. *Journal of Child Language*, *17*(2), 357–374.

Nelson, K. E. (1977). Facilitating children's syntax acquisition. *Developmental Psychology*, *13*, 101–107.

Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39–57.

Olguin, R., & Tomasello, M. (1993). Twenty-five-month-old children do not have a grammatical category of verb. *Cognitive Development*, *8*, 245–272.

Pickering, M., & Branigan, H. (1998). The representation of verbs: Evidence from syntactic priming in language production. *Journal of Memory and Language*, *39*, 633–651.

Pierrehumbert, J. (2001). Exemplar dynamics: Word frequency, lenition and contrast. In J. Bybee & P. Hopper (Eds.), *Frequency effects and the emergence of lexical structure* (pp. 137–157). Amsterdam: John Benjamins.

Pinker, S., Lebeaux, D., & Frost, L. (1987). Productivity and constraints in the acquisition of the passive. *Cognition*, *26*, 195–267.

Redington, M., Chater, N., & Finch, S. (1998). Distributional information: A powerful cue for acquiring syntactic categories. *Cognitive Science*, *22*(4), 425–469.

Saffran, E., & Martin, N. (1997). Effects of structural priming on sentence production in aphasics. *Language and Cognitive Processes*, *12*, 877–882.

Savage, C., Lieven, E., Theakston, A., & Tomasello, M. (2003). Testing the abstractness of children's linguistic representations: lexical and structural priming of syntactic constructions in young children. *Developmental Science*, *6*(5), 557–567.

Schütze, H. (1992). Dimensions of meaning. In *Proc. of Supercomputing '92* (pp. 787–796). Los Alamitos, CA: IEEE Computer Society Press.

Schütze, H. (1993). Distributed syntactic representations with an application to part-of-speech tagging. In *Proc. of the IEEE International Conference on Neural Networks* (pp. 1504–1509).

Schütze, H. (1995). Distributional part-of-speech tagging. In *EACL 7* (pp. 141–148).

Schütze, H., & Walsh, M. (2008, October). A graph-theoretic model of lexical syntactic acquisition. In *Proceedings of the 2008 conference on empirical methods in natural language processing* (pp. 917–926). Honolulu, Hawaii: Association for Computational Linguistics.

Schütze, H., Walsh, M., Wade, T., & Möbius, B. (2007). Towards a unified exemplar-theoretic model of phonetic and syntactic phenomena. In *29th Annual Meeting of the Cognitive Science Society* (pp. 1461–1466). Nashville.

Schweitzer, A., & Möbius, B. (2004). Exemplar-based production of prosody: Evidence from segment and syllable durations. In *Proceedings of the Speech Prosody 2004 Conference* (pp. 459–462). Nara, Japan.

Theakston, A. (2004). The role of entrenchment in children's and adult's performance on grammaticality judgment tasks. *Cognitive Development*, *19*, 15–34.

Tomasello, M. (2000). Do young children have adult syntactic competence? *Cognition*, *74*, 209–253.

Tomasello, M. (2006). Acquiring linguistic constructions. In R. Siegler & D. Kuhn (Eds.), *Handbook of child psychology: Cognitive development* (pp. 255–298). Wiley.

Tomasello, M., & Abbot-Smith, K. (2002). A tale of two theories: Response to Fisher. *Cognition*, *83*, 207–214.

Tomasello, M., & Brooks, P. (1998). Young children's earliest transitive and intransitive constructions. *Cognitive Linguistics*, *9*, 379–395.

Underwood, G., Schmitt, N., & Galpin, A. (2004). The eyes have it: An eye-movement study into the processing of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences* (pp. 153–172). Amsterdam and Philadelphia: John Benjamins.

Varley, R., & Whiteside, S. (2001). What is the underlying impairment in acquired apraxia of speech? *Aphasiology*, *15*, 39–49.

Wagner, P. (2008). *The rhythm of language and speech: Constraining factors, models, metrics and applications.* Habilitationsschrift, University of Bonn, Germany.

