Department of Computer Science,
University of Otago

UNIVERSITY of OTAGO

SAPERE AUDE

Te Whare Wānanga o Otāgo

Technical Report OUCS-2010-02

Generation of idioms in a simple recurrent network architecture

Authors:

Martin Takac, Alistair Knott, Lubica Benuskova

Department of Computer Science, University of Otago, New Zealand

Status:

Currently being rewritten to be submitted to Language and Speech journal

---

Department of Computer Science,
University of Otago, PO Box 56, Dunedin, Otago, New Zealand

http://www.cs.otago.ac.nz/research/techreports.php

Generation of idioms in a simple recurrent network architecture

Martin Takac, Alistair Knott, and Lubica Benuskova
Dept. of Computer Science, University of Otago, Dunedin, New Zealand

Abstract

Idioms are an ideal testbed for studying the interplay of lexical (content preparing) and syntactic (structure building) mechanisms in language production. This article contributes to the debate about the nature of these mechanisms and their relationship from the viewpoint of computational modeling. We present a neural network model of sentence generation, which is able to produce continuous and discontinuous idioms within regular compositional sentences. The model is a simple recurrent network extended to include a semantic episode representation as an extra input. Our main contribution consists in a detailed analysis of the representational space of the network's hidden layer, which shows that (1) an implicit structure-content division can arise as a result of internal space reorganization within a single SRN during learning, (2) idioms can be produced by the same general sequencing mechanism that works for regular sentences, (3) the production of idioms is modulated by content-specific mechanisms.

Introduction

While language allows words to be productively combined to convey new meanings, everyday language also contains many **fixed expressions**: sequences or patterns of words which occur together with particularly high frequency. Fixed expressions occupy a continuum, from songs and proverbs at the 'fixed' end of the spectrum, through idioms (e.g. *kick the bucket*, *take X to task*) and phrasal verbs (e.g. *come across*, *pull up*) to statistically-defined collocations at the other end (e.g. *good as gold*, *as soon as possible*); for a taxonomy, see Cowie (1998). What they have in common is that they are in some sense 'regularly occurring' structures of words: phrases whose component words occur together more frequently than one would expect by chance.

Fixed expressions can have various origins. Some fixed expressions have their origin in the fact that certain *situations* which need linguistic expression occur with particularly high frequency. For instance, the origin of the frequently-occurring expression *How are you?* is likely to be the fact that people frequently ask after one another's health. Other

This research was supported by a BuildIT postdoctoral fellowship grant for Martin Takac.

fixed expressions are patterns of words whose meanings have become conventionalised over time, so that they contribute their meaning collectively rather than through their individual component words. Of course, these two origins are not exclusive; high-frequency messages are good candidates for expression through conventionalised fixed expressions. (E.g. *Howya doin?* conventionally realises the message 'How are you?'.)

In the present paper, we will be focusing on fixed expressions which convey conventionalised meanings, which we will term **idioms**. We will define an idiom as a fixed expression which displays restricted syntax and whose meaning cannot be derived compositionally from its constituents. Idioms vary in degree of syntactic and lexical frozenness; some of them permit several transformations, e.g. *she kicked the bucket* or *she will kick the bucket*, but not *the bucket was kicked by her*, others are completely fixed. They also vary in the way their meaning is related to meanings of their components. In some cases, the original meaning of an idiom is quite easily recoverable; for instance if it has a metaphorical meaning which is still apparent (e.g. '*meet your maker*'). In other cases, the meaning is not easy to recover (e.g. above-mentioned *kick the bucket*).[1]

Estimates of the proportion of fixed expressions in the vocabulary of language users vary from a conservative 7% (Sprenger, 2003, results for Dutch) to almost half of the lexicon (Jackendoff, 1995). In either case, these estimates suggest that fixed expressions are nothing exceptional, but rather form a significant part of language. In terms of frequency of usage, fixed expressions which convey their meanings compositionally (e.g. *looking forward to* or *as soon as possible*) in fact outnumber idioms (Sprenger, 2003). However, in the current article, our interest is in fixed expressions which convey a particular semantic concept collectively, rather than compositionally.

Studying idioms is important for several reasons. First, idioms are surface patterns and have to be acquired and represented as wholes. The need to account for idioms places important constraints on the architecture of a syntactic theory. For instance, Jackendoff (2002) argues against generative theories in the Chomskyan tradition on the grounds that they cannot represent idioms. On the other hand, constructivist models of syntax (e.g. Goldberg, 1995; Tomasello, 2003) are expressly designed to deal with idioms as well as with productive grammatical rules. For constructivists, fixed expressions are qualitatively similar to regular grammatical rules: they are both learned associations between surface linguistic patterns and semantic patterns, which differ only in the complexity of the associated patterns. Constructivist accounts of language postulate a single unified mechanism for learning both types of structure. This learning mechanism is often modelled using connectionist architectures, which are especially suitable for extracting patterns from data (e.g. Elman et al., 1996; Christiansen & Chater, 2001). A first hypothesis we want to test is the constructivist hypothesis that idioms do not require a special treatment or postulation of a dedicated mechanism, but can be treated within a general framework of acquisition and production.

Second, idioms are an ideal testbed for studying the interplay of lexical and syntactic mechanisms in language production. Some leave these mechanisms undistinguished or intertwined; for instance Elman (1991) argues for a distributed representation of words that inherently also carries contextual/syntactic information. Others have postulated special

[1] Both of these phrases mean *to die*.

representational units carrying syntactic information specific for particular lexical concepts: **lemmas** for words (Levelt, Roelofs, & Meyer, 1999) and **superlemmas** for idioms (Levelt & Meyer, 2000; Sprenger, Levelt, & Kempen, 2006). Others (Chang, 2002; Chang, Dell, & Bock, 2006; Konopka & Bock, 2009) accept the need for lexically specific syntactical information, but argue for more general abstract syntactic mechanisms that map event structure to syntactic frames without being necessarily triggered by/dependent on lexical retrieval. For instance, in the connectionist models of Chang, there is a separate subnetwork that controls sequencing (of thematic roles in structural frames) and another one that supplies representational content. Chang argues that distributed architectures which lack this distinction would not be capable of structural generalizations that people apparently master without any problems. Again, a connectionist model of idiom production can help decide between these alternatives.

While we ultimately agree with Chang that there must be a structural division between sequencing and representational components of the network, we nonetheless want to explore to what extent this division can implicitly develop in the representational space of a single connectionist network as a result of learning from a language sample that contains a mixture of idiomatic and regular transitive sentences. For this purpose, we have devised a simple recurrent network (SRN) model of language production of both idioms and non-idioms and tested it on an artificial language. In Study 1, we explore the ability of our model to acquire both idioms and non-idioms as surface patterns and learn to produce them. In Study 2, we enhance the model with semantics and explore the model's ability to learn to produce syntactically well-formed and semantically correct sentences for given meanings. Then we compare both architectures and based on analysis of their state space draw conclusions about representation of different parts of language in their state space. To explore differences of language representation in the state space of SRN with and without semantics we make lesions to different parts of SRN and observe which part of language processing and production has been impaired.

The rest of the article is organized as follows. First we summarise the relevant aspects of existing theories of language and idiom production. Then, after presenting the results of Study 1 and Study 2, we give an analysis of the internal organization that developed within the neural network. To further analyse the interplay of syntactic and lexical mechanisms in the network, we report results of experiments with lesioning different parts of the model. We end with general discussion and directions for future research.

Theories of language production

In this section we briefly recapitulate the approach of V. S. Ferreira & Slevc (2009), who at first attempt to present a consensus model of what theories of language production have in common, then pinpoint aspects of the model that are still the matter of ongoing debates.

Language production is a process of transforming non-linguistic meanings (concepts) into a set of linguistic forms that can be phonologically encoded and ultimately spoken or written out. At a first approximation, it consists of two subprocesses: one which prepares the *content* of the utterance, and one which determines the *structure* of the target utterance. Each process has in turn two stages: a stage of *selection* of linguistic features from a

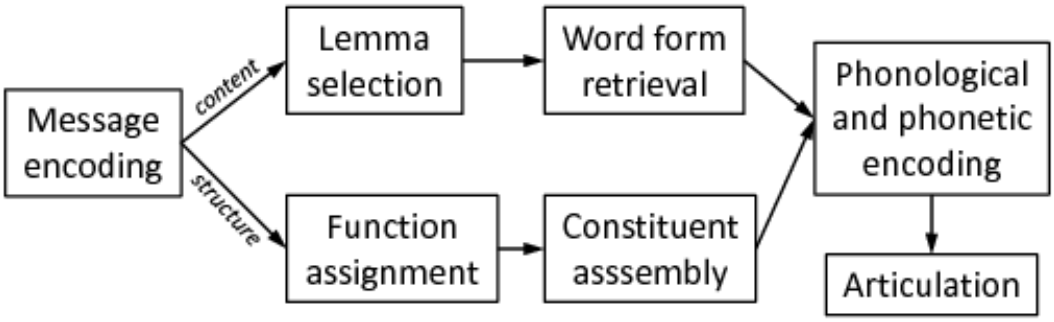

Figure 1. The consensus model of language production. Adapted from V. S. Ferreira & Slevc (2009).

candidate set, and a later stage of *retrieval* of the properties of the selected features, with possible overlap or feedback interactions between the stages (Figure 1).[2]

Preverbal meanings to be expressed (messages) enter the process of grammatical encoding as descriptions of events with thematic roles ('who did what to whom'). They have a *semantic aspect* (semantic features of the involved entities and actions or states), a *relational aspect* (mutual relations among the semantic constituents) and a *perspective aspect* (specifying the relative importance of components of the event description).

The first step of the content preparation process is *lexical selection* – choosing a set of lexical entries that appropriately express the semantic meaning. These entries, called *lemmas* (Levelt et al., 1999) are not yet fully specified words, but more like pointers to lexically-specific syntactic information (e.g. a form class, like noun or verb) of the word to be expressed.[3] The word itself in full detail of its morphophonological features is retrieved in the second step – *word form retrieval*.

Structure-building subprocess deals with syntactic information necessary for expressing relational and perspective aspects of the message. The first step, *functional assignment* (Bock, 1995), selects the appropriate grammatical functions, e.g. subject, direct object, modifiers, etc. The second step, *constituent assembly*, retrieves the constituent structures for the selected grammatical functions and orders them in an appropriate affixed sequence.[4]

If structure and content are processed separately, they must be reunited at some stage appropriately. How exactly this binding problem is solved is still an open problem (V. S. Ferreira & Slevc, 2009). In fact, there is considerable disagreement as to whether structure and content are indeed processed separately. Advocates of integrated processing (e.g. Levelt et al., 1999; Sprenger et al., 2006) can more easily explain lexically-specific syntactic phenomena (e.g. different structural privileges of different content words) and also do not have to deal with the above-mentioned binding problem. On the other hand,

[2]Levelt et al. (1999) maintain that these stages are discrete, i.e. lemmas have to be fully selected before the next stage can commence, others (e.g. Dell, 1986; Cutting & Ferreira, 1999) advocate more interactions, possibly allowing the latter stage to influence the former.

[3]Lexical selection only concerns with content words; function words (e.g. determiners) are produced by the structure-building subprocess.

[4]Affixes and function words are more important for languages with free word order.

proponents of content-independent abstract structures and mechanisms (e.g. Chang et al., 2006; Konopka & Bock, 2009) can more easily account for lexically-independent syntactic effects, especially structural priming (Bock, 1986).[5]

Production of idioms

Two major accounts of idiom production are Cutting & Bock (1997) and Sprenger (2003), Sprenger et al. (2006). Based on experiments with elicited speech errors (idiom blends), Cutting and Bock propose a hybrid model, where idioms have their own lexical-conceptual entry (among entries for non-idiomatic concepts) directly linked to the simple lemmas that constitute the idiom. Once the idiom entry activates its lemmas, further processing proceeds in a standard way and there is no difference between idioms and non-idioms thereafter. The syntax of an idiom is specified by the idiom entry activating an abstract phrasal frame independent of any lemmas. The frame has slots that can be filled by lemmas activated by the idiomatic concept. However, this mechanism leads to a binding problem for some idioms, e.g. *be a wolf in sheep's clothing* (Sprenger, 2003): the phrasal VP frame contains two slots that can be filled with nouns and the information specifying which slot should be filled with *wolf* and which with *sheep* is neither stored in the abstract frame nor with the nouns' lemmas.

This problem is amended in the *Superlemma theory* (Levelt & Meyer, 2000; Sprenger, 2003; Sprenger et al., 2006) that is the extension of Levelt's language production model (Levelt et al., 1999). In this theory, idioms have special holistic entries called *superlemmas* stored and competing with other (non-idiomatic) lemmas. A superlemma does not point to an abstract syntactic frame, but instead represents a syntactic structure of the idiom itself, together with pointers to particular lemmas. Thus selection of a superlemma predetermines which simple lemmas will fill each slot in its syntactic frame. Also, the syntactic properties stored with the superlemma modify or restrict those stored with simple lemmas.

Despite their differences, both theories stress the hybrid nature of idiom representation: holistic lexically-specific representation on a certain level,[6] but standard syntactic processing thereafter. Also, they both refuse to countenance a special mechanism of idiom production, instead incorporating idiom production within a standard model of 'regular' language production.

Language production and the brain

There are many accounts of the brain areas which are involved in the production of sentences. In this section we will briefly introduce the accounts we will refer to when considering how components of our network map onto to brain regions.

The research on what brain areas are involved in language production has been informed by two main sources: lesion data and, more recently, brain-imaging studies. Here we briefly review Indefrey (2009), who summarizes the results of a comprehensive meta-analysis

[5]Structural priming, also called syntactic persistence, is the tendency of speakers to reuse the same syntactic structures they used before, regardless of the content.

[6]Konopka & Bock (2009) provide experimental support for the claim that even if idiomatic expressions are stored in unitary form, their internal structure is accessible to and undergoes standard syntactic processing. Insignificant difference in structural generalizations from idiomatic and non-idiomatic phrasal verbs lead the authors to emphasize abstract structural processes in language production.

Table 1: Neural correlates of selected language production subprocesses according to Indefrey (2009).

Process	Brain locations
lemma selection	mid left medial temporal gyrus (MTG), left posterior superior temporal gyrus (STG)
word form retrieval	left posterior superior temporal lobe
sentence structure building	posterior part of Broca's area (left BA 44)

of many brain imaging studies. His conclusions about involvement of particular brain areas in various phases of language production are mostly based on the presence or absence of hypothesized processes in certain experimental tasks, for which brain-imaging data are available. For single-word production, he adopts the theoretical model of Levelt et al. (1999) consisting of stages of conceptual preparation, lexical (lemma) selection, morpho-phonological code (word form) retrieval, phonetic encoding, articulation and self-monitoring. On the sentence-level, he distinguishes between syntactic encoding of a sentence (that we named structure building above) and broader conceptual planning of a discourse. For the purposes of this article, we will only focus on his findings about neural correlates of lemma selection, word form retrieval and sentence structure building, which are summarized in Table 1 and Figure 2a. For comparison, Broca's and Wernicke's areas traditionally associated with language are shown in Figure 2b.

Study 1: Production of syntactically well-formed sentences containing idioms

Methods

In our first study, we focused on modeling production of grammatically correct sentences containing idioms, without any semantic input. We created a neural network model, trained it for next-word prediction on a sample of an artificial language containing various types of idioms, and then tested its (spontaneous) sentence generation ability. The grammatical correctness of the produced sentences was judged against the artificial language (i.e. a generated sentence was correct, if it belonged to the language).

This task was repeated 5 times with different random initializations of connection weights in the network and also with different training sets, as if having 5 persons exposed to different samples of the same language during their lifetime. Hence we will refer to these 5 instances of the model trained on different sets as the *model subjects* from now on. All results were averaged over the 5 model subjects.

Neural model. We used a Simple Recurrent Network (SRN, Elman, 1990) with localist coding of words on the input and output layers (see Figure 3). Thanks to recurrent connections to the hidden layer, the SRN is provided with a sort of a memory and its output does not depend on a current input only, but also on the previous context (via the context layer). SRNs have proven to be especially suitable for tasks involving learning of temporal dependencies and sequences (Elman, 1990) and language syntax in particular (Elman, 1991; Christiansen & Chater, 1999; Tino, Cernansky, & Benuskova, 2004). The

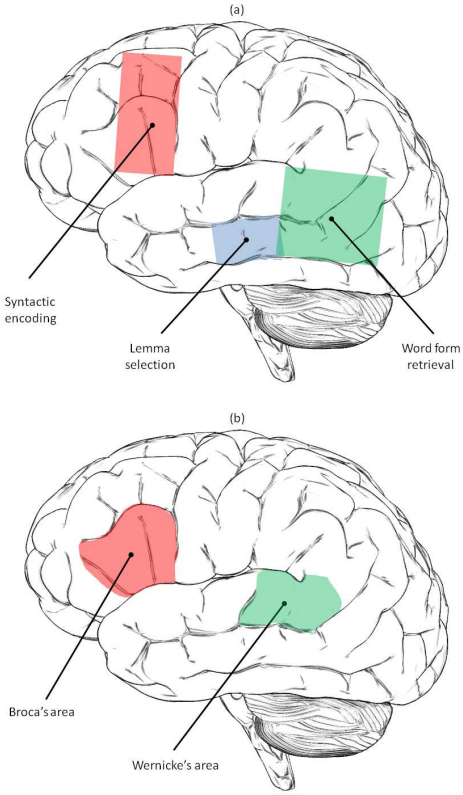

Figure 2. (a) Neural correlates of selected language production subprocesses according to Indefrey (2009). (b) Broca's and Wernicke's areas traditionally associated with language.

