

File ID 344016  
Filename 8: Learning grammar through episodic memory consolidation

---

SOURCE (OR PART OF THE FOLLOWING SOURCE):

Type Dissertation  
Title The neural basis of structure in language: bridging the gap between symbolic and connectionist models of language processing  
Author G. Borensztajn  
Faculty Faculty of Humanities  
Faculty of Science  
Year 2011  
Pages xv, 233  
ISBN 90-5776-233-1

FULL BIBLIOGRAPHIC DETAILS:

<http://dare.uva.nl/record/400259>

---

*Copyright*

*It is not permitted to download or to forward/distribute the text or part of it without the consent of the author(s) and/or copyright holder(s), other than for strictly personal, individual use.*

---

## Chapter 8

---

# Learning grammar through episodic memory consolidation

*This chapter completes the integration of the episodic grammar with the HPN framework in the episodic-HPN model. The enrichment of HPN with an episodic memory allows conditioning on sentence history, and at the same time learning from analogy among stored exemplars. The approach is motivated from the usage-based language acquisition literature, which shows that a similar process of analogy extraction drives the discovery of general, productive rules from specific utterances.*

*The episodic-HPN model assumes a bi-directional interaction between episodic and semantic memory in syntactic bootstrapping: on the one hand the topology (semantic memory) is used to compute distances between treelets where no episodic memory traces are present (hence, it can deal productively with unseen sentences), while on the other hand the episodic memory is used to find a ‘shortest derivation’ parse, which results in a gradual adjustment of the topology to episodic experiences. As such, learning a grammar in episodic-HPN parallels the process of memory consolidation and de-contextualization in the brain, whereby an abstract semantic memory is gradually extracted from concrete episodic memories. After presenting the formal model, I discuss how the episodic-HPN framework differs from existing connectionist models of memory consolidation, and I point out that it predicts a role for the hippocampus in dynamic binding.*

## 8.1 Introduction

This chapter introduces the episodic-HPN framework, which integrates an episodic memory with the HPN network. It is based on the *episodic grammar* formalism, which was presented in the previous chapters. By contrast to the episodic grammar, which is a supervised and symbolic system, episodic-HPN learns fully unsupervised and is connectionist. But unlike the ‘semantic’ HPN model of Chapter 5 it addresses the problem of conditioning on sentence context by using episodic memory.

In this chapter the two-way interaction between the episodic and semantic memory systems will be modeled. While the episodic grammar modeled language processing (parsing) as the retrieval from semantic memory of episodic memories (reconstructed from traces), episodic-HPN additionally models grammar acquisition as a process of memory consolidation from episodic to semantic memory. As discussed in Chapter 2 memory consolidation refers to the gradual construction of a relational semantic network out of episodic memories. This involves a search for structural analogies between episodes in order to infer shared semantic features, resulting in de-contextualization of the episodes.

Grammar acquisition, like memory consolidation, concerns the question of how abstract knowledge (e.g., the rules of a grammar) is extracted from experience (e.g., sentences). The central claim of this chapter is that grammar acquisition parallels the process of memory consolidation, i.e., it must be construed as a process of analogical inference from episodic to semantic memory. To motivate this claim I will review two well-known studies from the language acquisition literature, and I will formulate conditions for successful grammar induction based on an analysis of the BMM algorithm (section 3.2.2). Then I will present the formal model of episodic-HPN, which shows in technical detail how the consolidation process can be implemented in a connectionist network. Unfortunately, at this stage I have not yet been able to evaluate it quantitatively. However, I do assess qualitatively some predictions of the model relating it to the neuro-biology of memory, and to other models of memory consolidation.

## 8.2 The case for treating language acquisition as a memory consolidation problem

In this section I will review two examples from the literature of grammar acquisition that can be seen as instances of the claim that grammar acquisition, like memory consolidation, involves an analogical inference process based on episodic memories.

### 8.2.1 Example 1: Lieven et al. [2003]

The first example is a dense corpus study in the usage-based tradition by Lieven et al. [2003], aimed at tracing back the sources of creativity of children's speech. In this study Lieven et al. [2003] showed that most utterances produced in one day by a two year old child could be reduced to utterances produced in the previous 6 weeks by using only a single combinatorial operation. For each target utterance they searched the closest matching utterance produced by the child in the preceding weeks, and analyzed the ways in which the novel utterance differed from it. In particular, the number of operations needed to arrive from the closest match to the target utterance was determined. (Operations were 'substitute', 'add', 'drop', 'insert' and 'rearrange'.) It was found that of the target utterances that had not been said before in their entirety (37 % of the total) 74 % could be composed from previous utterances with a single combinatorial operation.