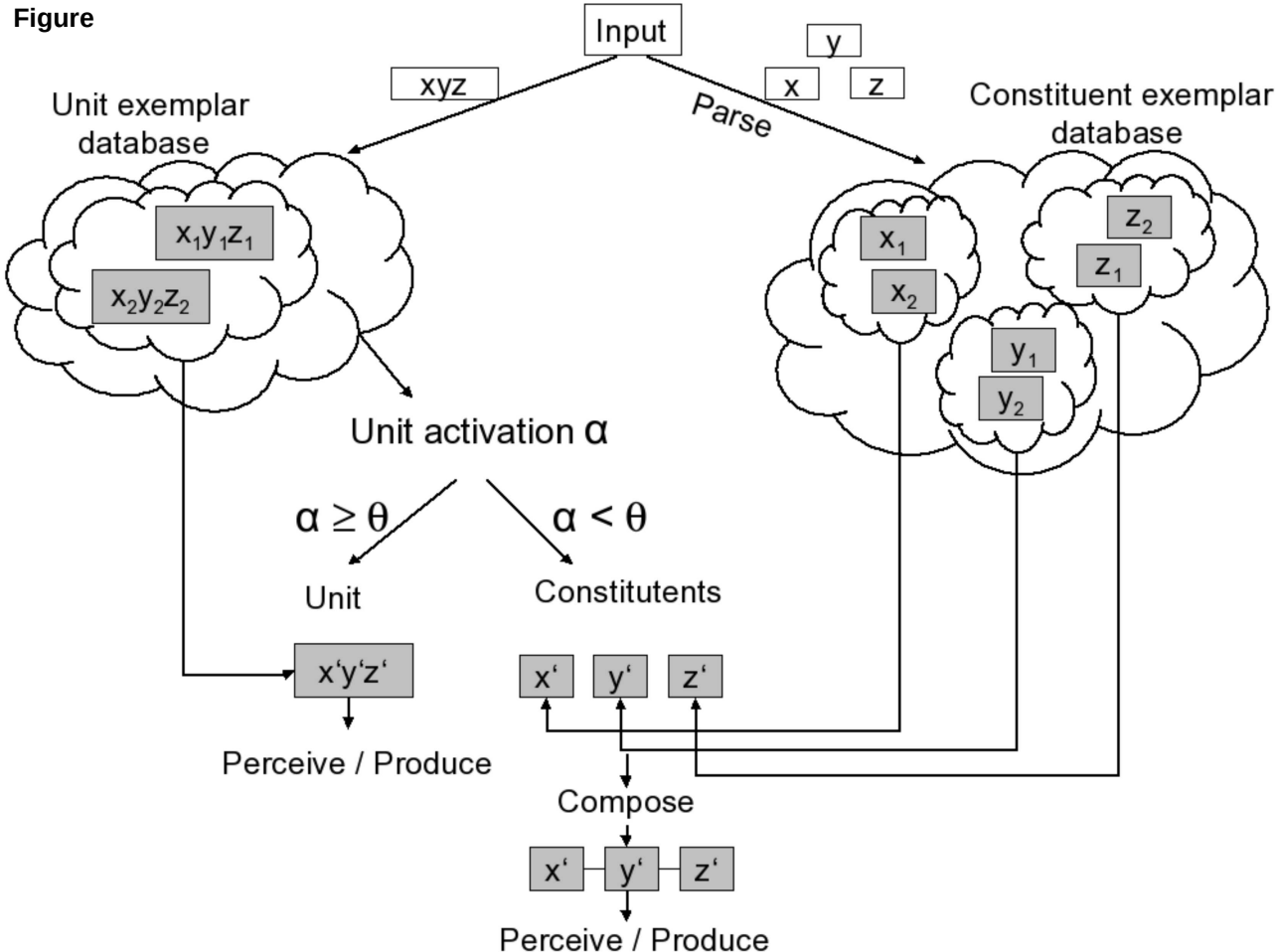
Walsh, M., Schütze, H., Möbius, B., & Schweitzer, A. (2007). An exemplar-theoretic account of syllable frequency effects. In *Proceedings of the 16th International Congress of Phonetics Sciences (ICPhS 2007)* (pp. 481–483). Saarbrücken, Germany.

Walsh, M., Schütze, H., Wade, T., & Möbius, B. (2007). Accounting for phonetic and syntactic phenomena in a multi-level competitive interaction model. In *ESSLLI Workshop on Exemplar Based Models of Language Acquisition and Use.* Dublin.

Whiteside, S., & Varley, R. (1998). Dual-route phonetic encoding: Some acoustic evidence. In *Proceedings of the 5th International Conference on Spoken Language Processing* (pp. 3155–3158). Sydney, Australia.