SRN can be understood as modelling circuits in posterior Broca's area—specifically, the posterior part of left Brodmann's Area (BA) 44—and adjacent premotor cortex. Several studies have suggested that posterior part of BA44 specialises in recognising patterns in sequential stimuli, including syntactic patterns in language (see e.g. Dominey, Hoen, & Inui, 2006). This same area, together with an adjacent area of premotor cortex, also appears to be involved in the generation of syntactic structures (see e.g. Indefrey et al., 2001). There are other studies which propose posterior Broca's area and adjacent articulatory premotor cortex as the locus of phonological representations—specifically of the phonological output buffer (see e.g. Paulesu, Frith, & Frackowiak, 1993; Henson, Burgess, & Frith, 2000). Since the SRN contains layers holding purely phonological representations (the current word and next word layers), it makes sense for these areas to hold a mixture of phonological and syntactic representations. However, we must also include left posterior STG as one of the areas holding phonological information, even during sentence generation; see e.g. Hickok et al. (2000).

Our SRN had 58 input units for 57 possible words plus 1 unit coding a sentence boundary (SB). In each time step, the unit corresponding to the current word had activity 1 and all other units had 0 (so-called **1-hot coding**). The network had 30 hidden units with sigmoidal activation functions, activities of which were copied back to the context layer in

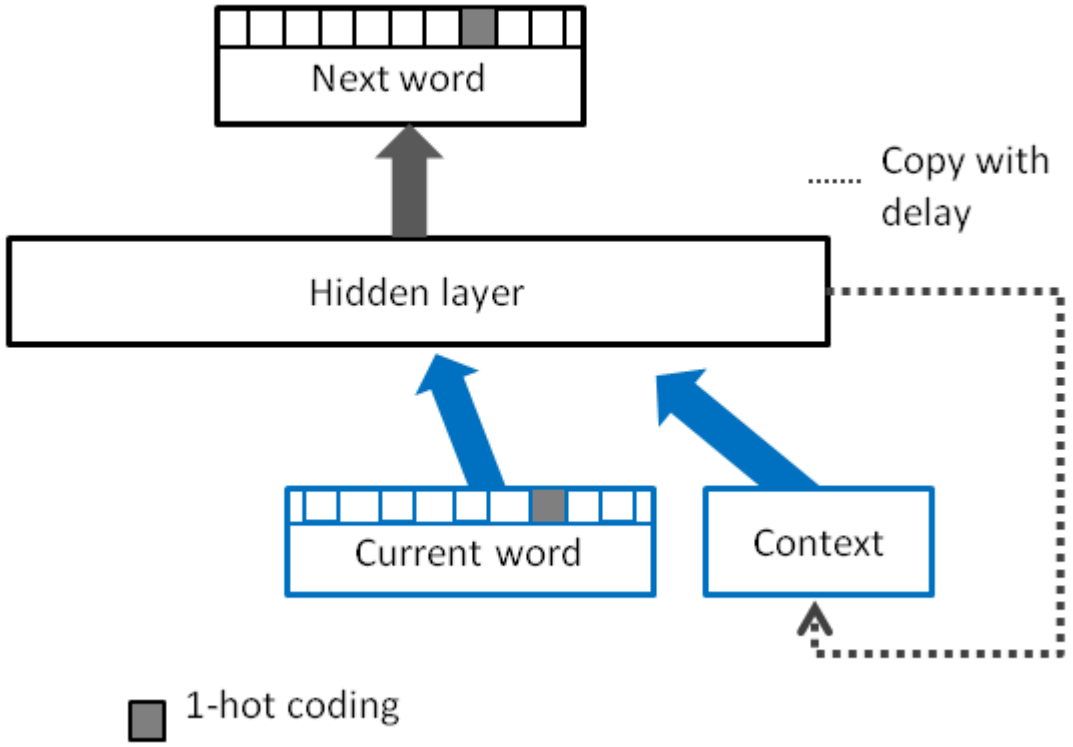

Figure 3. The SRN architecture for the task without semantics. The thick arrows mean fully connected layers. In language production mode, the output (predicted next word) is fed back to the input in the next time step.

the next time step. Sentences (separated by SB) were presented word by word on the input layer and the task of the network in each time step was to yield probability distribution of possible next words on the output layer. The output layer consisted of 58 linear units (one for each word) with softmax combination function (for technical details of computation, see e.g. Cernansky, 2007).

After training for next-word prediction, the network could run in a *sentence generation mode*. In this mode, we begin by setting SB as the current word and a conventional 'initial context' signal as the current context. Then at each iteration we feed activity forward to the output layer, stochastically choose an output word, and use the selected word as the current word for the next iteration, and the context layer activity as the current context layer input for the next iteration.[7] To prevent generation of unreasonably long sentences, we force SB on the input if it has not been selected as the next word within 20 words from the last SB.

For stochastic selection of the next word, we made use of the fact that the output units of a SRN trained for the next-word prediction task can be read as holding an estimate of the probability distribution of likely next words, provided they use localist coding (Elman,

[7]Note that we do not reset the context layer after each generated sentence; neither during sentence generation mode nor during training.

1990). We used a modified softmax selection policy called Boltzmann selection (Chambers, 1995):

$$P_i = \frac{\exp(a_i/T)}{\sum_{j=1}^{n} \exp(a_j/T)} ,$$

where P_i is a probability of selection of the i-th word, a_i is the activity of the i-th linear output neuron and T is a temperature. The probability of the selection of a neuron is proportional to its activity; the temperature T controls the stochasticity of the selection, from a deterministic choice of the most active neuron ($T = 0$) to a completely random selection ($T \to \infty$). We used $T = \frac{1}{3}$ in our experiments.

Artificial language. In both our studies, we investigated the performance of a SRN trained on an artificial language containing a mixture of idioms and syntactically regular constructions. Sentences in the language are syntactically homogeneous, in that they are all transitive; they differ only in their degree of idiomaticity. Our interest is in modelling how a child can learn to generate a mixture of productive and idiomatic sentences. In this first study, we only model syntax; in study 2, we will extend our model to consider semantics as well.

The core of our 57-word vocabulary consisted of words commonly used by 16-30 month-old toddlers according to the Child Development Inventory (CDI, Fenson et al., 1994). The grammar of our language allowed for regular transitive sentences and also for two types of idiom:

- continuous NP idioms (*teddy bear*, *Winnie the Pooh*, *ice cream*, *french fries*),
- discontinuous VP idioms (*kisses NP good bye*, *gives NP a hug*, *gives NP five*).

Note that these idioms do not all have the same degree of idiomaticity. For instance *gives NP a hug* is not fully idiomatic; it contains an NP 'slot' whose filler can have arbitrary (accusative) NP structure. Rather they exemplify the spectrum of possible idioms. However, there is some evidence that even phrases not considered idiomatic in adult language could be learned by children first as surface patterns or item-based constructions (Tomasello, 2003). Similarly, Pine & Lieven (1997) claim that although children use determiners with different noun types, there is no evidence for them possessing an adult-like syntactic category of determiners, which rather evolves gradually by broadening the range of lexically specific frames in which different determiners appear. Therefore, we omitted determiners (*a* and *the*) from our language, except for cases where they were part of an idiom, as in *give NP a hug* or *Winnie the Pooh*.

The language also featured semantic dependencies, in that some verbs could only be followed by animate objects and the verb *eats* could only be followed by a noun phrase denoting food. It also contained synonyms and lexical ambiguities (in that words for people could appear in both subject and object positions, the word *gives* could be a part of either *gives NP a hug* or *gives NP five* and the word *kisses* could be either a regular verb as in *grandpa kisses grandma* or a part of an idiom with a different meaning as in *grandpa kisses grandma good bye*). The complete language consisted of 2200 transitive sentences generated from a context-free grammar (see Table 2).

To allow for all mentioned phenomena, we made some extensions to the core CDI-based vocabulary. Out of the idioms used in our language, CDI explicitly contains *teddy*

Table 2: Top: Transcription rules for the language used in our simulations. All non-terminals are written in all capital letters (TRANSITIVE is the initial non-terminal), all terminals contain small letters. Period (.) is a terminal and stands for the sentence boundary (SB). Sentence type alternatives (right-hand side of the first rule) were assigned different probabilities of selection (not shown here). Bottom: Examples of sentences composed of single words, continuous idioms, and discontinuous idioms (note that a discontinuous VP idiom can be interleaved with a continuous NP one).

TRANSITIVE	→	SUBJ VERB_GEN OBJ_GEN . | SUBJ VERB_ANIM OBJ_ANIM . | SUBJ eats FOOD . | SUBJ kisses OBJ_ANIM good bye . | SUBJ gives OBJ_ANIM five . | SUBJ gives OBJ_ANIM a hug .
SUBJ	→	HUMAN
VERB_GEN	→	sees | loves | holds | bites | washes
VERB_ANIM	→	kisses | tickles | hugs
OBJ_GEN	→	HUMAN | THING | ANIMAL | FOOD
OBJ_ANIM	→	HUMAN | ANIMAL
HUMAN	→	mummy | daddy | Samko | Helen | Mia | grandma | grandpa
THING	→	ball | book | balloon | toy | doll | block | crayon | pen
ANIMAL	→	dog | kitty | duck | bunny | rabbit | cow | pig | bug | puppy | bee | monkey | teddy bear | Winnie the Pooh
FOOD	→	cookie | banana | apple | cheese | cracker | ice cream | bread | pizza | french fries

Mummy eats carrot. Mia loves ice cream. Helen tickles Winnie the Pooh. Grandpa gives grandma a hug. Daddy kisses teddy bear good bye.

bear, *ice cream*, *french fries* and *give me five*. It also contains single words *give*, *hug*, *kiss*, *good*, *bye* that we used in discontinuous idioms. We also added the word *rabbit* and the idiom *Winnie the Pooh*, which will feature as synonyms of *bunny* and *teddy bear* in Study 2, when we introduce semantics to the model.

Training. As a training set for the next-word prediction task, we used a symbolic sequence consisting of 500 randomly selected sentences presented word by word and separated by SB. The target vector at each time step was the next symbol in the sequence. We used a simple back-propagation training algorithm (Rumelhart, Hinton, & Williams, 1986) with entropy cost function, learning rate 0.1 and no momentum (for technical details see e.g. Cernansky, 2007). After each training epoch (i.e., one sweep through the training set), we evaluated the predictive performance of the network on a test set of 1,000 previously unseen sentences from the language using a normalized negative log-likelihood (NNL) function (see Appendix). The NNL is a measure of average surprise in prediction: 0 means no surprise i.e. 100% correct prediction, 1 corresponds to random guessing. We also separately evaluated the SRN's ability to spontaneously generate syntactically correct sentences (1,000 sentences were generated after each training epoch).

Results

Results of training for next-word prediction can be seen in Figure 4a. As the task is non-deterministic (multiple grammatically correct continuations of a sentence fragment are possible), a certain degree of error is inevitable (Elman, 1990). That is why the NNL for the test set does not tend to converge to zero. Even if the network successfully learns transition probabilities from the training set, the best it can do without semantics is to predict *categories*, not particular *words* (Cernansky, Makula, & Benuskova, 2007). Figure 5 shows this behaviour during sentence generation.

Without semantic input, the sentence generation is a product of the network's 'autonomous' regime, when a 1-hot coded word selected from the output is fed back to the input at the next time step. The variability in produced sentences is due to stochastic selection of the winning output word – if the selection was deterministic (i.e. winner-takes-all), the network would soon end up repeating the same subsequence. Figure 4b charts the syntactic performance of the network in sentence generation mode. We can see that the network was able to achieve almost perfect performance within just a few training epochs, not just for syntactically regular sentences, but for sentences containing idioms too. We took a closer look at sentence generation after 30 training epochs. Out of 1000 generated sentences, 99.9% were syntactically correct, with Standard Deviation 0.1%. The sentences contained on average 57 idioms[8] (SD=18.7), out of which 99.5% were syntactically correct (SD=1%). Here are examples of the generated sentences (asterisks denote syntactically incorrect sentences):

> '... *Samko gives rabbit a hug. Grandpa gives daddy five. Mia kisses mummy good bye. Daddy sees Winnie the Pooh. Helen bites french fries. *Mummy loves Winnie the Pooh five. *Grandma gives Winnie the Pooh . ...*'

Some of the mistakes were due to stochastic selection, when the most active neuron encoded a grammatically correct sentence continuation, but a different word was actually selected, others were genuine results of the network's internal organization. Nevertheless, we can conclude that a SRN can be trained to produce a mixture of syntactically well-formed regular and idiomatic transitive sentences, including variable-length phrases and discontinuous idioms, without any dedicated mechanism specially designed for idioms. As far as we are aware, this is a novel result.

Study 2: Production of syntactically well-formed and semantically correct sentences containing idioms

Although novel, the results about idioms from the previous section follow fairly straightforwardly from Elman's results (e.g. Elman, 1990, 1991). Syntactically, idioms are surface patterns, and it is well known that SRNs can learn such patterns. In this section, we consider the more contentious issue of how to add semantics to a SRN, to model the process of sentence generation rather than just knowledge of syntax.

[8]The sentences in training set contained on average 85.6 idioms (SD = 3.1), which is about 3 times more (when normalized by a sample size). However, without semantics, we had no means to influence the number of generated idioms during spontaneous production.

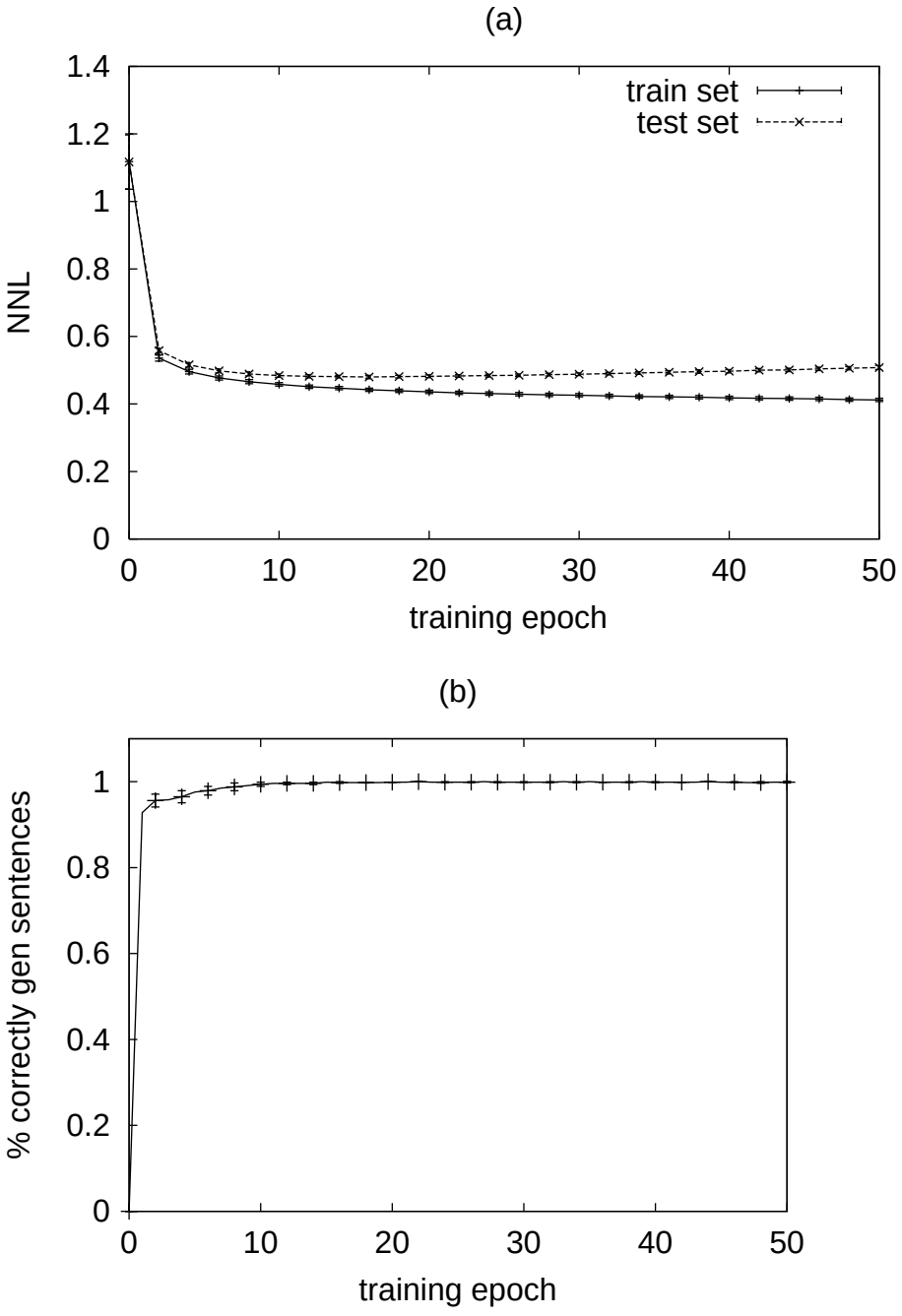

Figure 4. (a) Normalized negative log-likelihood for the next-word prediction task for SRN with syntactic input only. (b) Production performance of the network without semantics. Results are averaged over 5 model subjects (note that both graphs also show standard deviations, which are very small).