This result suggest that children vary their speech based on analogy with previous utterances. Analogical learning allows bootstrapping general rules, offering children a gradual path to an abstract grammar.

### 8.2.2 Example 2: Marcus et al. [1999] and Marcus [2001]

As a second case study consider the problem of generalization of sentences of the form

(8.1) *A rose is a rose.*

(8.2) *A lily is a lily.*

(8.3) *A tulip is a tulip.*

These examples were used by Marcus [2001] to train a Simple Recurrent Network (SRN). According to Marcus [2001, p. 50] the SRN could not generalize this to

(8.4) *A blicket is a ...*

Human infants, on the other hand, are able to make generalizations of this kind. In a much cited experiment Marcus et al. [1999] showed that 7-month-old infants can learn an artificial grammar of the form ABB, ABA or AAB, and generalize these simple patterns to patterns consisting of words they had not heard during the training session. According to Marcus, the reason why infants can and the SRN cannot generalize these examples correctly is that infants apparently possess a learning system that allows them to extract algebra-like rules that represent relationships between variables, such as identity, whereas the associative learning mechanism of the SRN is only sensitive to transitional frequencies.

While Marcus is correct that *distributed* networks cannot learn relations over variables, the work on HPN shows that there exist cognitively plausible connectionist learning algorithms that *can* simultaneously exploit rule-based structure and distributional information (see sections 4.7.3 and 4.8.1). The mere ability to represent relationships between variables (or invariants, as in HPN) however still does not explain how a system (or a child) actually discovers the ‘correct’ variables (as demonstrated in section 5.5.2, this is not a trivial issue for any learning algorithm).

I propose a different interpretation of the fact that, in contrast to the SRN, humans find it easy to generalize the above examples. According to this interpretation the SRN can only generalize based on *similarity* of the examples, whereas humans generalize based on *analogical* reasoning. Thus it seems that syntax learning exploits a basic cognitive capacity for discovering analogies between the internal representations of stored examples. Finding analogy is a higher order process than finding similarity, because analogy concerns similarity of relations: the discovery procedure requires performing pairwise comparisons between stored internal representations. To do so, a system has to keep track of the representations of all previously processed sentences, which in the SRN, or any other connectionist system without an episodic memory, are lost.

### 8.2.3 The use of analogy in computational models of grammar induction

To clarify the relation between grammar learning and discovery of analogies further, let us closely examine how a typical grammar induction algorithm, such as Bayesian Model Merging (BMM) [Stolcke and Omohundro, 1994, Stolcke, 1994], would succeed at learning the abstract relation underlying the above examples (for a short discussion of BMM see section 3.2.2). If the BMM algorithm is trained on the following sentences (from Stolcke [1994, p. 83])

$a a$	(10X)
$a a b b$	(5X)
$a a a b b b$	(2X)
$a a a a b b b b$	(1X)

then it will initially create a unique rule for every sentence, and unique nonterminal symbols for every occurrence of  $a$  and  $b$ , as in

$$\begin{aligned} S &\rightarrow A_1 B_1 \\ &\rightarrow A_2 A_3 B_2 B_3 \\ &\rightarrow A_4 A_5 A_6 B_4 B_5 B_6 \\ &\rightarrow A_7 A_8 A_9 A_{10} B_7 B_8 B_9 B_{10} \end{aligned}$$

Subsequently, it will perform a hill-climbing search for an optimal grammar, using *merge* and *chunk* operations to move in the space of possible grammars, while

after every operation it checks if it improves a certain objective function. If this objective function incorporates a preference for smaller grammars, as is the case for the Minimum Description Length (MDL), then it will reward a combination of merges and chunks if that uncovers an *analogy* that is hidden in the data. For instance, after merging all preterminals that rewrite to the same terminal, a subsequent chunk of  $(A A B B)$  into the single nonterminal  $X$  results in a reduction of the size of the grammar

$$\begin{aligned} S &\rightarrow A B \\ &\rightarrow X \\ &\rightarrow A X B \\ &\rightarrow A A X B B \\ A &\rightarrow a \\ B &\rightarrow b \end{aligned}$$