List of Figures

Figure

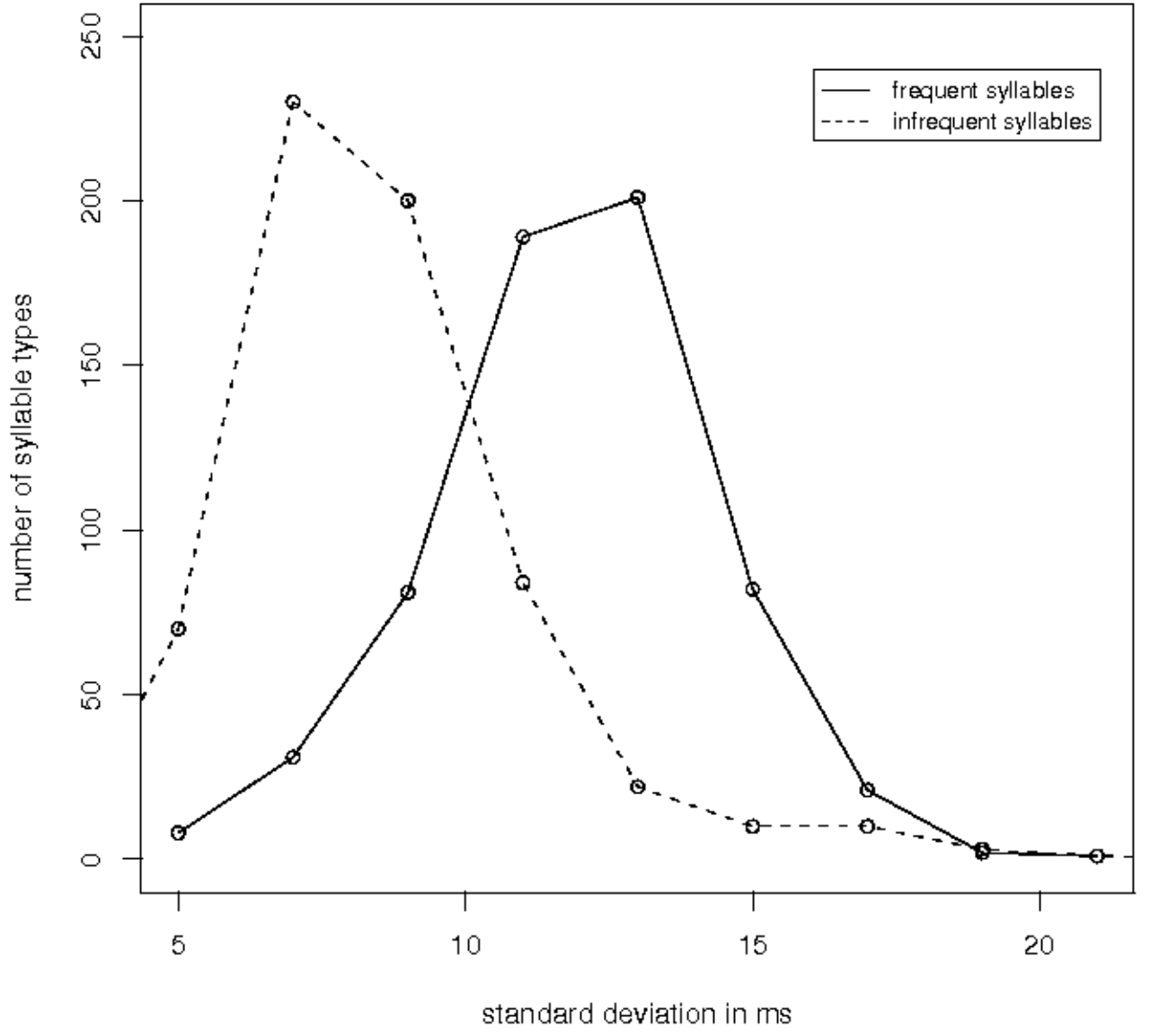

Input
xyz
y
x
z
Parse
Unit exemplar database
Constituent exemplar database
$x_1y_1z_1$
$x_2y_2z_2$
x_1
x_2
z_2
z_1
y_1
y_2
Unit activation α
α ≥ θ
α < θ
Unit
Constitutents
x‘y‘z‘
x‘
y‘
z‘
Perceive / Produce
Compose
x‘
y‘
z‘
Perceive / Produce

Table

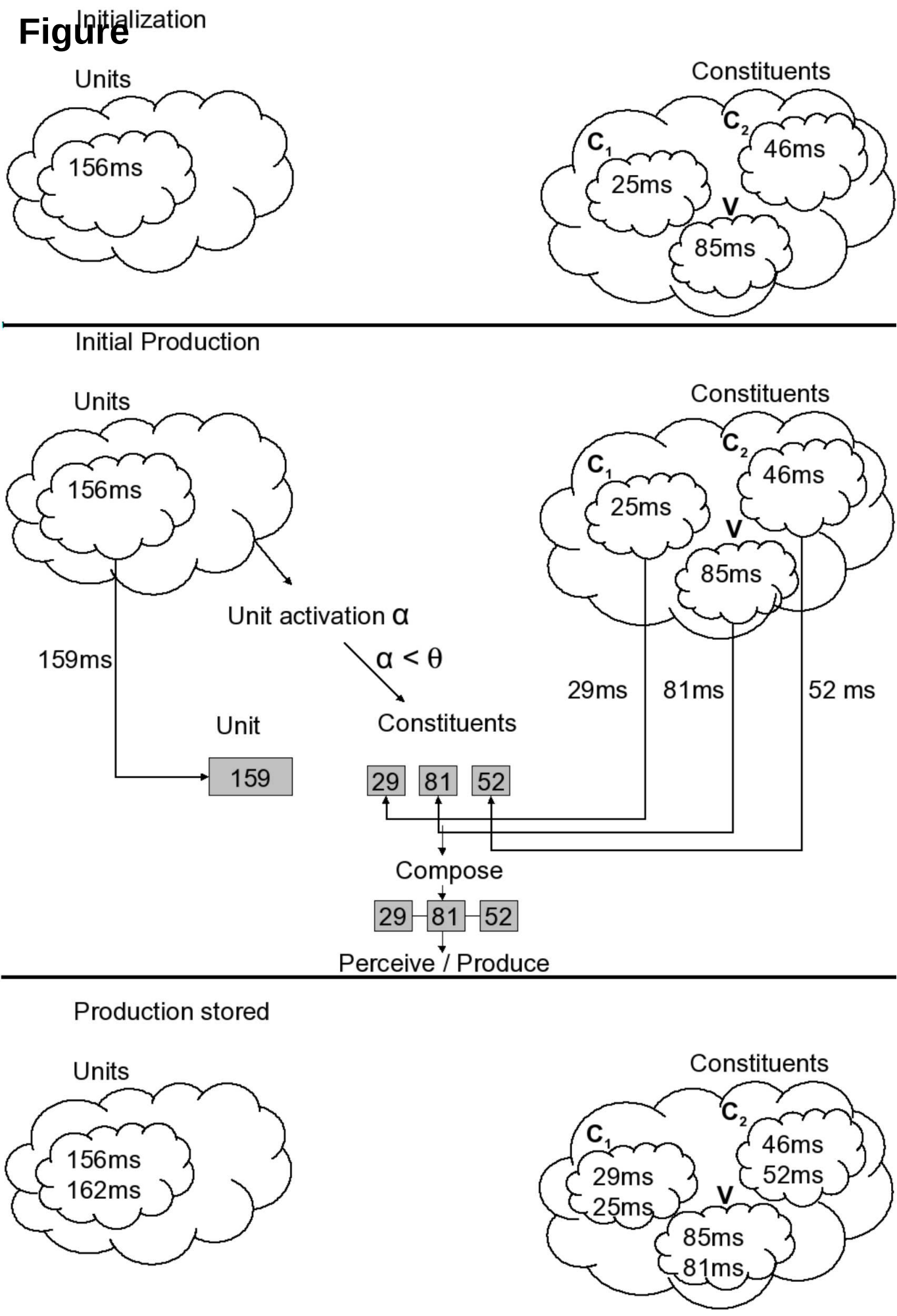
	syllable duration	grammaticality
stimuli	syllable to be produced	phrase (in perception)
constituents	segments	words
constituent representation	acoustics, duration	word's left / right context
similarity of constituents	sum of similarities of the components of the representation	
units	syllables	phrases
unit representation	sequence of constituents	
similarity of units	sum of similarities of the constituents of the units	
property inferred	duration of syllable	grammaticality of novel phrase