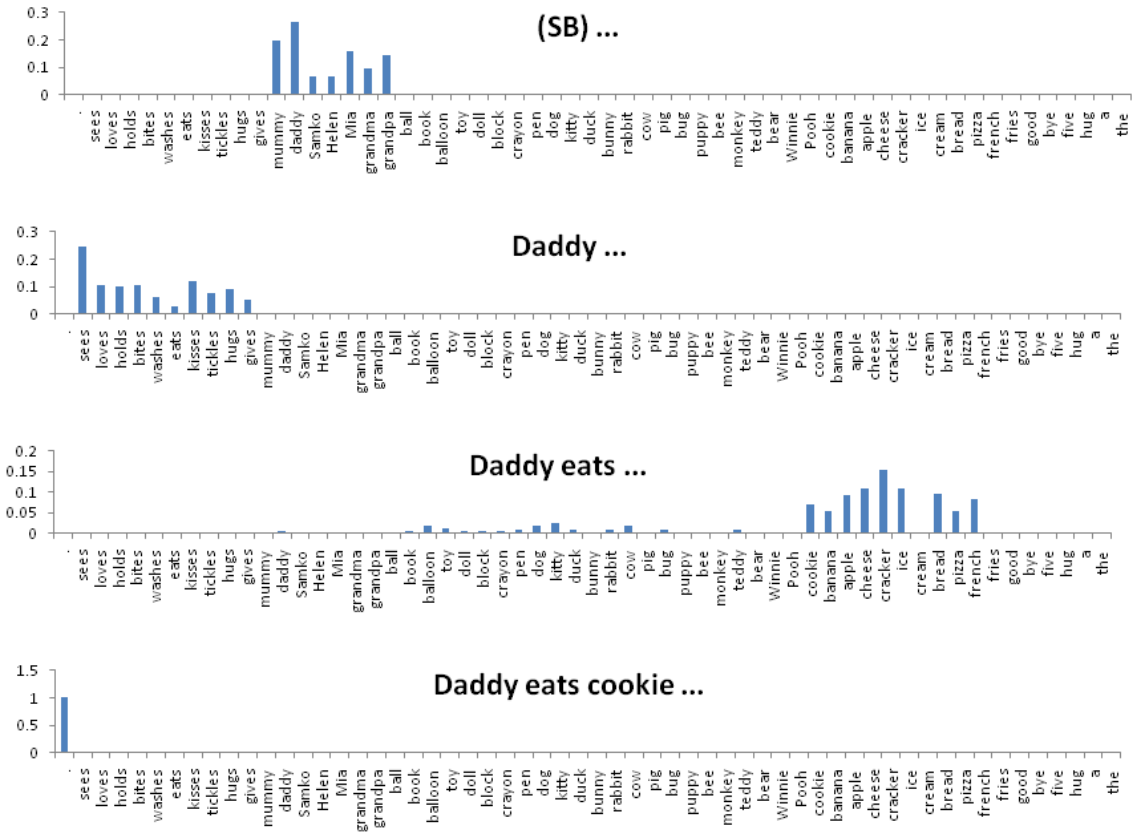

Figure 5. Activities on the output layer during a spontaneous sentence generation after 30 training epochs. Graphs show distribution of probabilities at the SRN output layer for each word from the dictionary with which is to follow the unfinished sentences in the graph title. We can see that the SRN is able to correctly predict the category.

Methods

A model of language production would be utterly implausible if it could only generate random sentences in autonomous regime. That is why we need to enhance our model in a way that it will produce a sentence which expresses a given meaning.

There are several ways in which semantics can be delivered to the input of language production device. Although we believe that it is ultimately necessary to deliver concepts that constitute the meaning of a sentence as a sequence (Knott, 2010), for methodological reasons we start with a much simpler representation, where 'static' semantic representations are presented to the production network.

Language and meanings. We used an artificial language with the same syntax as in Study 1. However, in the second study we added meanings to sentences of the artificial language. For simplicity, we represented the semantics of each transitive sentence by a theta-role frame with three roles AGENT, ACTION, PATIENT (abbreviated as AG, ACT,

Table 3: Synonyms present in our language.

Meaning	Expression 1	Expression 2
PAT: BUNNY	*rabbit*	*bunny*
PAT: TEDDYBEAR	*teddy bear*	*Winnie the Pooh*
ACT: HUG	*gives NP a hug*	*hugs NP*

PAT in the following examples). Each role was bound to a symbolic token representing a particular meaning, for example the token DADDY represented the meaning of the word *daddy* as in *daddy bites bread* represented by AG: DADDY, ACT: BITE, PAT: BREAD.[9]

The tokens were meant as crude abstractions of the large neural assemblies active when we entertain a particular meaning, which for concrete objects and actions are distributed over inferotemporal, motor and prefrontal cortices. While it is clear that these representations are distributed in nature, there is nonetheless some element of locality to them. For instance, animate objects and tools appear to be represented more strongly in different areas of inferotemporal cortex (Damasio, Grabowski, Tranel, Hichwa, & Damasio, 1996; Lu et al., 2002), while actions involving different body parts are preferentially represented in differerent areas of motor and premotor cortex (Hauk, Johnsrude, & Pulvermüller, 2004). Our localist scheme, while obviously just an approximation, can be seen as modelling the coarse-grained localism identified by experiments like these.

Single concepts were always represented by single tokens even if they were expressed by multi-word phrases, for example ICECREAM for *ice cream* (this follows from our initial working characterization of idioms that that they can be substituted by a single word). Meaning of a discontinuous idiom was represented by a single token bound to the ACTION role, for example ACT: FAREWELL for *kisses NP good bye* or ACT: HUG for *gives NP a hug* (meaning of the NP was represented by another token bound to the PATIENT role).

The language also contained several synonyms, i.e. multiple ways to express the same meaning (see Table 3).

Neural model. In our second study, we extended the network from Study 1 so that it receives inputs from semantic representations, as well as just producing syntactic structures. In sentence generation mode, a meaning to be expressed was presented in a new 'semantic input' layer, which was connected to the SRN's hidden layer (see Figure 6). This layer delivered the same tonically active input at each time step of the SRN's subsequent operation, until the network predicted another SB or an arbitrarily defined sentence-length limit was reached. At this point, a new semantic input was specified, and a new cycle started. The architecture of the network is basically the same as that of Chang's (2002) 'prod-SRN' network. Our interest is in whether this network can learn to produce idioms.

In relation to brain areas, the semantic layer with its connections to the hidden layer embodies the idea of a content preparing pathway (Figure 1, top pathway). We envisage that semantic representations of objects and actions are evoked in a range of neural areas. Objects are represented in working memory in anterior Broca's area (left BA45) and adjacent prefrontal areas BA46-47 (see Ungerleider, Courtney, & Haxby, 1998;

[9] We will distinguish concepts and words by capitalization and font: concepts will always be written in all CAPITAL letters, words in (usually) small letters and *italic*.

Table 4: Part of the training sequence for the next word prediction with semantics.

Time step	Semantic input	Current word	Target
1	AG: DADDY, ACT: SEE, PAT: TEDDYBEAR	SB	daddy
2	AG: DADDY, ACT: SEE, PAT: TEDDYBEAR	daddy	sees
3	AG: DADDY, ACT: SEE, PAT: TEDDYBEAR	sees	teddy
4	AG: DADDY, ACT: SEE, PAT: TEDDYBEAR	teddy	bear
5	AG: DADDY, ACT: SEE, PAT: TEDDYBEAR	bear	SB
6	AG: GRANDMA, ACT: HUG, PAT: MIA	SB	grandma
etc...			

Wagner, Maril, Bjork, & Schachter, 2001), as well as in inferotemporal cortex (Damasio et al., 1996), while actions are represented in left dorsolateral prefrontal cortex (Perani et al., 1999; Tranel, Adolphs, Damasio, & Damasio, 2001) as well as in motor and premotor cortex (Pulvermüller, Shtyrov, & Ilmoniemi, 2005). We assume that the circuits which map these representations onto phonological words involve parts of Wernicke's area, but other areas of temporal cortex (see e.g. Damasio et al., 1996; Lu et al., 2002), and connect to the syntactic network via links from Wernicke's area back to Broca's area (perhaps through the direct phonological route through Geschwind's territory proposed by Hickok & Poeppel, 2007).[10]

Recall that we represented concepts by symbolic tokens, and a sentence-sized proposition as a theta role frame with concepts bound to the roles AGENT, ACTION, PATIENT. These roles were coded using binding-by-space (McClelland & Kawamoto, 1986; Chang, 2002), where each role had its own field of units with a dedicated unit for each concept admissible in that particular role (as shown in Figure 6). Hence for example, there was a unit for a concept DADDY in the agent role, and a separate unit for DADDY in the patient role. This method of specifying propositions was criticized by Chang (2002) for its poor generalization ability: identical concepts in different roles are completely unrelated and cross-role generalization is impossible. We are aware of this issue and we return to it in the final discussion. There were 53 semantics units altogether (7 for AGENT, 11 for ACTION, and 35 for PATIENT role) and we increased the size of the hidden layer to 100 units.

Training. Training was similar to that in the previous study. Again, we had 5 different 'model subjects' with 500 sentences from the artificial language in their training sets and 1000 (different) sentences in the test sets. However, encoding of the sentences into a training sequence was now different because the sentences were accompanied by their meanings (see Table 4). The network input was encoded by a concatenated vector of 1-hot encodings of agent, action, patient, and current word (together 111 binary digits, out of which exactly four were equal to 1). A 1-hot-coded next-word binary target vector had 53 binary digits. We used the back-propagation training algorithm with the same parameters as in Study 1.

In order to evaluate production performance, *meanings* of the 1,000 test sentences

[10] In our SRN model, semantic representations connect with the hidden layer of the SRN rather than with phonological representations directly. But it is quite consistent with the above localisations to think of the hidden layer as carrying a mixture of phonological and syntactic information.

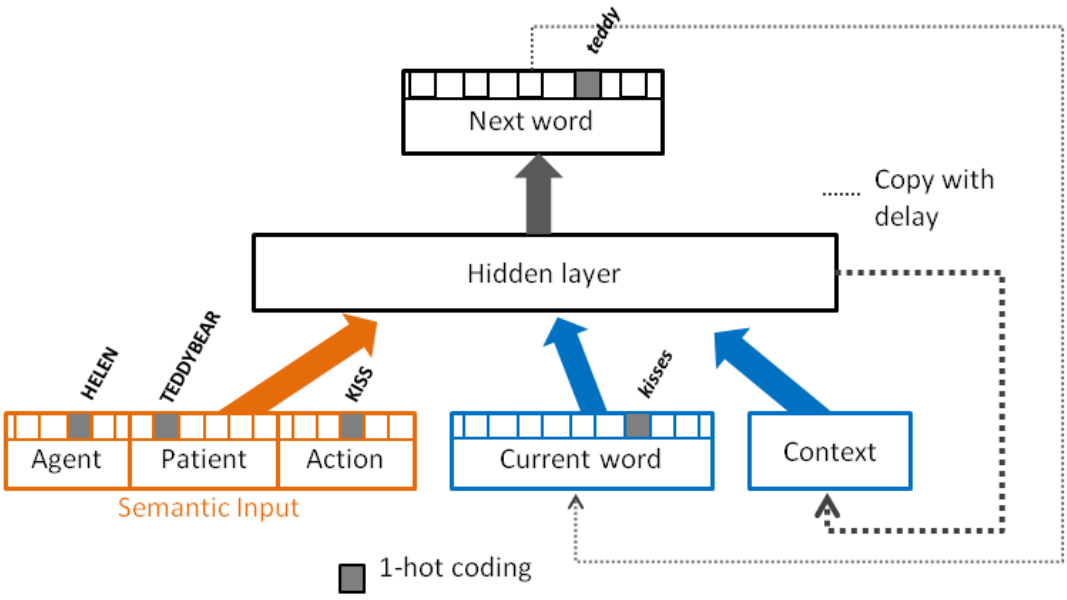

Figure 6. The SRN architecture enhanced with a semantic input. In language production mode, the meaning (e.g. AG:HELEN, PAT:TEDDYBEAR, ACT:KISS) is delivered to the semantic input and each predicted next word (e.g. *Helen*, *kisses*, *teddy*, *bear*, *SB*) is fed back to the input until the sentence boundary (SB) is predicted. Then a new sentence meaning is delivered in semantic input.

(together with SB) were presented to the network with frozen weights one by one after each training epoch. In each time step, the word to produce was again chosen stochastically and then fed back to the current word input. Sentence meaning stays unchanged until the network predicts SB, when the next sentence meaning is delivered to the semantic input (SB is forced if not predicted within 20 steps).

Results

Results of training for next-word prediction can be seen in Figure 7a. With the help of semantic information, the network learned to predict a specific word, not just a category. However, the task was not completely deterministic even now, due to possible synonyms (e.g. a concept PAT: BUNNY could be expressed as *rabbit* or *bunny*). Nevertheless, the NNL tended to get reasonably close to zero even on the test set.

In this network, sentence generation was no longer 'autonomous', but driven by the message on semantic input. A generated sentence was considered correct, if it was correct syntactically (i.e., it belonged to the language) *and* semantically (i.e., concepts in all three roles of the given message were expressed by appropriate words/phrases in appropriate syntactic positions). For some meanings, multiple correct sentences were possible due to synonymy. As we can see in Figure 7b, the network quite quickly achieved a high level of production performance. The stochasticity did not create so much variation as in the no-semantics case, as there is usually a clear winner on the output layer (see Figure 8) now.

Out of 1,000 sentences generated after 50 training epochs, 96.7% were correct

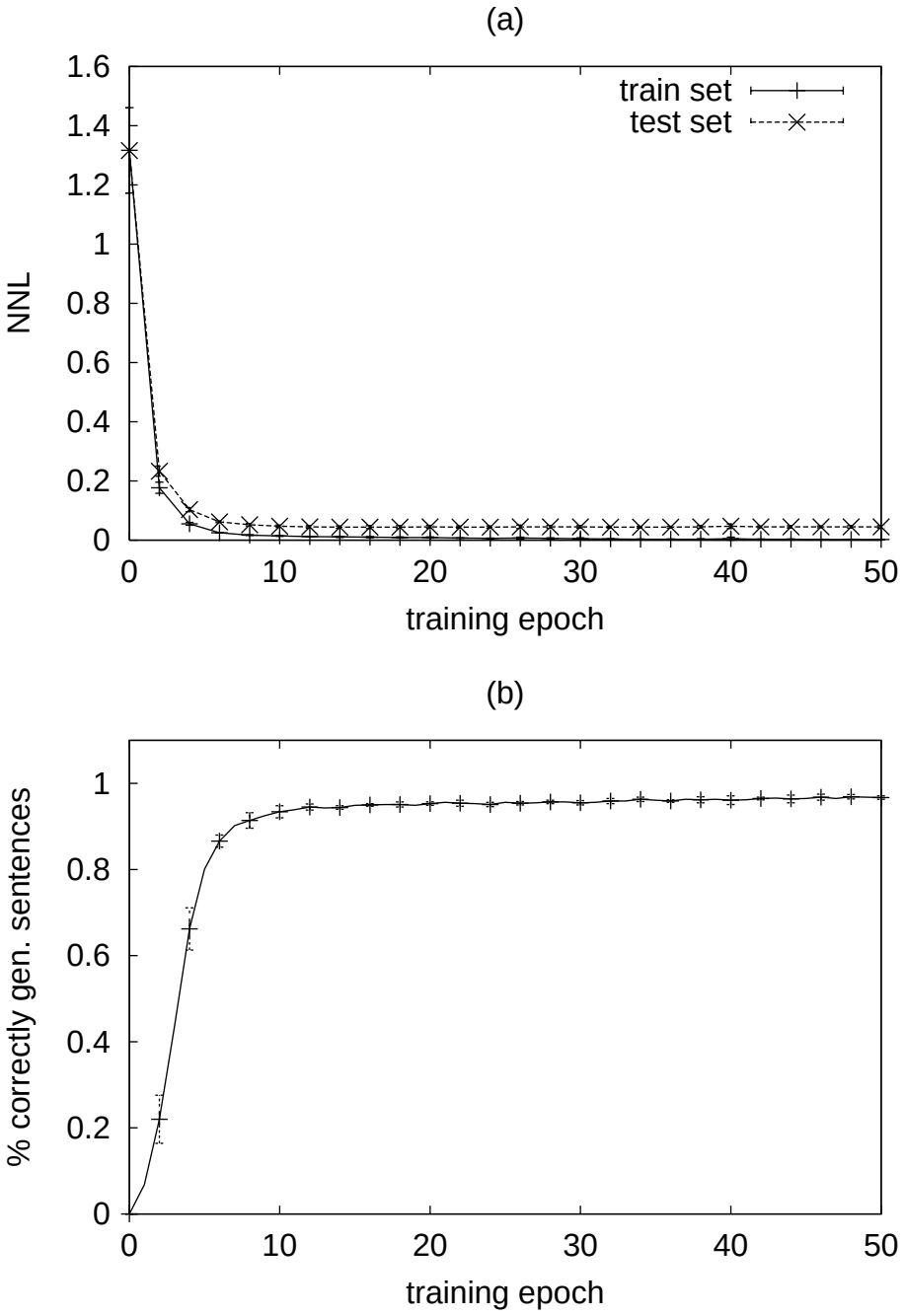

Figure 7. (a) Normalized negative log-likelihood for the next-word prediction task in the network with semantics. (b) Semantics-driven production performance. Results are averaged over 5 model subjects (note that both graphs also show standard deviations, which are very small).