The fact that MDL is able to discover recurring rewrite rules is thus due to the existence of analogical sentence pairs in the data. Finally,  $X$  can be merged with  $S$ , such that eventually one obtains the recursive grammar

$$\begin{aligned} S &\rightarrow A B \\ S &\rightarrow A S B \\ A &\rightarrow a \\ B &\rightarrow b \end{aligned} \tag{8.5}$$

It is important to note that thereby at every step BMM evaluates the objective function *globally*, on the entire corpus. From this analysis one may conclude that two things are necessary for a learning algorithm to find rules by analogy

1. *Simplicity bias.* A learning algorithm will not find the rules of the grammar if it is not somehow forced to find a more compact description of the data. This is possible if there are analogies hidden in the structure of the data; making these explicit as grammar rules reduces the number of rules needed to describe the data, hence the description length.
2. *Episodic memory.* A learning algorithm can only find analogies, or merges, if it tries out comparisons between all pairs of sentences. Therefore, for learning all previously processed data has to be available to the system.

I propose that both conditions must be fulfilled if one wants to build a successful connectionist model of grammar acquisition.

#### 8.2.4 Towards a connectionist model of memory consolidation in language

In essence, Marcus' results, showing that infants can generalize patterns to unseen words, and his subsequent demonstration that the SRN fails at the same task

can be reanalyzed and summarized as follows: language (or rather, grammar) acquisition should best be construed as a process of memory consolidation from an episodic to a semantic memory (and not as mere statistical learning of transition frequencies). The possession of an episodic memory is a necessary condition for a learning strategy that is based on the discovery of analogies, which seems to be the underlying strategy in language learning. The study of Lieven et al. [2003], as well as an analysis of the inner workings of the BMM algorithm point to the same insight.

From this it can be concluded that *connectionist networks that do not have a built-in episodic memory cannot learn a (phrase structure) grammar*, because they lack an ability for analogical inference. This applies specifically to recurrent networks such as the SRN, but also to ‘semantic’ HPN. Since these networks do not keep analyses of processed sentences, the induction of recursive, context free grammars from examples is theoretically impossible.

### 8.2.5 Discovering analogies via the principle of the shortest derivation

The analysis of the BMM algorithm in section 8.2.3 suggested that, given an episodic memory, a simplicity bias is still needed to drive learning toward an ‘optimal’ grammar. How can a preference for a smaller grammar be implemented in the brain (or in a connectionist network)?

A cognitively plausible solution, according to many, is to assume that the brain implements a principle of least cognitive effort, by using the *shortest possible derivation* of a sentence. The *principle of the shortest derivation* has been introduced in Data Oriented Parsing as a way to parse novel sentences in terms of fragments of earlier processed sentences, and as an alternative to probabilistic parsing [e.g., Bod, 2000]. It has also been used for unsupervised grammar induction with U-DOP [Smets, 2010].

By reusing existing fragments as much as possible the grammar is kept at minimal size; hence, this indirectly implements a simplicity bias. A learner that uses the shortest derivation will try to discover and reuse shared structure from examples. For instance, given the sentences from section 8.2.2 (*A rose is a rose*, etc.), it will prefer to reuse a rule such as ‘*X is X*’ rather than idiosyncratic rules in the derivation of new sentences. The study of Lieven et al. [2003] seems to indicate that children use a similar strategy (i.e., a minimal number of edit operations) to produce new sentences.

Parsing with the shortest derivation provides the brain with a tool for discovering analogies from a structurally organized episodic memory space. In general terms it involves a cognitive ability to analyze a new experience in terms of a minimal number of previously analyzed experiences. Presumably the ground work of the memory consolidation process, i.e., finding structure in the daily stream of

episodic experiences, can be traced back to a search for the shortest derivation also in non-linguistic domains. As shown in section 7.3.2, such a search procedure can be executed locally, conforming to the connectionist constraint, provided episodes are encoded as distributed traces in the network units (as proposed by the episodic grammar framework).

In the next section a similar local procedure for finding the shortest derivation (and with it, analogies) will be implemented in a connectionist version of the episodic grammar, episodic-HPN. It is expected that by virtue of a parsing strategy that prefers the shortest derivation an optimal (minimal) grammar can be bootstrapped from plain text, where the ‘semantic’ version of HPN (without episodic memory) failed (e.g., see the experiment in section 5.5.2).