Table 1: Components of the unified exemplar-theoretic model.

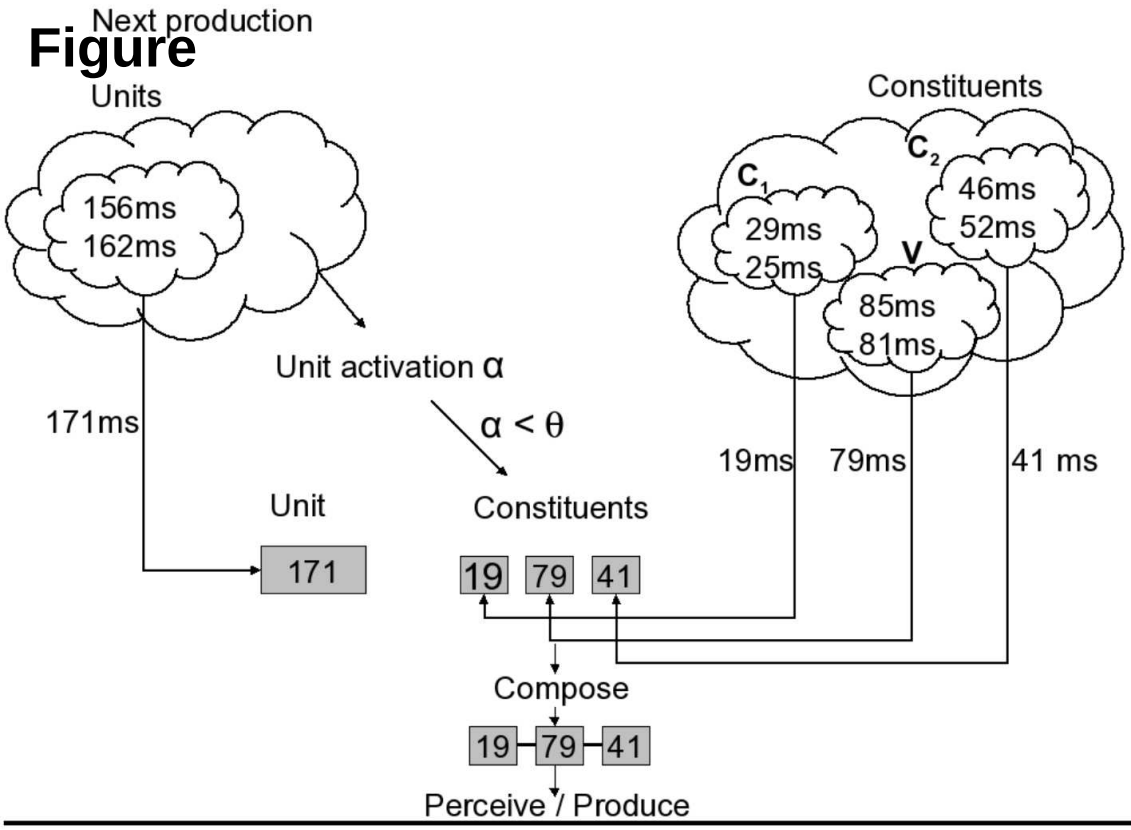
Figure

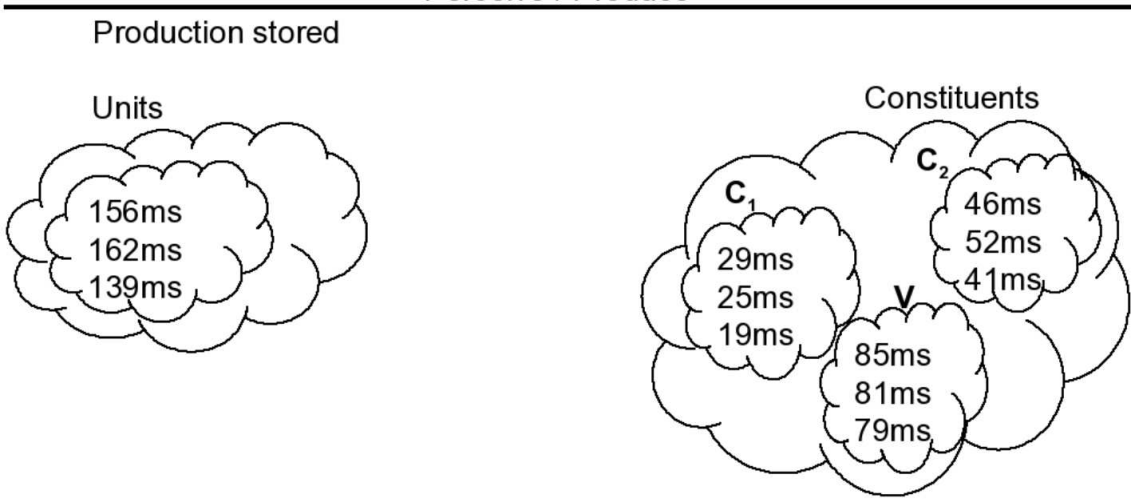

frequent syllables
infrequent syllables
number of syllable types
0
50
100
150
200
250
5
10
15
20
standard deviation in ms

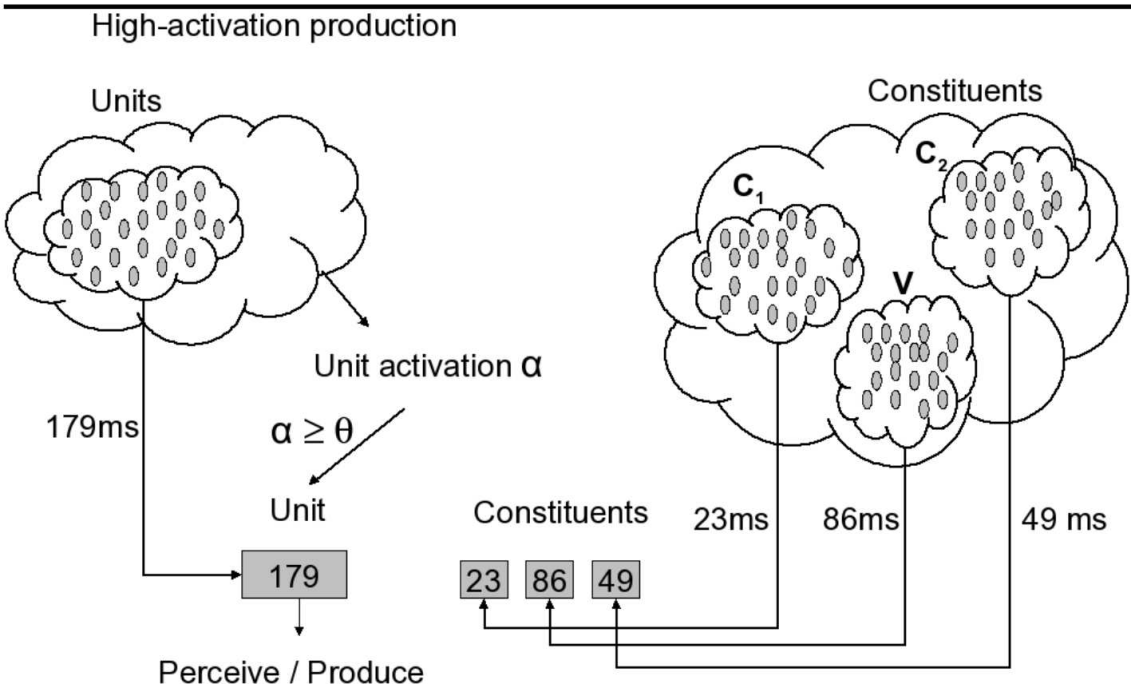
Figure


Initialization
Units
156ms
Constituents
C1
25ms
C2
46ms
V
85ms
Initial Production
Units
156ms
Constituents
C1
25ms
C2
46ms
V
85ms
Unit activation α
α < θ
159ms
29ms
81ms
52 ms
Unit
159
Constituents
29
81
52
Compose
29
81
52
Perceive / Produce
Production stored
Units
156ms
162ms
Constituents
C1
29ms
25ms
C2
46ms
52ms
V
85ms
81ms