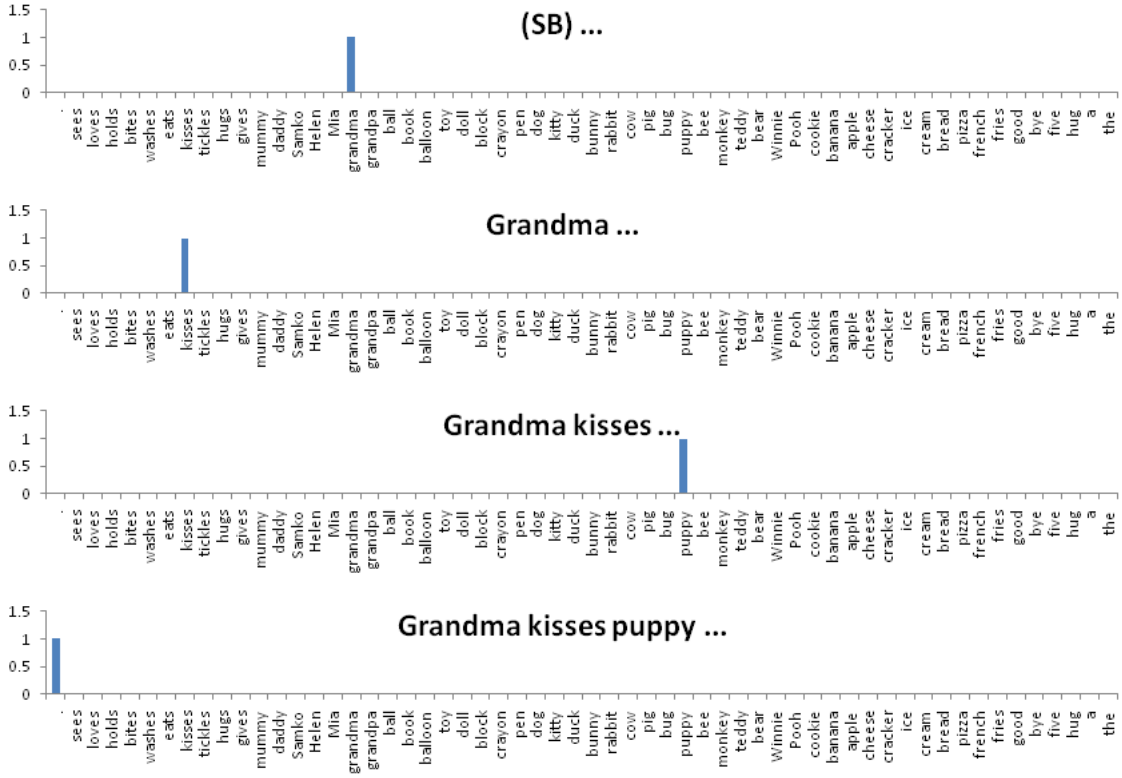

Figure 8. Activities on the output layer during a sentence generation for a sentence meaning AG:GRANDMA, ACT:KISS, PAT:PUPPY after 50 training epochs. Graphs show distribution of probabilities at the SRN output layer for each word from the dictionary with which is to follow the unfinished sentences in the graph title. The semantic input helps the SRN predict a specific word, not just a category (compare Figure 5).

(SD=0.4). The generated sentences contained on average 206.4 idioms[11] (SD=11.6), out of which 93.7% were correct (SD=1.7). Here are examples of sentences produced for given messages:

AG: GRANDPA, ACT: BITE, PAT: CRAYON → *Grandpa bites crayon.*
AG: HELEN, ACT: HUG, PAT: BEAR → *Helen gives Winnie the Pooh a hug.*
AG: GRANDMA, ACT: FAREWELL, PAT: RABBIT → *Grandma kisses rabbit good bye.*
AG: GRANDMA, ACT: HUG, PAT: BEAR → **Grandma hugs Winnie the Pooh a hug.*

Discussion

The SRN has mastered its task of learning to produce semantically correct and syntactically well-formed regular and idiomatic sentences to express given meanings. This task

[11]For comparison, the test set based on the same meanings that were used for generation contained on average 205.6 idioms (SD=8.7).

can be analysed as having several different components. The task of 'completing a continuous idiom' can be seen as a purely syntactic one. Once we have started generating an idiom for a given meaning, subsequent words neither add any new semantic content to the idiom's whole meaning, nor require additional semantic information, so a regular Elman network with no semantic input should be able to accomplish this task comfortably. The task of predicting words in a regular transitive sentence requires a semantic input in each step to disambiguate next-word prediction. At each time step, the updating context units select words which are syntactically appropriate in the current context, while the semantic input units provide a tonic bias towards words which are semantically appropriate. The task of generating a discontinuous VP idiom appears the most difficult: the network must start generating a VP idiom, then put this process 'on hold' while generating an intervening NP, before finally resuming the 'syntactic' task of idiom completion. (Note that generating the intervening NP itself requires a mixture of semantic and context information—and may even contain a nested idiomatic structure of its own.) However, these apparently heterogeneous tasks are all accomplished within a single SRN architecture, by a single sentence production mechanism.

Our architecture seems suitable for capturing the constructivist idea that there is no clear boundary between syntactically productive constructions and idiomatic ones, but rather a continuum of gradations between these two extremes. However, in order to assess this hypothesis in more detail, it is important to gain deeper understanding of what is happening 'inside' the network: how exactly the internal organization of distributed representations on the hidden layer contributes to the overt behaviour. We deal with this issue in the next section.

Hidden space organization

Methods

The hidden space of an Elman network is a N-dimensional space, where N is the number of neurons in the hidden layer. Each *state* – a particular configuration of activities of the hidden-layer neurons is a point in this space and a temporal sequence of states can be thought of as a trajectory in the hidden space (Elman, 1991). Because of its recurrent connections, the hidden layer has a special function of encoding previous history in the dynamic of the network. The actual configuration of the activities reflects not only the current and immediately previous state, but the whole of its history. Even an untrained SRN has a property called *architectural bias* in that points representing temporal sequences with a common history are close to each other in the hidden space (Cernansky et al., 2007). However, while the representation of history is theoretically unlimited, in practice it is limited by the numerical precision with which neuron activities are represented, and thus fades away with time.

If the activation functions of the output layer are linear (as is the case of our model), each output neuron computes a scalar product of its weight vector with the hidden layer activity configuration, hence effectively responds to those configurations that are most similar (have the least angle) with its weight vector. If each output neuron represents one of the possible next words in the dictionary, it is desirable that the sequences (partial sentences) that can be completed with a particular word be represented by similar hidden layer config-

urations. The difference between an untrained and a trained SRN is that, during training, the hidden space reorganizes so that categories of sequences that should yield a similar outputs are clustered together (Cernansky et al., 2007).

For this reason, it is very instructive to study what internal representations the SRN has developed to solve a particular problem. Elman (1991) even argues that hidden layer activities in a trained SRN can serve as distributed representations of words that reflect all their characteristic syntactic and semantic properties without the necessity of encoding these properties manually and explicitly. In this section, we describe our findings about the hidden space organization in the models without (Study 1) and with (Study 2) semantic input.

A useful way of studying the nature of internal representation consists in hierarchical clustering of hidden layer activities, which can be visualised by dendrograms that organize the closest clusters in a tree-like structure (e.g. Elman, 1990). After training, the SRN's weights are frozen, the complete training set is swept through the network word by word again, and the hidden layer activity patterns are recorded. As each word appears many times in the training sequence (in different contexts), the patterns are grouped by words and all activity patterns for a particular word are averaged to obtain 57 centroids of regions of activities elicited by the 57 words in the dictionary. Dendrograms show the hierarchical clustering of these 57 vectors.

Results

Figure 9 shows the dendrogram for the network without a semantic input (Study 1). We can see that all objects group together, as well as all people,[12] verbs, and idioms without their first words. The first words of idioms are not clustered with other parts of idioms; the effect of previous context keeps them close to their respective categories of verbs and objects, nevertheless they are separate, because they set off a special idiomatic trajectory. Interestingly, within the cluster for idioms, we can observe subclusters corresponding to words occurring in particular ordinal positions, regardless of the idiom type. For instance, there is a cluster of words which occur as the second word of a two-word idiom – the second words of the NP idioms *teddy bear*, *ice cream*, *french fries* cluster together with the second word of the discontinuous VP idiom *give five*. Also, there is a cluster of words which occur as the third words of a three-word idiom – the word *Pooh*, which is part of the continuous NP idiom *Winnie the Pooh*, groups with words *hug* and *bye*, which are parts of discontinuous VP idioms *give a hug* and *kiss good bye*. The second words of the three-word idioms (*the*, *a*, and *good*) belong to the large cluster of idioms too, but their representations are rather distant from each other, because they set off separate idiomatic trajectories. This suggests that the SRN has internally developed something like a counter that keeps track of the position within an idiomatic phrase.

Traces of a counter-like structure can also be seen in the dendrogram for the model with semantics (Study 2, Figure 10), but some clusters are now different because of a strong influence of semantic information. While the basic categories of people, objects, verbs, and idioms are retained, idioms group internally more by meaning now, i.e. *the*, *Pooh* and *bear*

[12] We use the term 'people' rather than 'subjects', because people can appear also in the object role (while no member of the category named 'objects' cannot appear in the subject role).

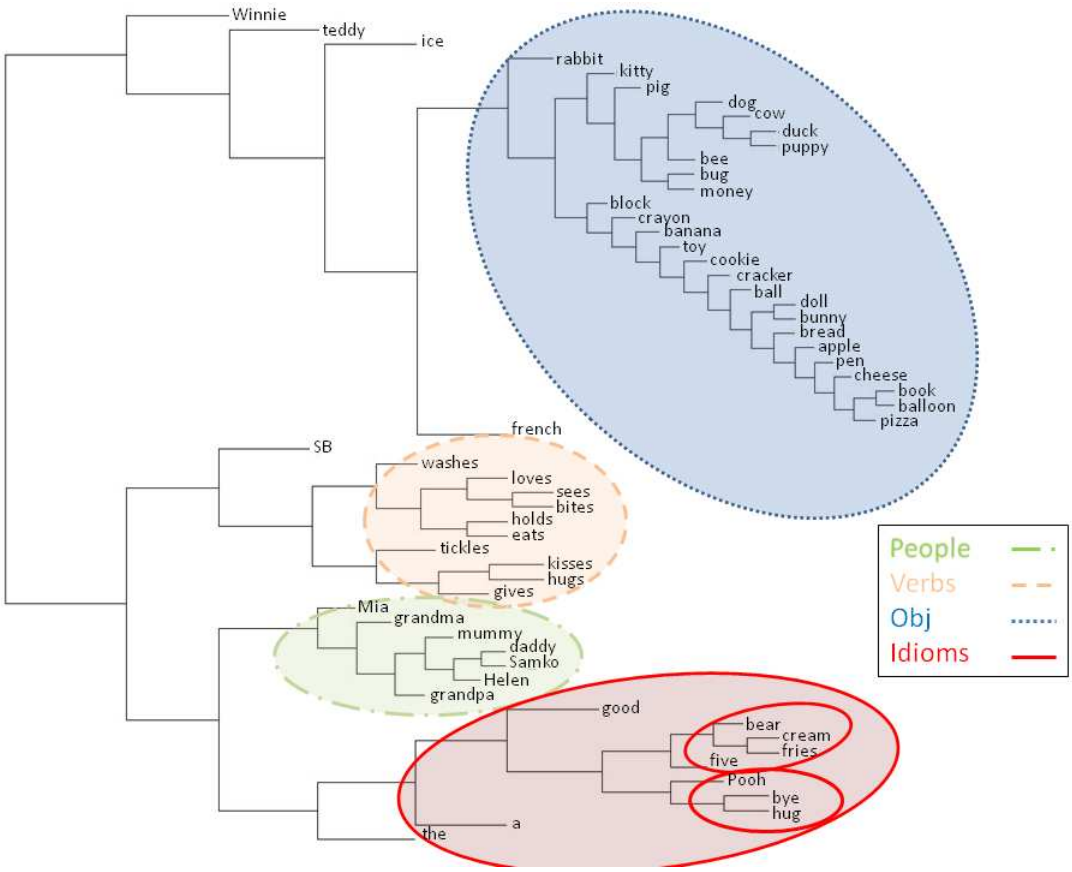

Figure 9. Dendrogram of average activity patterns for each input word in the network *without* a semantic input (Study 1). Clusters corresponding to categories of people, verbs, objects and idiomatic phrases are visible. Category of idioms further subclusters according to position of a word within an idiom (regardless of the idiom type).

form a cluster, as well as *good* and *bye*, or *teddy* and *Winnie*. Note also that synonyms *rabbit* and *bunny* are together, as well as verbs *hugs* and *give* (being part of *give a hug*).

While dendrograms schematically depict clustering relations between *centroids* of actual activation patterns for words in their contexts, it would be useful to analyse the actual activation patterns themselves. For this purpose, we use 2-dimensional projections of the hidden space based on Principal Component Analysis (PCA). PCA is a method for finding directions of maximal variance in the data (principal components) and projecting the original data vectors to the space defined by these components (Haykin, 2008). Components are usually ordered by their magnitude (importance), the first component being the most important. As data vectors for obtaining principal components, we use the same hidden layer activation patterns that we used for the dendrograms just discussed (but now without averaging by words). After acquiring principal components, we can visualize patterns for particular sentences, categories or other items of interest in 2D spaces (planes) defined by selected principal components. As the same set of patterns can appear differently in different planes, we can even speculate about the interpretation of particular principal components.