### 8.3 The episodic-HPN model

The episodic-HPN grammar is based on the episodic left corner shifting grammar (e-LCSG; see section 7.3). The primitive units of the grammar are treelets containing episodic traces. However, instead of reading off treelets from the CFG rules of a treebank, in episodic-HPN the treelets are based on the compressor nodes and input nodes of the HPN grammar; hence they have no labels. The parser is built on top of the episodic shortest derivation left corner chart parser that was developed in Chapter 7. The shortest derivation parse is selected for the reasons discussed in the previous section. Learning is integrated with parsing: fast, instant learning occurs as episodic traces are added to the treelets involved in a derivation after successfully parsing a sentence. In addition, slow, statistical learning follows the algorithm for updating the representations of units across bindings, adopted from HPN (section 5.4). I will first discuss the initialization of the episodic-HPN grammar, then the parser and then the learning algorithm.

#### Initialization

When episodic-HPN is initialized, episodic treelets are created for every compressor node in every possible register position.

**Definition 7** (HPN compressor node treelet). An HPN compressor node treelet is a triple  $\mathcal{T} = \{X, n, E\}$ , where  $X$  denotes a unique compressor node,  $n$  denotes the ordinal number of the *active* slot in  $X$  (i.e., the register position), and  $E$  is a set of traces from sentences that have visited the treelet, initially empty.

A distinct *shift treelet* is created for every combination of an HPN input node (corresponding to a word) and shift slot of a compressor node (see the section ‘Language model’ below).

**Definition 8** (HPN shift treelet). An HPN shift treelet is a 4-tuple  $\mathcal{T} = \{X, n, W, E\}$ , where  $X$  denotes a unique compressor node,  $n$  denotes a *shift*

slot of  $X$ ,  $W$  denotes an input node in HPN corresponding to word  $w$ , and  $E$  is a set of traces from sentences that have visited the treelet, initially empty.

## Parsing

Like in the symbolic case, episodic HPN is implemented as a shortest derivation episodic left corner chart parser. The construction of the chart (involving *shift*, *project* and *attach* operators) is described in section 7.1.3; when a new treelet state is added to the chart all the traces are copied from the ‘treelet type’ to the treelet state, and receive a certain activation (i.e., a value for the Shortest Derivation Length (SDL)). HPN treelet states are defined in analogy to symbolic treelet states (see section 7.3).

**Definition 9** (HPN treelet state). An HPN treelet state  $q$ , associated with a treelet  $\mathcal{T}$ , is a 4-tuple  $q = \{\mathcal{T}, i, j, E_q\}$ , where  $\mathcal{T}$  is either a compressor node treelet or a shift treelet;  $j$  is the left span index,  $i$  is the right span index, and  $E_q$  is a set of *activated* traces. If  $\mathcal{T}$  is a shift treelet then  $i = j + 1$ .

The shortest derivation parser uses two levels of tie-breaking in case of equal derivation length. The first (optional) level of tie-breaking is the probabilistic left corner model, as estimated from the relative frequencies of the sentences processed up to the current point. Note that while for the symbolic, supervised parser relative frequencies are estimated only once, from an annotated treebank, in episodic-HPN the frequency counts are updated after every parse and probabilities have to be re-normalized. This level computes project, attach and shift probabilities conditioned on a left corner and goal category, which in case of HPN are identified with a root vector and goal slot. Initially these probabilities will be zero for most events.

The second level of tie-breaking consists of back-off probabilities computed from the HPN metric (i.e., distances between root and slot vectors in substitution space), which are conditioned on the left corner (i.e., the root of a compressor node or a word unit) alone. The back-off probabilities must again be recalibrated (renormalized) after every parse, because the root and slot vectors of nodes that are involved in the parse may have changed as a result of learning.

Although it may seem as though the HPN metric only contributes in a minor way to the parse decision, as it is only used for tie-breaking (i.e., for computing back-off probabilities), in fact in the early stages of learning the role of the metric will be dominant, because there are still very few traces (exemplars) to derive the shortest derivation with, and most first level probabilities will be zero. The metric, however, yields non-zero probabilities for all events right from the start because node representations are initialized with random values.