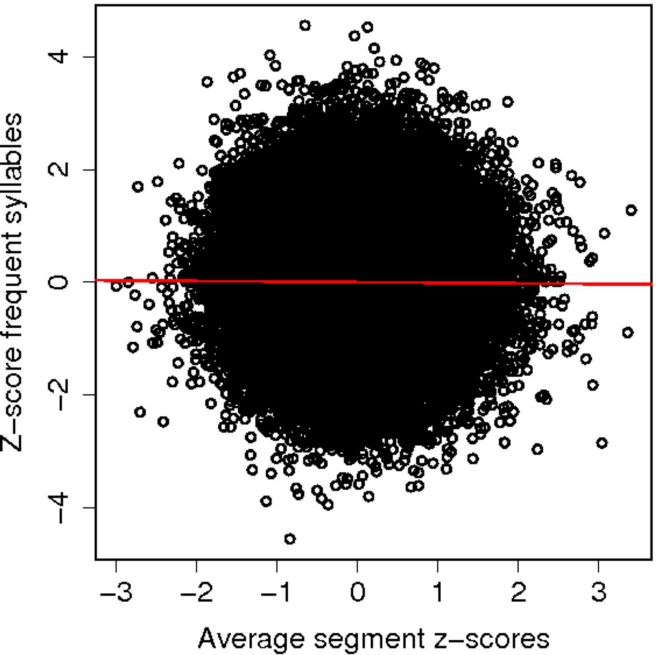
Figure

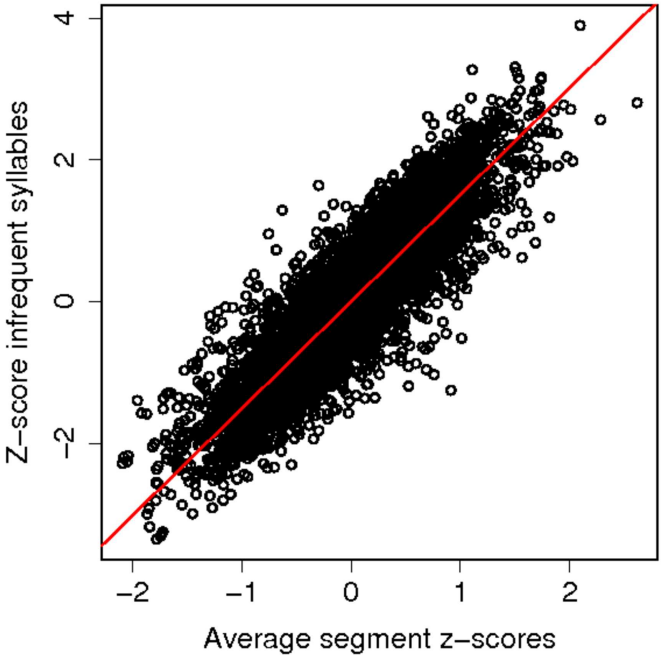

Next production
Units
156ms
162ms
Unit activation α
171ms
α < θ
Unit
171
Constituents
19
79
41
Compose
19 – 79 – 41
Perceive / Produce
Constituents
C1
C2
V
29ms
25ms
46ms
52ms
85ms
81ms
19ms
79ms
41 ms


Production stored
Units
156ms
162ms
139ms
Constituents
C1
C2
V
29ms
25ms
19ms
46ms
52ms
41ms
85ms
81ms
79ms

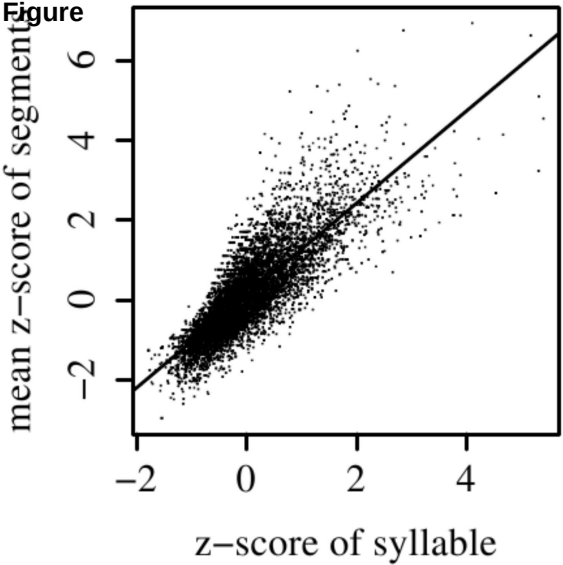

High-activation production
Units
Unit activation α
179ms
α ≥ θ
Unit
179
Perceive / Produce
Constituents
C1
C2
V
Constituents
23
86
49
23ms
86ms
49 ms