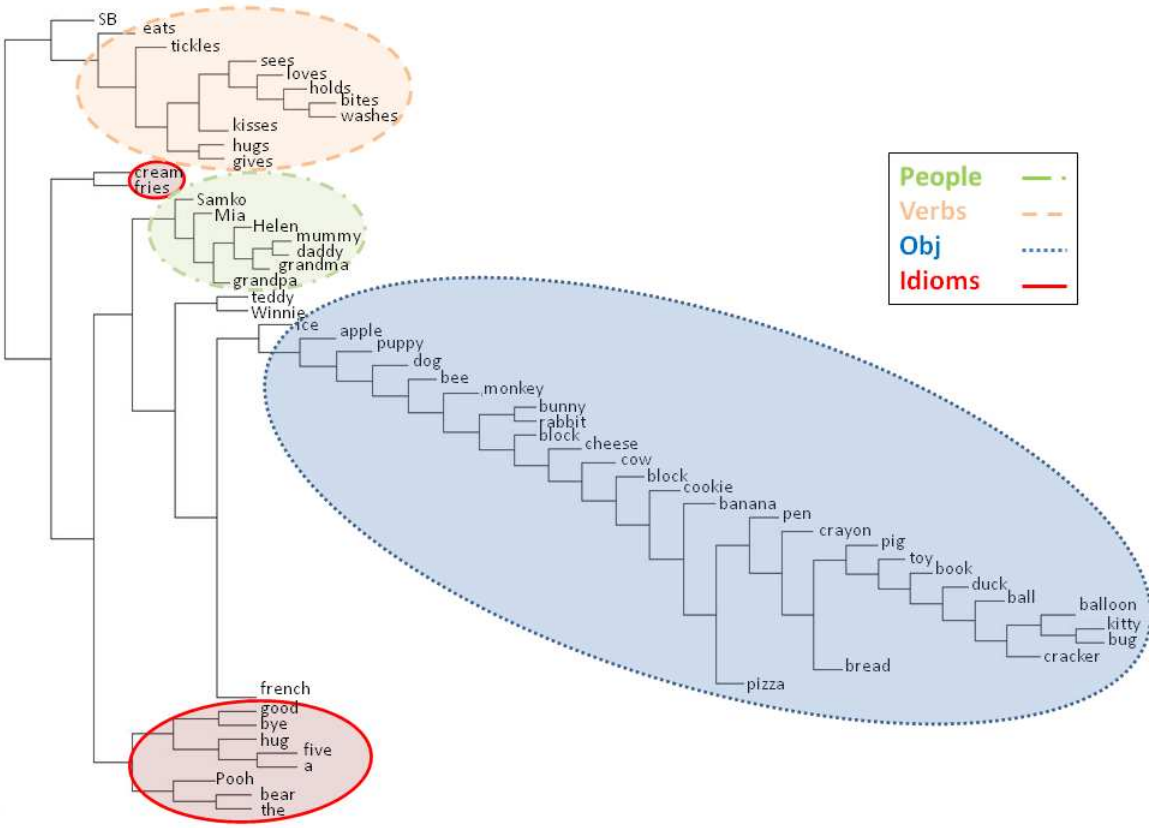

Figure 10. Dendrogram of average activity patterns for each input word in the network *with* a semantic input (Study 2). Clustering by syntactic role is now influenced by meaning (synonyms are clustered together).

We focus on the results obtained for activation patterns in Study 2, i.e. the model with semantic input. Figure 11 shows all activation patterns in the plane of the first two principal components, color/texture coded by categories defined by structural position of words that elicited them in the sentence (i.e., all SBs have the same color/texture, as well as all subjects, verbs, 1st, 2nd and 3rd words of object, and 2nd and 3rd words of discontinuous idioms). We can see that they cluster nicely, which implies that the first two principal components reflect structural information. If we viewed the same structural categories in e.g. planes of 5-8 principal components (not displayed here), we could not observe any clusters and all color codes would be intermingled. This is because these components mostly reflect semantic information as can be seen on Figure 12. Now the same color/texture code is used for all the words of sentences with the same meaning, e.g. activity patterns elicited by words *SB*, *grandpa*, *eats*, *apple* while expressing the sentence meaning AG:GRANDPA, ACT:EAT, PAT:APPLE share the same code. Note that the sentence meaning AG:MUMMY, ACT:HUG, PAT:RABBIT can be expressed by synonymous sentences *Mummy hugs rabbit*, *Mummy hugs bunny*, *Mummy gives rabbit a hug* and *Mummy gives bunny a hug*, so the patterns elicited by all these words while expressing this particular

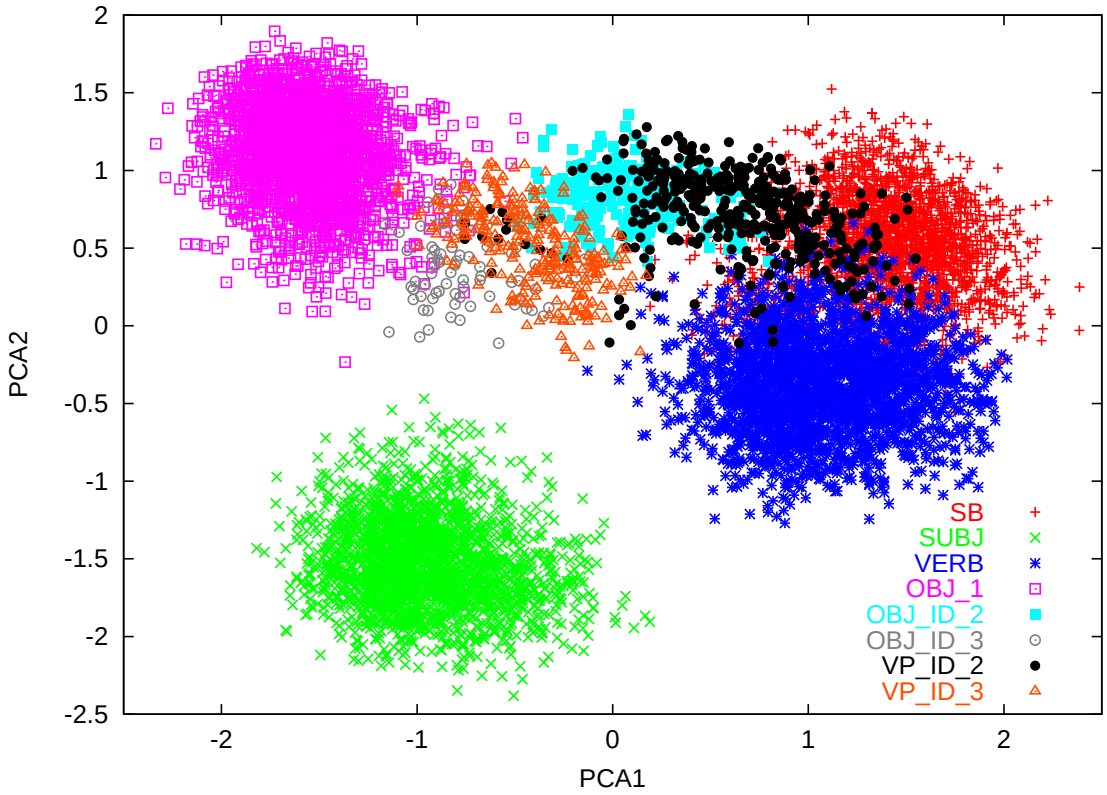

Figure 11. Activity patterns elicited by all words in training sequence (not averaged) shown in PCA1-2 plane. Words with identical syntactic positions in sentences share the same color/texture codes. OBJ-1 = 1-word object or the first word of a NP idiom, OBJ-ID-2/3 = second/third word of a NP idiom, VP-ID-2/3 = second/third word of a discontinuous VP idiom. Syntactic clustering is clearly visible.

meaning[13] share the same code too. Although the words appear in different structural positions within their sentences and some of them are parts of idioms while others are not, we can see that their activities cluster quite nicely in the plane of 5-6th principal components.

Discussion

The dendrograms and principal component analyses just discussed lead us to the hypothesis that different subspaces of the hidden space encode different types of information during sentence generation. Sentences with similar structures will follow similar trajectories when projected to a subspace of components that encode mostly structural information; sentences with similar semantics (i.e. sharing some semantic roles) will have similar trajectories when projected to components that encode semantic information. Remember that a role of each output neuron that represents a particular next word is to compare the angle of its weight vector with the hidden-layer activity pattern vector. The activity pattern

[13]The same words can also participate in expressing different meanings; the activation patterns of these were excluded from consideration.

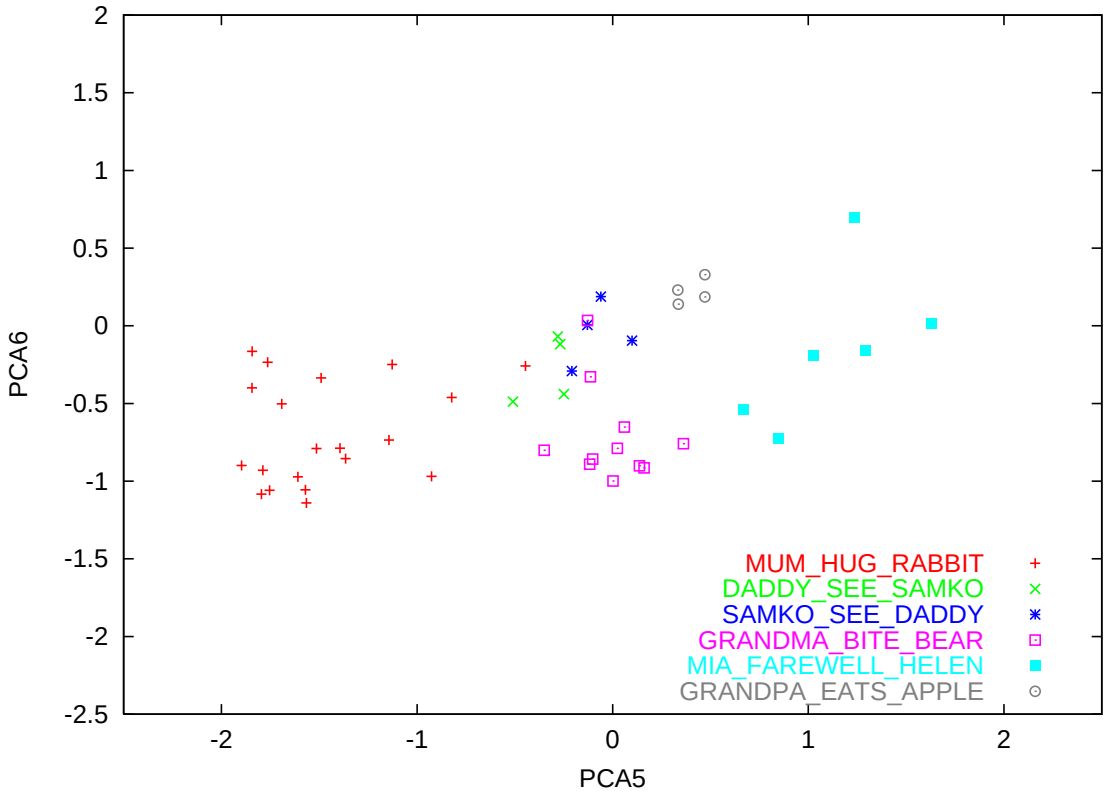

Figure 12. Activity patterns of all words generated for selected sentence meanings shown in PCA5-6 plane. All words of all synonymic variants of a sentence share the same color/texture codes regardless of their syntactic position. Semantic clustering (by sentence meanings) is visible.

vector is equal to a linear combination of its principal components (with earlier principal components having the most influence); which means that syntactic and semantic aspects jointly contribute to the decision. As we will show in the next section, representation of semantic and structural aspects can be selectively damaged by lesioning different parts of the network.

Lesioning the network

Impairing neural networks, whether by adding excessive noise, zeroing or decreasing connection weights, has been widely used in many connectionist models in order to show graceful degradation of performance, to fit empirical data or to show double dissociations. In this section, we consider the effects of separately lesioning the 'semantic' and 'recurrent' inputs of our network. The basic idea of reproducing symptoms of Broca's and Wernicke's aphasia by selective lesions of this sort is common currency in neural network models. However, to our knowledge, no-one has examined in detail the effects of such lesions in a simple network like ours.[14] We first demonstrate that we can indeed produce Broca- and

[14]Chang (2002) discusses the effect of selective lesions in a more complex network, but not in the simple 'prod-SRN' model.

Wernicke-like symptoms by selective lesions in our simple network. However, our main focus is on seeking to explain how these symptoms result from changes in the organisation of the network's hidden space, drawing on the hidden space analysis given earlier in the paper. In addition, we want to investigate the network's ability to generate idioms in the presence of selective lesions. It turns out that different lesions have different effects on the network's performance on idioms, creating some testable predictions for our account of idioms.

Methods

We only performed lesion studies with the SRN with semantic input (Study 2). As motivated in the section describing the architecture of this model, we will denote the connections from the semantic input to the hidden layer as *semantic connections* and those from current word and context units to the hidden layer as *syntactic connections*. As detailed in the next sections, our basic result is that lesioning syntactic connections leads to selective syntactic deficits, while lesioning semantic connections leads to selective semantic deficits.

We took the network connections after 50 epochs of training and selectively lesioned (zeroed) a certain ratio of (randomly chosen) semantic or syntactic connections. Then we froze the weights and measured the sentence generation performance on the same set of meanings that were used for the unlesioned network. For each ratio of lesions, we repeated lesioning 5 times, one for each model subject, each time doing an independent random selection. All the results were averaged over all 5 model subjects thereafter.

To be able to better assess the degree of performance deterioration, we needed to evaluate semantic and syntactic correctness separately and adopt finer measures. A sentence produced for some meaning (delivered on semantic input) was considered *syntactically correct*, if it belonged to the language defined by the rules in Figure 2 (regardless of the relevance to the meaning that had to be expressed). As a degree of *semantic relevance*, we used the relative proportion of input semantic roles that could be traced among words in the sentence (regardless of their order, completeness or grammatical correctness). For example, in the phrase '*ice ice daddy ice eats ice*' produced for the meaning AG: DADDY, ACT: EAT, PAT: ICECREAM we could trace attempts to express concepts in all the three roles (sem. relevance 1), while in '*Daddy eats pizza.*' produced for the same meaning just two roles were expressed correctly (sem. relevance 0.66).

Results of lesioning syntactic connections

Quantitative results of lesioning different proportions of syntactic connections can be seen in Figure 13. We can see that the total performance deteriorates with increasing number of lesioned connections. Syntactic competence follows the total performance – indeed, if we damage 60% of the connections, the network hardly ever produces a syntactically correct sentence. However, the semantic relevance is almost 0.6, which means the network is still able to express on average 2 out of 3 semantic roles. If we completely cut off the current word and context connections (and the network stops being recurrent at all), the network still produces a correct word for one of the roles on the semantic input (sem. relevance around 0.3). As the network is no longer recurrent, it cannot get past the first generated word (except for quite rare variations caused by stochastic selection) until the sentence length limit (20 words) is reached and a new meaning is delivered on the semantic input.

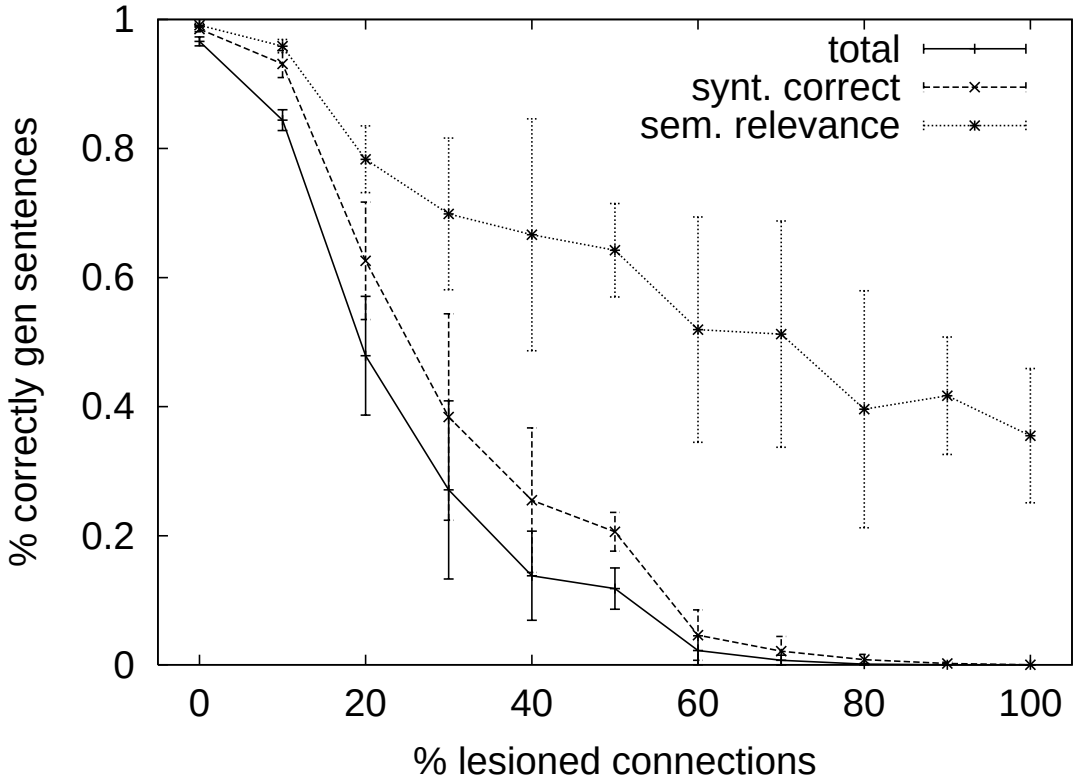

Figure 13. Sentence production in a network with lesioned *syntactic* connections. *Total* means the relative number of syntactically *and* semantically correct sentences. *Semantic relevance* means the average number of relevantly expressed semantic roles per a sentence. All results are also averaged over the 5 model subjects.