## Learning

Learning in unsupervised episodic HPN occurs after every parse, and involves the following steps:

1. The shortest derivation of the sentence is found, or in case of ties the most probable shortest derivation.<sup>1</sup>
2. Fast, one-shot learning: Traces for the current sentence are stored in the treelets along the derivational path of the winning parse (episodic memory consolidation).
3. Slow, semantic learning: Vector representations are updated for compressor nodes and word nodes that participated in the winning parse, as described in section 5.4; back-off probabilities are renormalized from the updated metric.
4. (optionally) Frequency counts for project, attach and shift transitions are updated, and project, shift and attach probabilities are renormalized.

## Language model

In order to cope with one of the limitations of HPN identified in section 5.6.3, namely its inability to compute string probabilities, HPN was modified in a way that enables it to represent and learn shift transitions to a word. To this end, compressor nodes are equipped with an extra set of *shift slots*, apart from the regular slots, which are invoked at every shift operation. (Thus, there is one shift slot between every two regular slots.) Input nodes in episodic-HPN also have a slot, unlike in the previous version of HPN, where the shift slots of the compressor nodes bind to. Thus, in episodic-HPN the input nodes are not actually terminal nodes anymore. During learning, the vector representations of shift slots and input node slots that were involved in the shift bindings of the most probable parse are updated, just like the regular roots and slots in the earlier version of HPN. This means that a complementary, independent substitution space is needed: the *shift space*. Its dimensionality equals the number of words in the lexicon.

## Implementation

I have fully implemented the episodic-HPN model to confirm that the described components together constitute a complete model. *Quantitative* evaluation and optimizing the parameters of the model for its use in typical computational linguistics tasks is left for future work. However, we can already evaluate *qualitatively* the predictions the model makes for cognitive neuroscience.

---

<sup>1</sup>As before, it should be taken care of that previously processed sentences that are identical to the currently processed sentence are excluded from participating in the shortest derivation.

## 8.4 Predictions of episodic-HPN for memory consolidation

### 8.4.1 What episodic-HPN says about the transformation (de-contextualization) from episodic to semantic memory

An important contribution of the episodic-HPN framework is that it focuses attention on the parallels between language acquisition and memory consolidation, as it implements a model of grammar induction that is casted in terms of a transfer of linguistic knowledge from specific episodic representations to abstract semantic representations.

To recapitulate, in episodic-HPN a derivation of a processed sentence is instantly stored in the form of episodic memory traces distributed over semantic network units (one-shot learning). These traces are then recruited to compute the shortest derivation in subsequent processing of novel sentences. Thereby they may force a preference for certain parses that are compatible with previous experiences (exemplar-based processing). As a consequence the bindings of network units that participate in the preferred parse are strengthened, resulting in adjustment of their representations, which in turn affects the topological organization of the network. The topology is important for dealing with unseen events in a productive way.

As this demonstrates, the formation of a topology of syntactic categories (the so-called substitution space) is strongly influenced by the episodic shortest derivations, because the latter are a major factor in determining the selected parse for which the root-slot bindings are updated. In terms of memory consolidation this interaction represents a de-contextualization process: it shows how a (context free) semantic memory (the topology) is gradually shaped from contextually bound episodic experiences, until eventually it comes to reflect an individual's personal (linguistic) experience in the form of an abstract grammar.

The proposed approach to memory consolidation, and its neural interpretation depart drastically from previous computational models of memory consolidation [e.g., McClelland et al., 1995, O'Reilly and Rudy, 2001]. In section 8.5.1 I will discuss these models, and contrast them with the current proposal.

### 8.4.2 The role of the hippocampus according to episodic-HPN

As discussed in section 2.7.1, in the neuroscience literature a special role is reserved for the hippocampus in memory consolidation. There are two aspects of hippocampal function that are intimately related [Eichenbaum, 2004]:

- First, the hippocampus is involved in episodic memory encoding and consolidation. For encoding it binds sequences of discontinuous semantic elements into episodes, which are structurally organized in so-called ‘relational networks’; consolidation involves replay of episodic sequences, thereby strengthening semantic relations.
- Secondly, the hippocampus is involved in processing novel configurations by flexible association of semantic elements, that are shared among episodes through the relational networks.