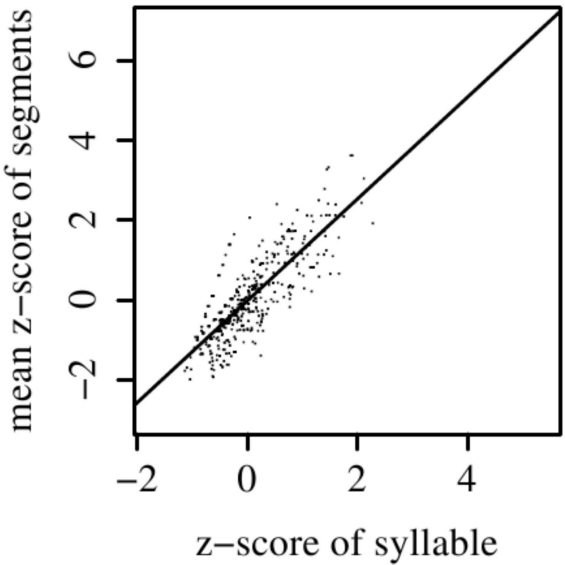
Figure


Z−score frequent syllables
4
2
0
−2
−4
−3
−2
−1
0
1
2
3
Average segment z−scores


Z−score infrequent syllables
4
2
0
−2
−2
−1
0
1
2
Average segment z−scores

Figure


mean z−score of segments
6
4
2
0
−2
−2
0
2
4
z−score of syllable

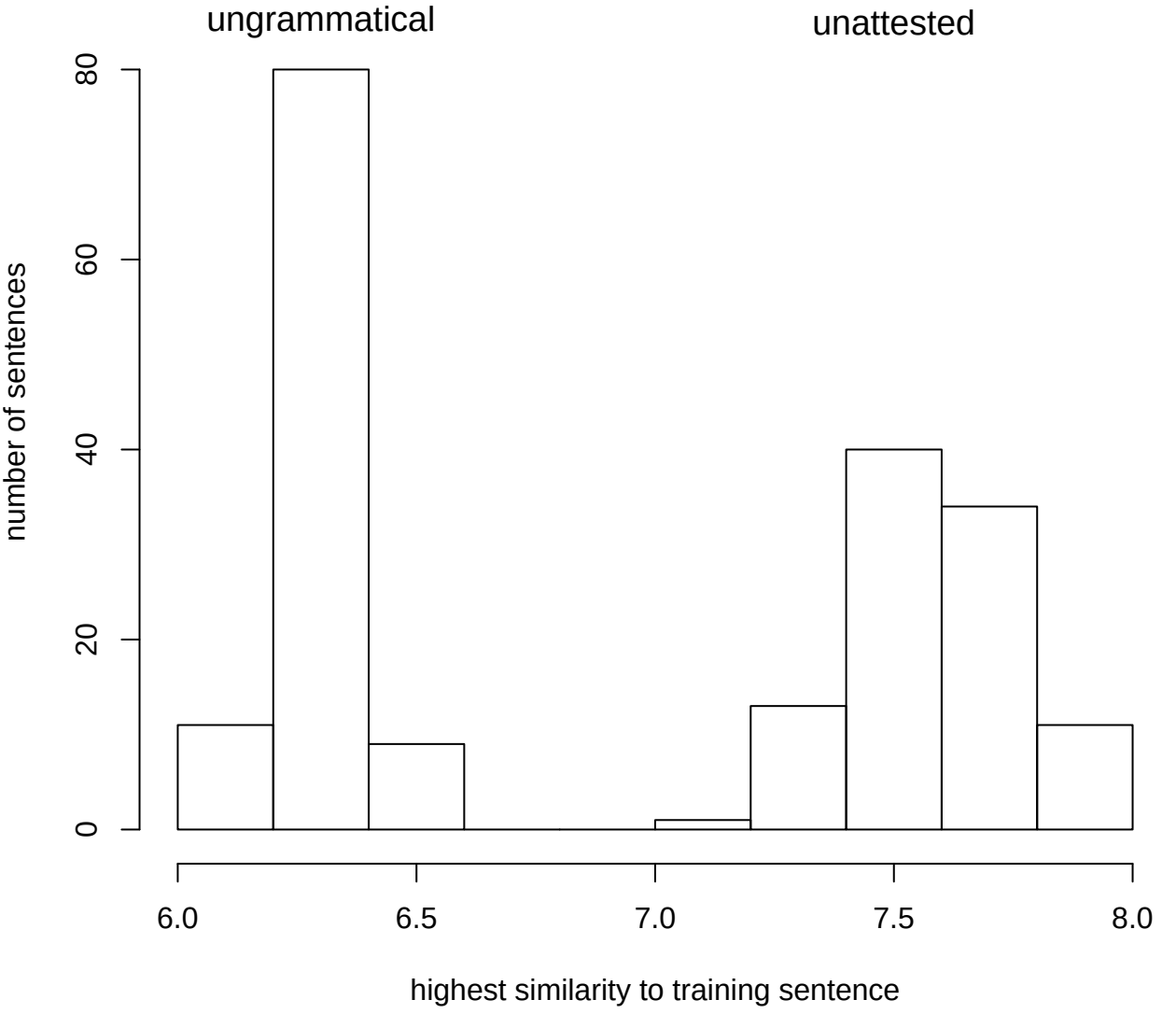

mean z−score of segments
6
4
2
0
−2
−2
0
2
4
z−score of syllable

Table

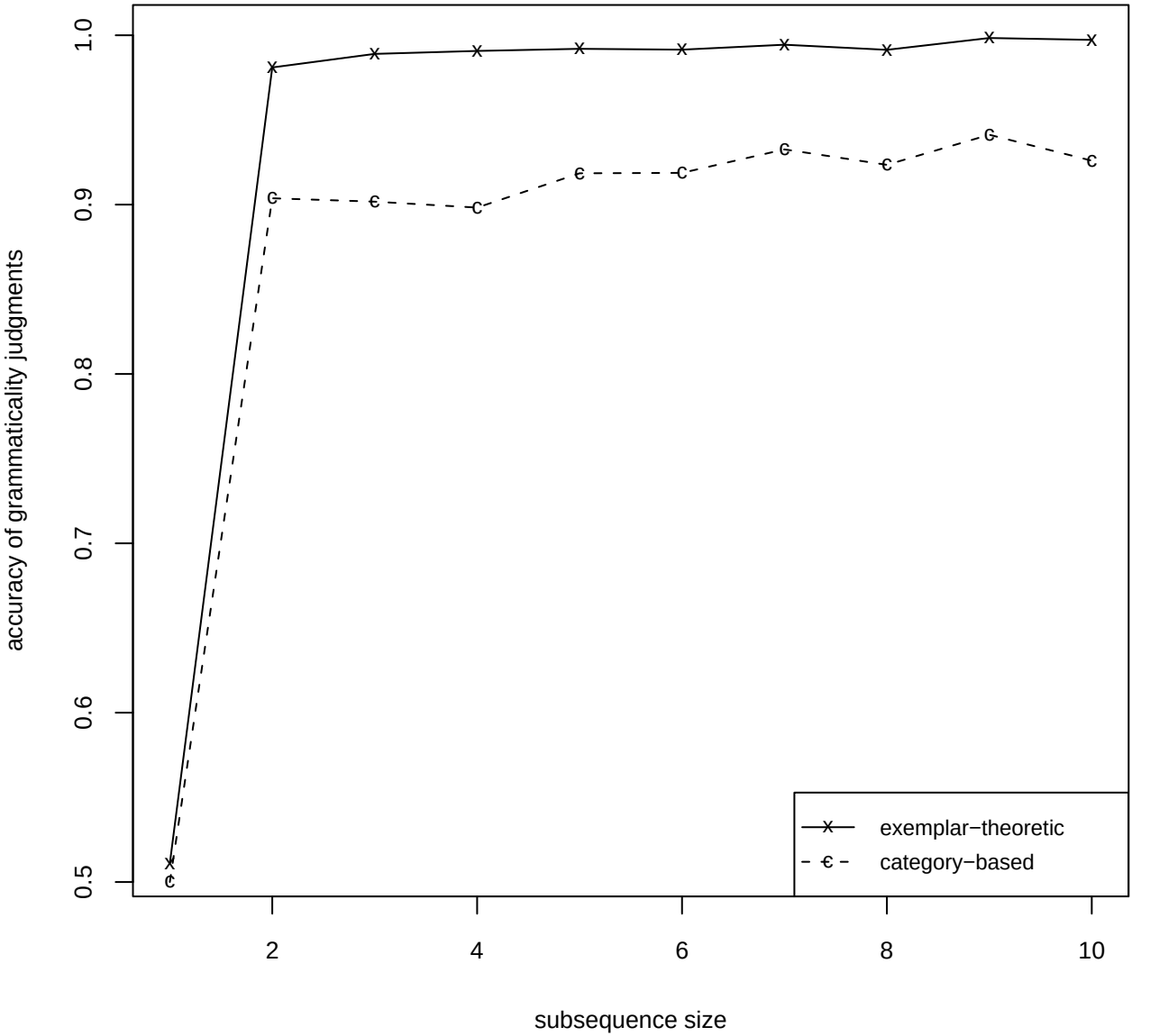
	people	chocolate	beside
$like_r$	0.0054	0.0004	0.0000006
$love_r$	0.000098	0.0072	0.000007
sit_r	0.0000001	0.0000001	0.0045

Table 2: Subsection of a bigram statistics matrix. Each row represents part of the probability distribution for the corresponding right half-word given in column 1.

Figure


ungrammatical
unattested
number of sentences
0
20
40
60
80
6.0
6.5
7.0
7.5
8.0
highest similarity to training sentence

Figure


accuracy of grammaticality judgments
0.5
0.6
0.7
0.8
0.9
1.0
2
4
6
8
10
subsequence size
exemplar−theoretic
category−based

Table

exemplar-theoretic model			**category-based model**		
	grammatical	ungrammatical		grammatical	ungrammatical
accepted	1973 (TP)	10 (FP)	accepted	1697 (TP)	104 (FP)
rejected	27 (FN)	1990 (TN)	rejected	303 (FN)	1896 (TN)

Table 3: Number of errors (false positives (FP) and negatives (FN)) and correct decisions (true positives (TP) and negatives (TN)) of the two models for the sentences used in the qualitative analysis.

Table

	test sentence	type	closest subsequences		type of problem
			test	train	
1	higgledy piggledy my	FN	piggledy (l/r) my (l/r)	beep (l/r) my (l/r)	unattested sentence ungrammatical