Here are some examples of sentences generated for given meanings by a network with 90% of syntactic connections lost:

AG: SAMKO, ACT: WASH, PAT: BOOK → * *Samko.*
AG: DADDY, ACT: LOVE, PAT: COW → * *Loves cow cow loves loves loves loves loves cow cow loves loves cow cow loves cow loves loves cow cow* (sentence length limit reached)
AG: SAMKO, ACT: WASH, PAT: FRENCHFRIES → * *Samko fries french french Samko french fries fries french fries french fries fries washes fries french french fries* (sentence length limit reached)

We can see that a SRN with seriously lesioned syntactic connections is unable to produce syntactically correct sentences. It often repeats just one or two words. The produced words are semantically appropriate. This roughly corresponds to the clinical picture of Broca's aphasia. Broca's aphasics have difficulties with grammar and syntax; the rules of grammar and word order are impaired, especially in longer utterances (Tanner, 2008).

The speech is non-fluent, agrammatic and often telegraphic (Alexander, 2005).[15] According to Kolk (2006), these symptoms are partly the result of corrective adaptation to problems with temporal coordination of syntactic elements, e.g. nonfluency and repetitions could correspond to a covert restart of the sentence production process.

Results of lesioning semantic connections

Figure 14 shows the results of lesioning *semantic* connections, i.e. those from the semantic input to the hidden layer. Again, the total performance deteriorates with increasing number of lesioned connections, but now the picture is different from the previous case. The network maintains high degree of syntactic correctness,[16] while at the same time the sentences become less and less semantically relevant to the meanings that should be expressed. We can see it clearly from the following examples of generated sentences (the network had 90% of semantic connections lost). A hash (#) denotes semantically incorrect sentences (i.e, with semantic relevance less than 1).

AG: SAMKO, ACT: LOVE, PAT: BEAR → # *Mia loves cow.*
AG: MUMMY, ACT: EAT, PAT: APPLE → # *Mummy kisses puppy.*
AG: MIA, ACT: LOVE, PAT: PEN → #* *Mummy bites Winnie.*
AG: MUMMY, ACT: HUG, PAT: DUCK → #* *Samko gives Helen.*

We can conclude that a SRN with seriously lesioned semantic connections produces mostly syntactically correct, but semantically inadequate sentences. This roughly corresponds to the clinical picture of Wernicke's aphasia. The speech of Wernicke's aphasics is usually fluent and to a large extend grammatical; but content may be extremely paraphasic[17] or even empty (Alexander, 2005).

The performance results of simulations with lesioned networks confirmed that syntactic and semantic competences are related to connections in parts of the network that we denoted as syntactic and semantic, and can to a large extent be selectively damaged. In the following section, we will try to look for the evidence of decomposition of representation in the hidden layer.

Hidden space organization in lesioned networks

In the section on hidden space organization in the unlesioned network, we showed that syntactic and semantic information is visible in different principal components of the space of hidden layer neuron activations. Recall that Figure 11 shows clustering of syntactic categories in PCA1-2 and Figure 12 shows clustering of semantic categories in PCA5-6 subspace. By way of comparison, Figures 15 and 16 show activation patterns for exactly the same stimuli and with the same color/texture coding, but for the network obtained from the original one by lesioning 100% of syntactic/semantic connections respectively.

As we can see, clustering of syntactic categories is well preserved if we lesion *semantic* connections (Fig. 15a). The loss of semantic information causes that the clusters for SB and

[15] Our model is not detailed enough to account for other symptoms of Broca's aphasia such as omission of function words, morphological and phonological errors, differences in noun/verb production etc.

[16] But see also section on idioms in lesioned network below.

[17] Again, our model is not detailed enough to reproduce particular symptoms, such as semantic paraphasias (substitutions of semantically related words).

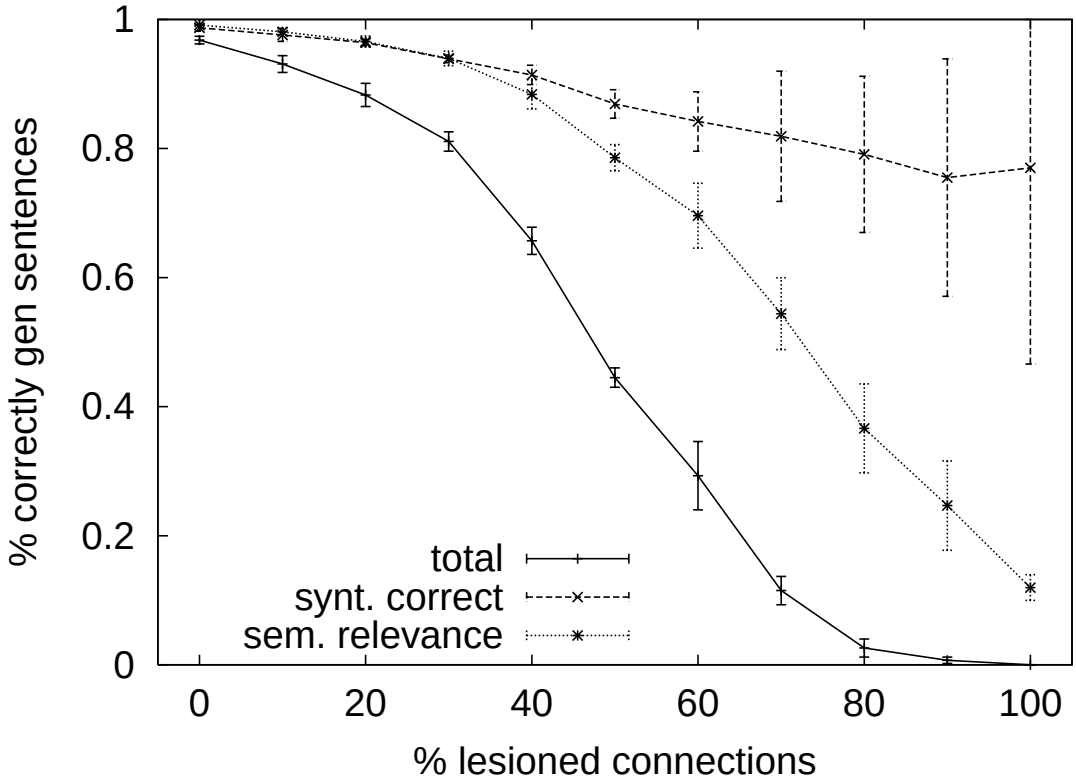

Figure 14. Sentence production in a network with lesioned *semantic* connections.

SUBJ are now divided into smaller subclusters that represent the influence of the previous sentence (this influence would be eliminated in unlesioned network by arrival of a new meaning on the semantic input). The clusters are now smaller in size, but they generally remain in areas within the original clusters of the unlesioned network (compare Figure 11)[18] (when projected to principal components that encode mostly syntactic information). If, on the other hand, we lesion syntactic connections, clusters of syntactic categories collapse and are intermingled (Fig. 15b).

Now we consider the effects of lesions to hidden space organization along principal components that mostly encode semantic information. If we lesion semantic connections, clustering of semantic categories is destroyed (Fig. 16). If we lesion 100% syntactic connections, original clusters reduce to points for the first produced word (as without a context signal, the SRN cannot get past producing the first word of a sentence). However, these 1-point activation patterns across synonymous sentences that express the same meaning (e.g. *Mummy hugs rabbit*, *mummy hugs bunny*) are highly consistent.

Taken together, these results suggest that the SRN has developed largely independent representations of syntactic and semantic knowledge in its hidden space, which can be selectively damaged by lesioning connections in the respective parts of the network.

[18]However, the shift to subareas of original clusters affects production of idioms; we will discuss this problem later.

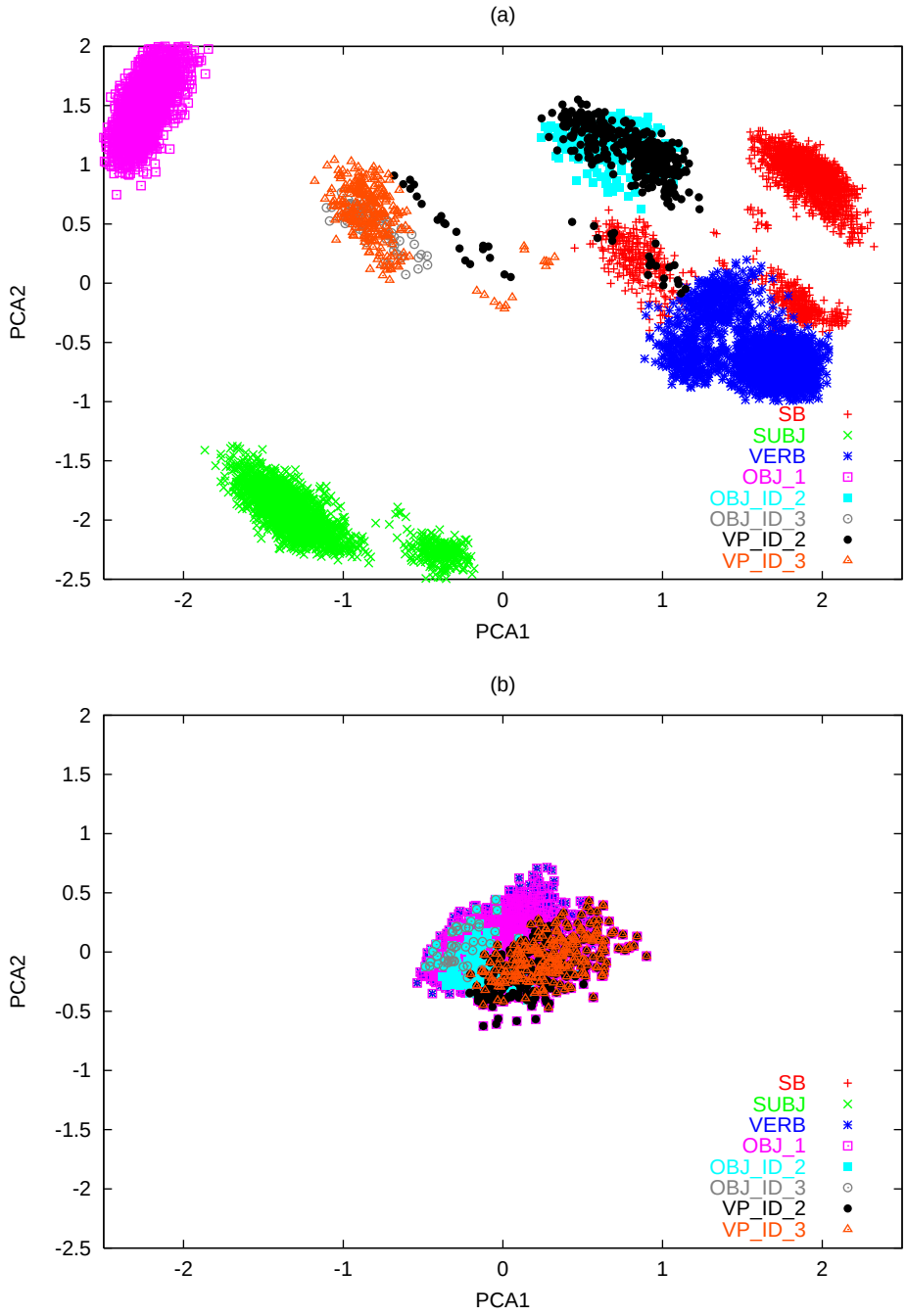

Figure 15. Activity patterns of words in the training sequence color/texture-coded by syntactic roles in the network with 100% semantic (a) and syntactic (b) connections lost shown in the PCA1-2 plane. Clusters are preserved in the former case and destroyed in the latter (compare Figure 11).

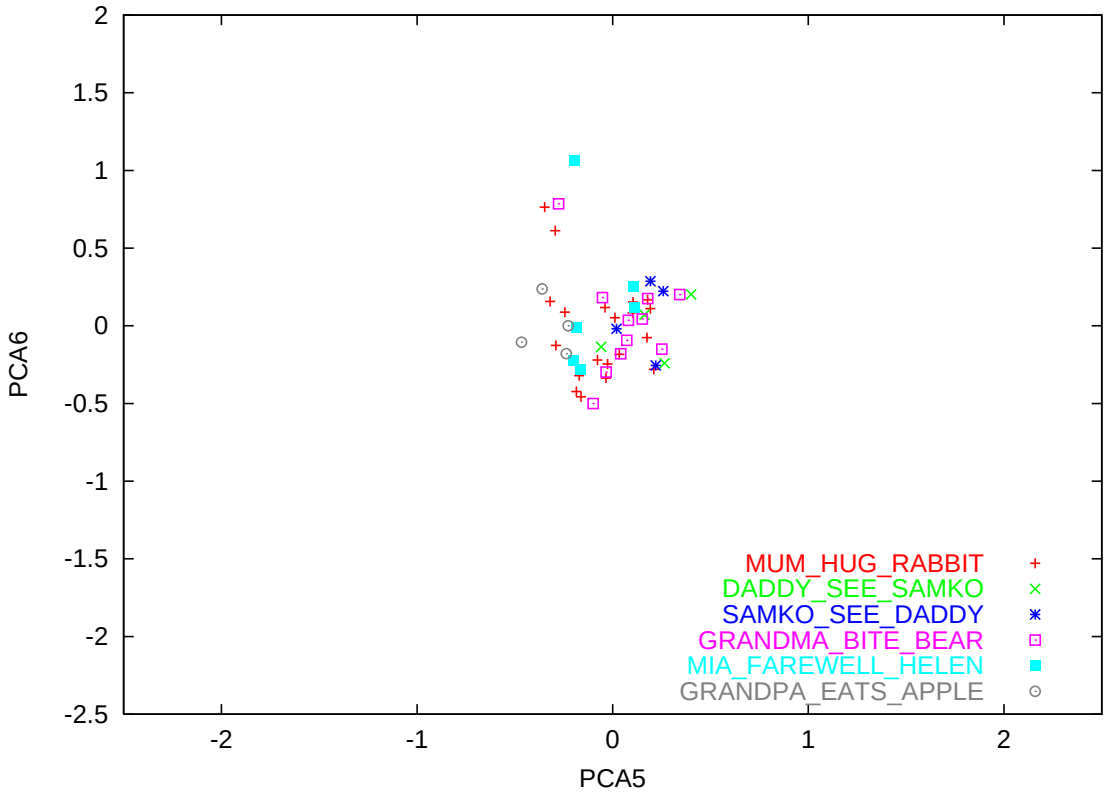

Figure 16. Activity patterns of words for selected meanings color/texture-coded by sentence meanings in the network with 100% semantic connections lost shown in the PCA5-6 plane. Clusters are destroyed (compare Figure 12).

Idioms in lesioned networks

As the syntax and structural flexibility of idioms is to a large degree arbitrary, it is often conjectured that syntactic constraints of particular idioms are stored as a dictionary-like entries, i.e. in a lexically specific way (Jackendoff, 1997; Sprenger et al., 2006). On the other hand, based mostly on experiments with structural priming, others (Konopka & Bock, 2009) maintain that even in the case of idioms more general abstract syntactic mechanisms guide their production. Brain-imaging studies that we reviewed above suggest that areas related to conceptually driven lexical (lemma) selection are left mid MTG and left posterior STG. These areas are often damaged in Wernicke's aphasics, who make semantic errors (Alexander, 2005). More general sentence-level syntactic encoding is related to the posterior part of Broca's area (BA 44), which is often damaged in Broca's aphasics (Alexander, 2005). We are unaware of quantitative data on the production of idioms in aphasics;[19]

[19] *Some* Broca's aphasics are good at producing overlearned phrases and automatisms (Tanner, 2008). However, the data are not detailed enough to tell whether these phrases contained idioms too. A patient with damaged Broca's area and preserved ability to produce idioms would present a challenge to our claim that, besides lexically-specific information, general sentence structure building (sequencing) mechanisms are necessary for idiom production.