### Dynamic encoding of episodic memories in episodic-HPN

Let us first consider how episodic memory replay is accounted for in episodic-HPN. In the episodic grammar framework it was assumed that episodes can be reconstructed from sequences of traces that are encoded with follow-up numbers. However, the question of how successive traces within a derivation are localized in the cortex was for that moment ignored.

A similar problem, concerning how the brain can efficiently recover the address of a unit where a ‘tag’ is stored in a pending derivation, was addressed in HPN. There the solution involved a switchboard construction that implements an addressor system, as part of the *dynamic binding* approach (see section 4.8.2). The same solution can be adopted for episodic memory replay from traces.

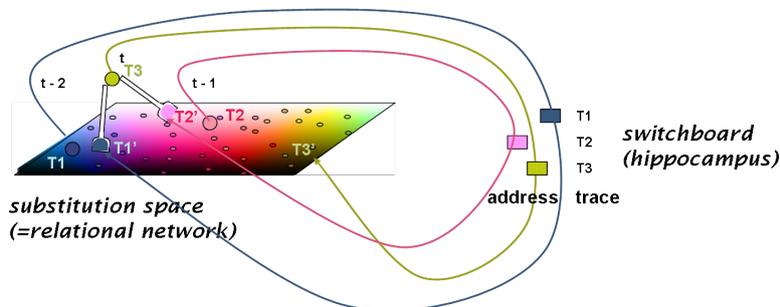


Figure 8.1: Replay of an episodic memory by the hippocampus. When a unit in the substitution space is activated both its address (color coded) and a trace are retrieved. As in dynamic binding, the address is serially transmitted to the switchboard, which tries to match it to a slot. The trace information is used to filter for slots with a matching trace.  $\langle T1, T1' \rangle$ ,  $\langle T2, T2' \rangle$  and  $\langle T3, T3' \rangle$  are pairs of identical traces stored inside slots and bound semantic units.

Figure 8.1 repeats a simplified version of Figure 5.10 from section 5.7, to illustrate how episodic memories are replayed from their traces. Suppose that trace  $T1$  is primed (e.g., by the first word of a sentence) then, as explained in section 5.7, its topological address (encoded in the root of the unit) is serially transmitted to the switchboard, which projects to a neighborhood in the substitution space

where the matching slot (with trace  $T1'$ ) can be found (assuming  $T1'$  and  $T1$  are near).<sup>2</sup> The same procedure is repeated for the successor trace  $T2$ , which is linked via the compressor node, etc., until all the original bindings of the episode have been restored.

### The hippocampus implements a switchboard function

If, as predicted by episodic-HPN, episodic memories are encoded as distributed traces that are dynamically bound, then that explains that a single system, the hippocampus, is responsible for both replay of stored episodes and flexible association of semantic elements in processing novel events, because both functions involve dynamic binding. This would imply that the hippocampus implements a switchboard function, and that is indeed consistent with the fact that it is situated at the central ‘gateway’ of the brain: the switchboard must be connected to semantic elements that are distributed throughout the entire cortex in order to be able to dynamically bind them.

From this perspective, *encoding* an episodic memory amounts to making the temporary tags, that are involved in dynamic binding of an event, persistent as episodic memory traces. (Recall from section 5.7 that a critical component of the switchboard solution for dynamic binding is a tagging system, whose function is to attach a unique ‘tag’ to the units that participate in a binding, such that the bindings are kept ‘alive’ for some time in working memory.) Specifically, in episodic-HPN the temporary tags that bind the most probable derivation of a sentence are turned into traces (of a stored derivation) by converting them from short term into long term memories.

This idea is consistent with recent neuro-biological findings on memory consolidation, concerning the ‘illegibility’ of synapses for long-term potentiation [e.g., Izhikevich, 2007]. For instance, according to the ‘synaptic tagging and capture hypothesis’ [Redondo and Morris, 2010] long-term memory potentiation follows a two-step mechanism, whereby in the first step a so-called ‘tagged state’ is induced that only creates the *potential* for a lasting change in synaptic efficacy.

### The HPN substitution space is an instance of a ‘relational network’

According to [Eichenbaum, 2004, e.g.] the hippocampus structurally organizes episodic memories in ‘relational networks’, by linking them through semantic elements of episodes that share the same context (see Figure 2.7 in section 2.7.1).