2	little bird cannot speak she's got a worm in her beak	FN	cannot (r) speak (l/r) she (l)	you (r) speak (l/r) NAME (l)	missing sentence boundary
3	you've got a stuffed nose haven't you	FN	a (r) stuffed (l/r) nose (l)	probably (r) stuffed (l/r) in (l)	rare word stats unreliable
4	chocolate tummy broke	FP	chocolate (l/r) tummy (l/r)	chocolate (l/r) pudding (l/r)	random sentence grammatical
5	somewhere bottle rubble	FP	somewhere (r) bottle (l/r) rubble (l)	one (r) bottle (l/r) there (l)	long-distance dependency

Table 4: Qualitative analysis of false positive and false negative errors of the exemplar-theoretic model. Each line gives a test sentence, whether it was a false positive (FP) or a false negative (FN), the subsequence s_{test} of the test sentence whose greatest similarity to any subsequence in the training set was smallest, the training sentence subsequence s_{train} that s_{test} was most similar to, and a short description of the reason the error was made. Words in subsequences are marked with l (left half-word), r (right half-word) or l/r (both left and right half-words), depending on which of the word's half-words were part of the subsequence.

Table

		category-based			
		correct		incorrect	
		GR	UN	GR	UN
exemplar-theoretic	correct	1689	1890	284	100
exemplar-theoretic	incorrect	8	6	19	4

Table 5: Number and type of decisions that the two models agree and disagree on. GR = grammatical, UN = ungrammatical. For example, there were 6 ungrammatical sentences for which the category-based model made a correct decision (TN) and the exemplar-theoretic model made an incorrect decision (FP).