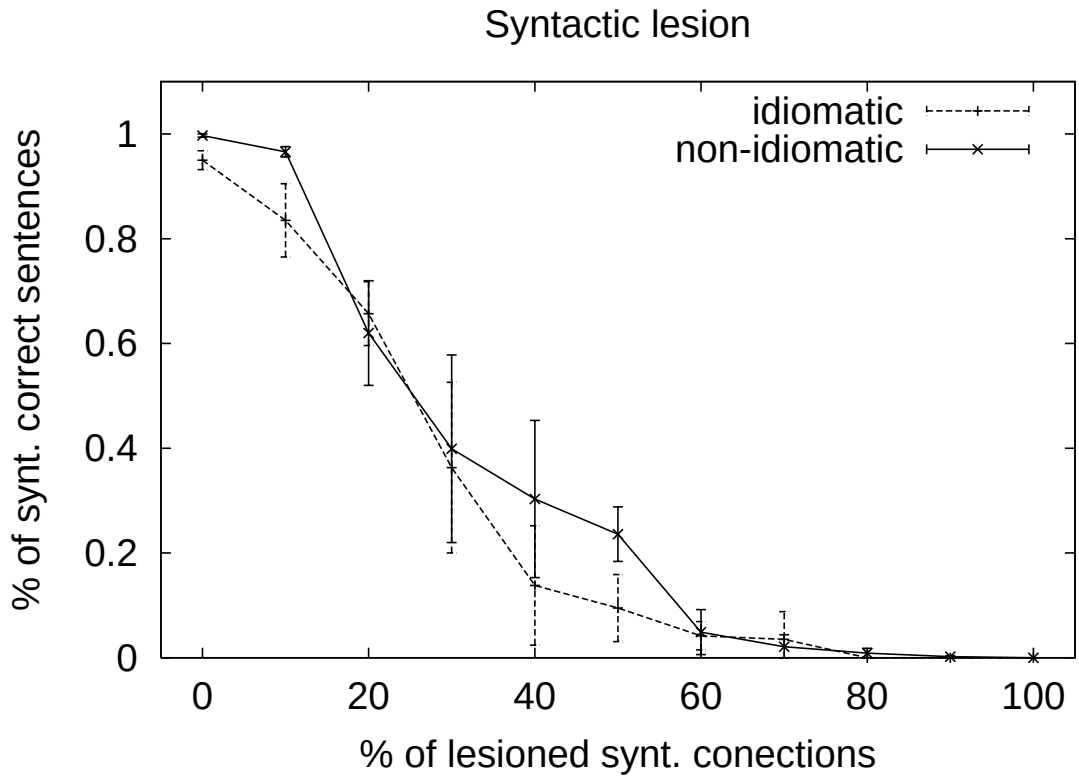

Figure 17. Percentage of syntactically well-formed sentences with and without idioms generated by the network with lesioned syntactic connections.

however, our model of idiom production can make some predictions. If the production of idioms depends on general as well as lexically-specific mechanisms (the claim which is embodied in our model, and which follows from Sprenger et al., 2006), we would predict that production of both idioms and non-idioms would be impaired in Broca's aphasics, while in Wernicke's aphasics, production of non-idiomatic sentences would be better than that of idioms (i.e., if a lexically-specific mechanism is involved in constructing the syntax of idioms, then Wernicke's patients should have more problems with producing idiomatic than non-idiomatic sentences).

To test this prediction 'in silico', we reviewed our lesioning experiments, now evaluating the syntactic performance of lesioned networks separately for sentences that contained idioms and for completely productive sentences.

First, we focused on the model with lesioned syntactic connections, which showed performance symptomatic of Broca's aphasia. Figure 17 shows the percentage of syntactically well-formed idiomatic and non-idiomatic sentences for different lesion ratios. We can see that both idiomatic and non-idiomatic sentences show similar trends of decrease in performance with the degree of impairment, and that, for sufficiently large damage of syntactic connections, production of both idiomatic and non-idiomatic sentences is severely impaired. This is in line with our prediction.

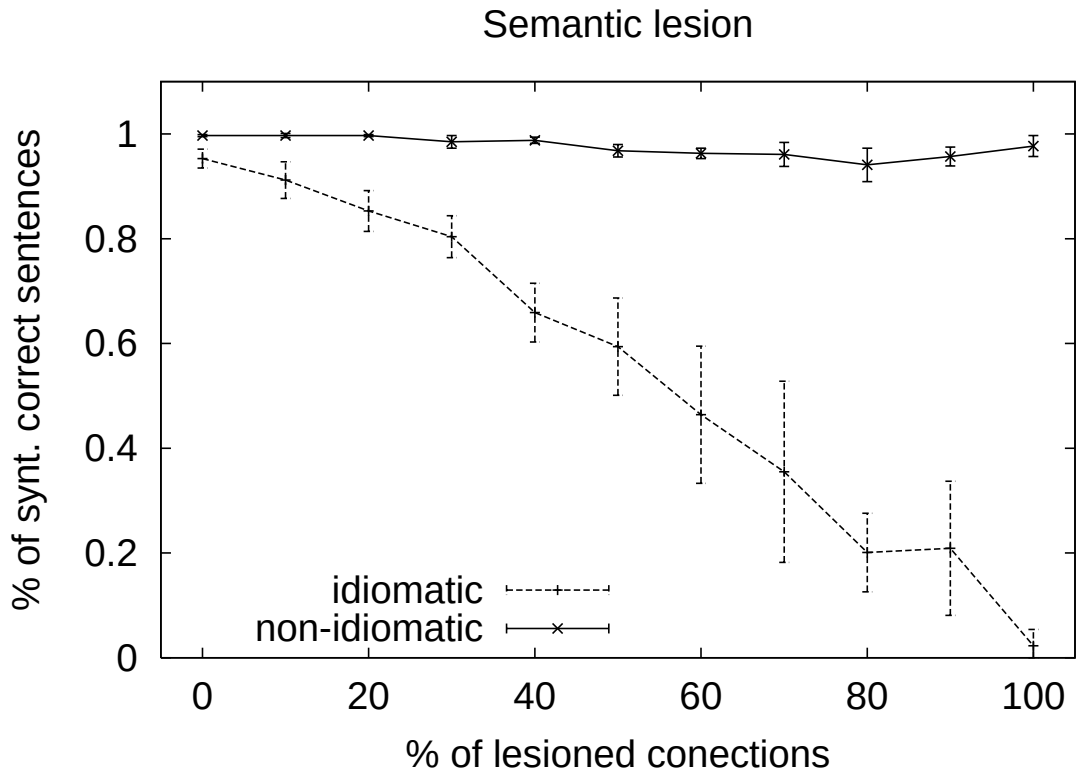

Figure 18. Percentage of syntactically well-formed sentences with and without idioms generated by the network with lesioned semantic connections.

Figure 18 shows the percentage of syntactically well-formed idiomatic and non-idiomatic sentences for different proportion of damaged semantic connections (as in the previous case, we only evaluated syntactic well-formedness, and ignored semantic relevance of produced sentences to the meaning to be expressed). Now the picture is very different: while syntax of non-idiomatic sentences is maintained almost perfect regardless of the lesion degree, syntax of idioms deteriorates proportionally to the ratio of damaged connections. Again, this is in line with our prediction that if we damage semantic connections, syntax of non-idiomatic sentences will be preserved, while production of idioms will be impaired.

To see in detail why idioms are impaired, we analyze what kind of errors were in idiomatic sentences. Almost all of them were 'regularization' errors, in which the sentence followed SUBJ VERB OBJ pattern with 1-word only in each syntactic role, i.e. idioms were reduced to their first word. For instance, *Mummy sees Winnie* was generated instead of *Mummy sees Winnie the Pooh*, and *Daddy gives Helen* was generated instead of *Daddy gives Helen five*. To find the cause of these errors, we compared hidden space activation patterns for the representation of 1-word objects and the first word of multi-word NP phrases in the object role, as well as representation of regular verbs and first words of discontinuous VP idioms, in the normal and lesioned networks. Apart from semantic differences, these activation patterns differed along principal components encoding syntactic properties (roughly

PCA 1-4) in the unlesioned network (Figures 19a and 20a). In the lesioned network, the idiomatic and non-idiomatic cluster shrank, collapsed together and occupied a location that corresponded to non-idioms in the unlesioned network (Figures 19b and 20b). As the semantic differences are missing too, the information about idiomaticity gets lost and the network makes regularization errors.

Discussion. The fact that we lose idioms if we destroy conceptually specific information (mediated through semantic connections in our model) supports the claim that syntax of idioms is bound to their concepts and stored in the lexicon as a whole in the form of superlemmas (Sprenger et al., 2006). However, this does not rule out general sequencing mechanisms; indeed, our SRN uses a single sequencing mechanism based on recurrent connections to the context layer for constructing both idiomatic and non-idiomatic sentences. Levelt & Meyer (2000), as well as Sprenger et al. (2006) claim that after superlemmas and lemmas are selected, they undergo standard grammatical and phonological encoding.

On the other hand, our result that syntax of non-idiomatic sentences was preserved in spite of severity of damage to semantic connections seems contrary to Levelt's claim that syntactic privileges for even single words are stored in a lexically specific way as lemmas (Levelt et al., 1999), and in accord with theories emphasizing more general abstract syntactic mechanisms (Chang et al., 2006; Konopka & Bock, 2009). However, this result could be a too-good artifact, because all our regular sentences were transitive sentences of identical SUBJ VERB OBJ structure, hence semantics of the sentences made no difference.

General discussion

In this article, we tested the constructivist conjecture that language can be acquired as surface patterns. We specially focused on production of idioms, namely the hypothesis that idioms form a continuum with other linguistic forms and can be treated within a general framework of acquisition and production.

Idioms are an ideal testbed for studying the interplay of lexical and syntactic mechanisms in language production. Among students of language production, there has been a lot of disputes about roles of content preparing and structure building mechanisms, some arguing for separate pathways (Chang, 2002), others for integrated processing (e.g. F. Ferreira, 2000; Sprenger et al., 2006). We explored to what extent the content-structure division can implicitly develop within a single connectionist network as a result of learning. Our approach was to gain a deep understanding of the mechanisms and internal organization of a very simple network (before proceeding to a more elaborate model).

Our main findings follow. In Study 1, we validated for idioms a general result from Elman (1990)—namely that a simple recurrent network can acquire language as surface patterns. Our SRN was successfully trained to produce a mixture of syntactically well-formed regular and idiomatic transitive sentences, including variable-length phrases and discontinuous idioms, without any dedicated mechanism specially designed for idioms.

In Study 2, we proved the ability of the SRN enhanced with a semantic input layer to produce syntactically well-formed and semantically correct sentences for given meanings. The network was able to learn to produce both idioms and regular sentences. Detailed analysis of the representational space of the network's hidden layer showed that different

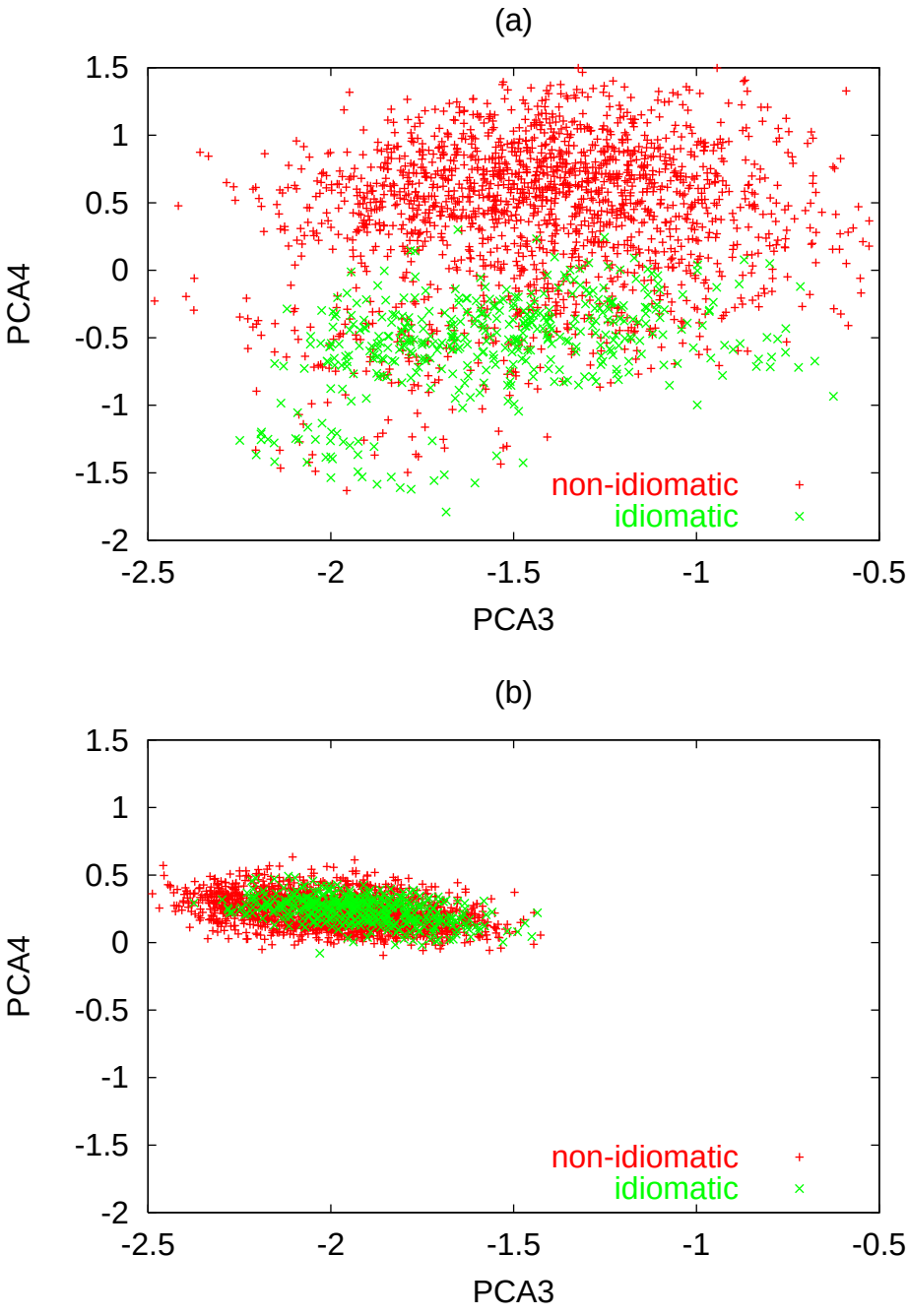

Figure 19. Activity patterns of non-idiomatic and idiomatic verbs in the unlesioned network (a) and the network with 100% of semantic connections lost (b). Clusters shrank in the latter case and occupy a position within or close to the area originally occupied by non-idiomatic verbs in the unlesioned network.

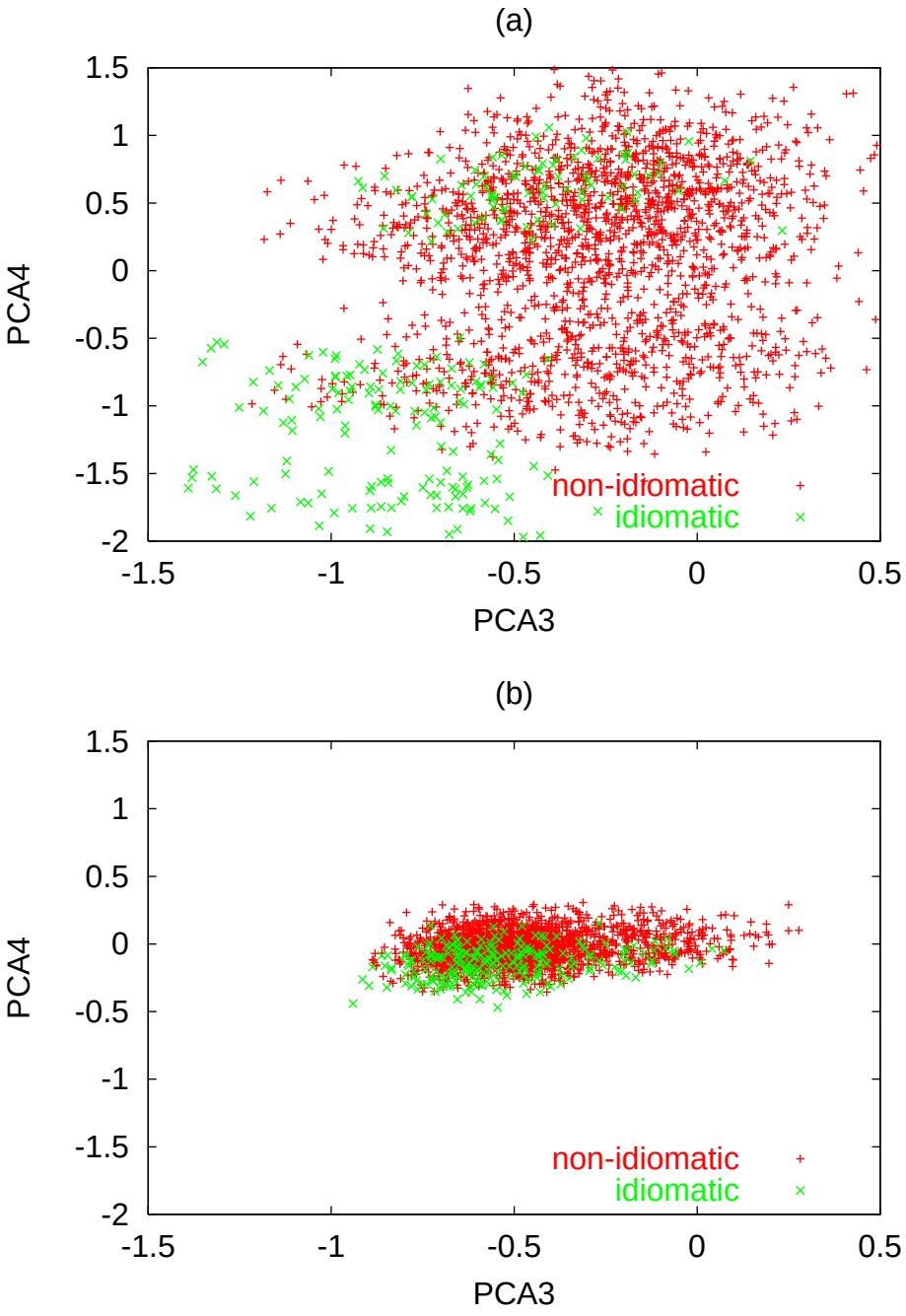

Figure 20. Activity patterns of non-idiomatic and idiomatic objects in the unlesioned network (a) and the network with 100% of semantic connections lost (b). Clusters shrank in the latter case and occupy a position within or close to the area originally occupied by non-idiomatic objects in the unlesioned network.

subspaces encoded different aspects of the sentence generation process: some encoded structural information (regardless of content), others encoded content (regardless of structure). Because a hidden layer activity pattern can be expressed as a linear combination of its components, syntactic and semantic aspects encoded in different subspaces jointly influence the generation process.