Such an organization allows for transitive inference through flexible combination of episodes (which explains why the hippocampus is needed for novel problem

---

<sup>2</sup>An alternative solution is that within the switchboard an episodic archive is stored, consisting of a list of addresses indexed by episodic trace numbers. In that case the address of a matching trace  $T1'$  can be found by querying the list with  $T1$ , even if the topological locations of the units where  $T1$  and  $T1'$  are stored are remote.

solving). An example of a relational network is the navigational (cognitive) map, found in the hippocampal place cells of rats, which enables them to navigate their way in a maze (see [e.g., O’Keefe and Nadel, 1979] and Figure 2.8).

The *substitution space* of episodic-HPN, in which the units are enriched with episodic traces, can be regarded as an instance of a relational network, or navigational map, for the language domain. Like relational networks, it organizes episodes, distributed as traces, in a structured memory space, and links them by shared semantic units (containing multiple traces). This allows priming and flexible association with other episodes, allowing for productive use of language (as exemplified in the episodic grammar). Transitive inference (e.g., for finding new routes in a navigational map) is operationalized through the topology of the substitution space. Thus, like rats learning to navigate a maze by structuring spatial episodes (recall Figure 2.8), language learners construct a ‘navigational’ (topological) map of language by reorganizing episodic linguistic experiences. This process is modeled by episodic-HPN.

Further, episodic-HPN supports at the implementational level the claim of Eichenbaum [2004] that structuring of episodes in relational networks is instrumental in memory consolidation. In section 8.4.1 I explained that the topology of the substitution space (i.e., a semantic memory) is shaped from episodic memory as a result of selecting the shortest derivation parses, and adjusting the topology accordingly.

## 8.5 Discussion

### 8.5.1 Relation to other neural network models of memory consolidation

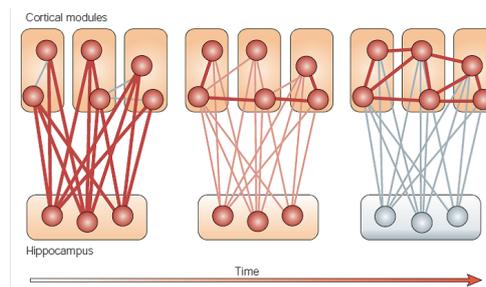


Figure 8.2: Standard consolidation model. The top layer shows the cortical modules containing distributed episodic representations. These are linked to sparse representations in the hippocampus in the bottom layer. (From Frankland and Bontempi [2005].)

Since the current work purports to be a domain general model of memory

consolidation, it is interesting to place it in the context of other modeling work in this field. In the computational neuroscience literature most modeling work on memory consolidation is based on one of two major theories [e.g., see the review by Frankland and Bontempi, 2005]: one is the so-called *standard consolidation model* [e.g., Squire and Alvarez, 1995], and the other is called the *multiple trace model* [e.g., Nadel and Moscovitch, 1997]. According to the standard consolidation model (see Figure 8.2), episodic memories are initially stored in connections between the hippocampus and the cortex. As the hippocampus *replays* the memories (presumably during sleep), the cortico-cortical connections are strengthened, while the dependency on the hippocampus is gradually diminished. This accords with retrograde amnesia studies that show that after a hippocampal lesion, or stroke, episodic memories are lost retro-actively, but memories from a long time before the accident are often preserved, apparently because those have moved out of the hippocampus.

Computational models of memory consolidation in this tradition typically hypothesize that the neocortex and hippocampus act as ‘two complementary learning systems’ [e.g., McClelland et al., 1995, Battaglia and Pennartz, 2011, O’Reilly and Norman, 2002, Tse et al., 2007]: while the hippocampal system is a fast learner, which stores episodes in one shot as they are processed, the neocortex is a slow learner, which gradually assimilates the episodic experiences within a semantic memory system that represents general, statistical knowledge. This division of tasks is usually motivated by the argument that the different requirements that the human memory system has to cope with, namely learning specifics about the environment (i.e., episodic memory) versus extracting generalities (i.e., semantic memory) are apparently mutually incompatible. According to O’Reilly and Rudy [2001] a single representation cannot simultaneously capture both generalities and specifics, nor can a single learning system combine slow, statistical (integrative) learning with fast automatic recording. In the same vein McClelland et al. [1995] motivate the complementary learning systems approach from the problem of ‘catastrophic interference’ — this is the phenomenon that, beyond a certain threshold, old memories are overwritten by new ones, which is a known problem for parallel distributed neural networks.

To deal with the trade-off between rapid learning of episodic events and slow learning of statistical structure O’Reilly and Rudy [2001], O’Reilly and Norman [2002] propose a modular network architecture, consisting of several hippocampal and cortical networks (schematically illustrated in Figure 8.2). While the hippocampal networks form sparse representations of episodes (allowing for pattern separation), neurons in the hippocampus are conjunctively bound to distributed representations of the same episodes in the cortical networks (allowing for pattern completion).

In the second tradition, the multiple trace theory (MTT) [Nadel and Moscovitch, 1997] holds, in contrast to standard consolidation theory, that episodic memory traces remain in the hippocampus forever. According to MTT each time an

episodic memory is reactivated this happens in an different context, and consequently a new memory trace is created, with overlapping features in the neocortex, but with a distinctive pattern in the hippocampus, where ‘context’ is encoded. As a result memories that are often reactivated are associated with a larger number of traces, hence can be retrieved from multiple cues and become more stable. Neural network models in this tradition also assume a modular approach, in which episodic memories are encoded in conjunctive connections between a hippocampal module (with sparse encoding), and a distributed pattern in the neocortical network module [e.g., Nadel et al., 2000]. In this respect also the MTT can be regarded as an instance of the ‘complementary learning systems’ approach.

### **Critique of the complementary learning systems approach**

A challenge for the complementary learning systems framework is the massive number of connections between the hippocampus and the cortex that result from conjunctive coding of episodes, since every new episode (many thousands a day) recruits at least one dedicated binding neuron in the hippocampus, and must establish connections to the cortex.

Another, more fundamental problem for conjunctive binding of episodes is that it lacks the flexibility to process unseen events by associative expression of stored memories. Yet, as discussed in section 2.7.1, this ability has been suggested by Eichenbaum and Cohen [2001] to be the driving force behind memory consolidation. Hummel et al. [2004] argue that a major limitation of conjunctive coding is that it does not afford to make relational inferences (nor to generalize) beyond specific stored role-filler conjunctions, because conjunctive coding represents all elements of the binding as a single, indivisible entity.

A related problem is that conjunctive bindings provide no temporal or hierarchical structure to episodes. Yet, structure is required for instance to encode causal relations (used for predictions) [e.g., Eichenbaum and Fortin, 2009], syntactic and linear precedence relations in language, or to distinguish the different relations between roles and participants in an event [e.g., Shastri, 2002].

### **Comparison to episodic-HPN: efficient and flexible encoding of episodes using dynamic binding**

The view of memory consolidation that is implied by episodic-HPN differs in important respects from the complementary learning systems approach, because in episodic-HPN episodic and semantic memory are fully integrated within a single system. The learning algorithm of episodic-HPN demonstrates that, contrary the claim of O’Reilly and Rudy [2001], the requirements of fast capture of specific detail versus slow, statistical learning of generalities are *not* incompatible, but can be satisfied simultaneously. Yet, this trade-off was in fact the main motivation for having two complementary learning systems.

The argument of McClelland et al. [1995], that the hippocampus is needed as a buffer in a complementary learning system to prevent catastrophic interference, does not hold either. This argument only makes sense if one assumes that the neocortex is a parallel distributed neural network, yet in this thesis I have defended a localist network view of the neocortex. According to the episodic-HPN model of memory consolidation the reason that the hippocampus is involved as a ‘gateway’ for episodic memory encoding is not to prevent catastrophic inference, but for its role in dynamic binding, as explained in section 8.4.2.

One of the advantages of dynamically binding episodes is that it saves dedicated binding neurons and connections, hence it does not suffer from the ‘massiveness of the binding’ problem. Whereas in the conjunctive binding approach every individual episode requires hardwired connections between the hippocampus and the cortex, in the dynamic binding approach of episodic-HPN the same hippocampal element (that is, a single unit in ‘substitution space’) can be shared by all episodes, and a single projection from this element to the cortex suffices for activating the sensory circuits associated with the element. Internal connections of episodes are handled by the switchboard and the traces.

Further, dynamic binding of episodes in episodic-HPN is responsible for ‘flexible expression of memories’, which is a condition for the ability to make analogical inference in support of memory consolidation.