Our lesion study showed that by lesioning different part of the network, we could reproduce some symptoms of patients with Broca's aphasia (grammatical errors, when lesioning connections of the basic sequencing SRN) and Wernicke's aphasia (semantic errors, when lesioning connections from the semantic input layer to the hidden layer of the sequencing SRN). Analysis of the hidden layer showed that these lesions indeed had effects in the corresponding subspaces representing content and structure. Hence, an implicit structure-content division can arise as a result of internal space reorganization within a single SRN during learning and the two aspects can be to a large extent damaged selectively without the necessity of having explicit structural (architectural) division in the model.

This situation is not new in connectionist modeling; for example in the domain of word reading, dual-route architectures were proposed to account for reading of regular words and pseudowords[20] as well as irregular words. Interestingly enough, there is analogy between this problem and ours: pronunciation of regular words and of pseudowords must rely on general phonological rules of a language; pronunciation of irregular words (exceptions) needs to be learned and remembered as dictionary-like entries (similar to idioms in our model). Dual-route models (e.g Coltheart & Rastle, 1994; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) solved this problem by suggesting a separate sublexical or phonological (grapheme-to-phoneme) pathway for reading regular words and a lexical pathway (sometimes divided to semantic and non-semantic) for reading irregular words. These models were also successful in simulating different kinds of dyslexias by lesioning respective pathways. On the other hand, models declaring themselves to be instances of Parallel Distributed Processing (PDP) approach challenged the established dual-route models by showing how different patterns of reading and different dyslexias could be accounted for within a single connectionist network with no explicit structural division between lexical and sublexical routes (Seidenberg & McClelland, 1989; Plaut, McClelland, Seidenberg, & Patterson, 1996). Similarly, Gordon & Dell (2002) published a simple feedforward connectionist model accounting for deficits in light and heavy verb production. Although the model was structurally uniform, it could be selectively lesioned in a way that produced syntactic and semantic aphasic symptoms. All these models demonstrate that a 'division of labour' can develop implicitly in a distributed representational space as a result of learning.

Our lesion study also contributes to the discussion between those who emphasize lexically-specific processes in generation of idiom syntax (e.g. Sprenger et al., 2006) and those who argue for abstract structure generating mechanisms (Konopka & Bock, 2009). When we lesioned connections related to general structure building (as shown by analysis of the unlesioned network), the syntax in produced idioms and regular phrases was impaired to the same degree. When we lesioned connections related to providing content, the syntax of regular sentences was almost intact, while idioms were erroneously regularized. This result is in line with the Superlemma approach of Sprenger et al. (2006) in that, in our model:

[20] Pseudowords are strings of characters that obey phonological rules of a language, but are not words of that language.

(1) idioms are produced by the same general sequencing mechanism that works for regular sentences; (2) the production of idioms is modulated by content-specific mechanisms. This generates interesting predictions that could be tested empirically: finding a Broca's aphasics with agrammatic symptoms who could produce idioms correctly would be a challenge to our claim 1; finding a Wernicke's aphasics with lexical/anomic deficits who could produce idioms correctly would be a challenge to our claim 2.

Limitations

Our model was not intended as either a complete or a detailed account of language production and is oversimplistic in many ways. First of all, use of localistic representation on semantic input layer implies that all concepts are equally distant/similar to each other; therefore it prevents from accounting for semantic effects, such as semantic paraphasias[21]. Also, since we abstracted away from a model of phonological encoding and articulation, we cannot account for phonological errors typical in some aphasics. We also deliberately abstracted from morphology issue: our language only contained verbs in one morphological form (3rd person, singular, present tense) that could easily be acquired as a surface pattern. Also, because our language almost exclusively contained content words (e.g. determiners only occurred as part of idiomatic phrases), we could not replicate differences in production of function versus content words typical in many Broca's aphasics. We also did not intend to account for any kind of priming (structural or semantic) typically tested in psycholinguistic experiments with idiom production. Needless to say, we abstracted from related issues of language comprehension, reading and writing.

However the most serious critique applicable to our model is that of Chang (2002), whose Prod-SRN architecture is very similar to ours – it is a SRN enhanced with a semantic input that uses binding-by-space for 4 roles AGENT, PATIENT, ACTION, GOAL, with localist coding of concepts in the roles. Chang criticizes the model for poor generalization ability and reports that his Prod-SRN achieved just slightly above 10% of correct sentences on the test set even after 4000 epochs of training. Our prod-SRN-type network apparently has better generalization performance; it achieves very high accuracy in the sentence production task, despite the fact that the set of meanings used for testing and sentence generation had zero overlap with those in the training set. In fact, our network's apparently higher performance was probably due in part to the more complex grammatical constructions used in Chang's training/testing sentences, and in part due to different ratio between train set size and the total size of the language (500 : 2200 in our case, vs. 501 : 75330 in his case).

However, Chang's argument is not just about the empirical performance of a prod-SRN-type architecture. His most important argument is from an analysis of the 'binding-by-space' scheme which the network uses to represent propositions. If different semantic roles are represented using separate banks of semantic units, each role is connected to the hidden layer with its own independent set of connections; hence, training experience with e.g. DOG in AGENT semantic role has no influence at all on DOG in PATIENT semantic role. Therefore, the Prod-SRN is not able to generalize a word to a role where it has never been observed before. We accept this objection. Our network was able to generate new

[21] Semantic paraphasias are substitutions of words by semantically related words.

combinations, e.g. generate a correct sentence for AG:MUMMY, ACT:EAT, PAT:BREAD after being taught to express AG:MUMMY, ACT:EAT, PAT:APPLE and AG:DADDY, ACT:SEE, PAT:BREAD; however, it could not generate TEDDYBEAR as agent, because it has never appeared in that role. Currently, we are preparing a new model, where semantics is not coded by space, but in a temporally structured sequence of activations of a single bank of units. This way, the same connections are trained during the experiences with same concept in AGENT and PATIENT roles. In fact, the network presented in the current paper is a component of this new model, so the analysis in the current paper is an analysis of some components of the more complex system we envisage.

In the end, we want to put our model in a developmental perspective. According to constructivist linguists (e.g. Goldberg, 1995; Tomasello, 2003) syntax is learned from surface patterns in the form of 'item-based' constructions and later becomes abstract through accumulated experience. We ultimately intend the model presented in the current paper to be a model of the item-based stage in language development. The additional mechanisms which are required to model mature sentence generation are the focus for our current research.

References

Alexander, M. P. (2005). Aphasia I: Clinical and anatomical issues. In *Patient-based approaches to cognitive neuroscience* (pp. 181–198). Cambridge, MA: MIT Press.

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

Bock, K. (1995). Sentence production: From mind to mouth. In J. L. Miller & P. D. Eimas (Eds.), *Handbook of perception and cognition* (Vols. 11: Speech, language and communication, pp. 181–216). Orlando, FL: Academic Press.

Cernansky, M. (2007). *Comparison of recurrent neural networks with markov models on complex symbolic sequences*. Unpublished doctoral dissertation, Slovak Technical University, Bratislava, Slovakia.

Cernansky, M., Makula, M., & Benuskova, L. (2007). Organization of the state space of a simple recurrent neural network before and after training on recursive linguistic structures. *Neural Networks*, *20*(2), 236–244.

Chambers, L. (1995). *Practical handbook of genetic algorithms* (Vol. 2). Boca Raton, FL: CRC-Press.

Chang, F. (2002). Symbolically speaking: a connectionist model of sentence production. *Cognitive Science*, *26*(5), 609–651.

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

Christiansen, M. H., & Chater, N. (1999). Toward a connectionist model of recursion in human linguistic performance. *Cognitive Science*, *23*(2), 157-205.

Christiansen, M. H., & Chater, N. (Eds.). (2001). *Connectionist psycholinguistics*. Westport, CT: Ablex.

Coltheart, M., & Rastle, K. (1994). Serial processing in reading aloud: Evidence for dual-route models of reading. *Journal of Experimental Psychology: Human Perception, and Performance*, *20*(6), 1197–1211.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, *108*(1), 204-256.

Cowie, A. (Ed.). (1998). *Phraseology: Theory, analysis and applications*. Oxford, UK: Clarendon Press.

Cutting, J. C., & Bock, K. (1997). That's the way the cookie bounces: Syntactic and semantic components of experimentally elicited idiom blends. *Memory and Cognition*, *25*, 57–71.

Cutting, J. C., & Ferreira, V. S. (1999). Semantic and phonological information flow in the production lexicon. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(2), 318–44.

Damasio, H., Grabowski, T., Tranel, D., Hichwa, R., & Damasio, A. (1996). A neural basis for lexical retrieval. *Nature*, *380*, 499–505.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, *93*(3), 283–321.

Dominey, P., Hoen, M., & Inui, T. (2006). A neurolinguistic model of grammatical construction processing. *Journal of Cognitive Neuroscience*, *18*(12), 2088–2107.

Elman, J. L. (1990). Finding structure in time. *Cognitive Science*, *14*(2), 179–211.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195-224.

Elman, J. L., Bates, E. A., Johnson, M. H., Smith, A. K., Parisi, D., & Plunkett, K. (1996). *Rethinking innateness: A connectionist perspective on development.* Cambridge, MA: MIT Press.

Fenson, L., Dale, P., Reznick, J. S., Thal, D., Bates, E., Hartung, J., et al. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, *59*(5), i–185.

Ferreira, F. (2000). Syntax in language production: An approach using tree-adjoining grammars. In L. Wheeldon (Ed.), *Aspects of language production* (pp. 291–330). Hove, UK: Psychology Press.

Ferreira, V. S., & Slevc, L. R. (2009). Grammatical encoding. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 453–469). Oxford, UK: Oxford University Press.

Goldberg, A. E. (1995). *Constructions: A construction grammar approach to argument structure.* Chicago: University of Chicago Press.

Gordon, J., & Dell, G. (2002). Learning to divide the labor between syntax and semantics: a connectionist account of deficits in light and heavy verb production. *Brain and Cognition*, *48*(2-3), 376–81.

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*, 301–307.

Haykin, S. (2008). *Neural networks and learning machines* (3rd ed.). Upper Saddle River, NJ: Prentice Hall.

Henson, R., Burgess, N., & Frith, C. (2000). Recoding, storage, rehearsal and grouping in verbal short-term memory: an fMRI study. *Neuropsychologia*, *38*(4), 426–440.

Hickok, G., Erhard, P., Kassubek, J., Helms-Tillery, A., Naeve, S., Velguth, Struppf, J., et al. (2000). A functional magnetic resonance imaging study of the role of left posterior superior temporal gyrus in speech production: implications for the explanation of conduction aphasia. *Neuroscience Letters*, *287*, 156–160.

Hickok, G., & Poeppel, D. (2007). The cortical organization of speech processing. *Nature Reviews Neuroscience*, *8*(5), 393–402.

Indefrey, P. (2009). Brain-imaging studies of language production. In G. Gaskell (Ed.), *Oxford handbook of psycholinguistics* (pp. 547–564). Oxford, UK: Oxford University Press.

Indefrey, P., Brown, C. M., Hellwig, F., Amunts, K., Herzog, H., Seitz, R. J., et al. (2001). A neural correlate of syntactic encoding during speech production. *Proceedings of the National Academy of Sciences*, *98*(10), 5933–5936.

Jackendoff, R. (1995). The boundaries of the lexicon. In M. Everaert, E. van der Linden, A. Schenk, & R. Schreuder (Eds.), *Idioms, structural and psychological perspectives* (p. 133-166). Hillsdale, NJ: Lawrence Erlbaum.

Jackendoff, R. (1997). *The architecture of the language faculty.* Cambridge, MA: MIT Press.

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution.* Oxford, UK: Oxford University Press.

Knott, A. (2010). A sensorimotor characterisation of syntax, and its implications for models of language evolution. In *Proceedings of the 8th international conference on the evolution of language (EVOLANG8).* (to appear)

Kolk, H. (2006). How language adapts to the brain: An analysis of agrammatic aphasia. In L. Progovac, K. Paesani, E. Casielles, & E. Barton (Eds.), *The syntax of nonsententials: Multi-disciplinary perspectives* (pp. 229–58). London, UK: John Benjamins.

Konopka, A., & Bock, K. (2009). Lexical or syntactic control of sentence formulation? Structural generalizations from idiom production. *Cognitive Psychology*, *58*(1), 68–101.

Levelt, W. J. M., & Meyer, A. S. (2000). Word for word: Multiple lexical access in speech production. *European Journal of Cognitive Psychology*, *12*(4), 433–452.

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

Lu, L., Crosson, B., Nadeau, S., Heilman, K., Gonzalez-Rothi, L., Raymer, A., et al. (2002). Category-specific naming deficits for objects and actions: semantic attribute and grammatical role hypotheses. *Neuropsychologia*, *40*, 1608–1621.

McClelland, J. L., & Kawamoto, A. H. (1986). Mechanisms of sentence processing: assigning roles to constituents. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition* (Vol. 2: Psychological and biological models, pp. 272–325). Cambridge, MA: MIT Press.

Paulesu, E., Frith, C., & Frackowiak, R. (1993). The neural correlates of the verbal component of working memory. *Nature*, *362*, 342–345.

Perani, D., Cappa, S., Schnur, T., Tettamanti, M., Collina, S., Rosa, M., et al. (1999). The neural correlates of verb and noun processing - a pet study. *Brain*, *122*, 2337–2344.

Pine, J. M., & Lieven, E. V. M. (1997). Slot and frame patterns in the development of the determiner category. *Applied Psycholinguistics*, *18*(123–138).

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56–115.

Pulvermüller, F., Shtyrov, Y., & Ilmoniemi, R. (2005). Brain signatures of meaning access in action word recognition. *Journal of Cognitive Neuroscience*, *17*(6), 884–892.

Rumelhart, D. E., Hinton, G. E., & Williams, R. J. (1986). Learning internal representations by error propagation. In D. E. Rumelhart & J. L. McClelland (Eds.), *Parallel distributed processing: explorations in the microstructure of cognition* (Vol. 1: Foundations, pp. 318–362). Cambridge, MA, USA: MIT Press.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568.

Sprenger, S. A. (2003). *Fixed expressions and the production of idioms.* Unpublished doctoral dissertation, University of Nijmegen, Netherlands.

Sprenger, S. A., Levelt, W. J. M., & Kempen, G. (2006). Lexical access during the production of idiomatic phrases. *Journal of Memory and Language*, *54*, 161–184.

Tanner, D. C. (2008). *The family guide to surviving stroke and communication disorders* (Second ed.). Sudbury, MA: Jones and Bartlett.

Tino, P., Cernansky, M., & Benuskova, L. (2004). Markovian architectural bias of recurrent neural networks. *IEEE Transactions on Neural Networks*, *15*(1), 6–15.

Tomasello, M. (2003). *Constructing a language: A usage-based theory of language acquisition.* Cambridge, MA: Harvard University Press.

Tranel, D., Adolphs, R., Damasio, H., & Damasio, A. (2001). A neural basis for the retrieval of words for actions. *Cognitive Neuropsychology*, *18*(7), 655-674.

Ungerleider, L., Courtney, S., & Haxby, J. (1998). A neural system for human visual working memory. *Proceedings of the National Academy of Sciences*, *95*, 883–890.

Wagner, D., Maril, A., Bjork, R., & Schachter, D. (2001). Prefrontal contributions to executive

control: fmri evidence for functional distinctions within lateral prefrontal cortex. *NeuroImage*, *14*, 1337–1347.

Appendix
Normalized negative log-likelihood

Normalized negative log-likelihood (NNL) is calculated over the sequence of symbols (words) from time step $t = 1$ to T as

$$NNL = -\frac{1}{T}\sum_{t=1}^{T} \log_N p_i(t), \tag{1}$$

where the base of the logarithm N is the number of possible symbols (size of the dictionary), and the $p_i(t)$ is the probability of predicting the i-th symbol in the time step t. Value $p_i(t)$ was calculated using soft-max combining function of output neurons activations o_j as:

$$p_i(t) = \frac{\exp(o_i(t))}{\sum_{j=1}^{N} \exp(o_j(t))}. \tag{2}$$