Confabulation Theory

Robert Hecht-Nielsen

Confabulation Theory

The Mechanism of Thought

With 150 Figures and 2 DVD-ROMs

Springer

Author

Robert Hecht-Nielsen

University of California, San Diego
La Jolla, CA 92093-0407
USA
r@ucsd.edu
http://r.ucsd.edu

Library of Congress Control Number: 2007929728

ACM Computing Classification (1998): I.2, I.4, I.5, I.6, I.7, F.1, H.1, H.5

ISBN 978-3-540-49603-8 Springer Berlin Heidelberg New York

The publisher and the author accept no legal responsibility for any damage caused by improper use of the instructions and programms contained in this book and the DVD-ROMs. Although the software has been tested with extreme care, errors in the software cannot be excluded.

Springer is a part of Springer Science+Business Media

springer.com

Typesetting and Production: LE-TEX Jelonek, Schmidt & Vöckler GbR, Leipzig
Cover design: KünkelLopka Werbeagentur, Heidelberg

Printed on acid-free paper 45/3180/YL - 5 4 3 2 1 0

Preface

This book introduces a body of recent discoveries regarding the mechanism of thought. These discoveries together constitute the first concrete, detailed, and comprehensible scientific theory of how thought works: *confabulation theory*. Experiments with computer-simulated versions of the theory's proposed neurophysiological components are presented. These demonstrate information-processing characteristics and capabilities that support the theory's scientific viability.

This book is designed to serve both the neuroscience and computational intelligence communities in assimilating confabulation theory. The book is inherently interdisciplinary – with all readers expected to absorb all of the material. This is predicated on the assumption that the futures of both communities will be significantly dependent upon cross-community inspiration.

In organizing this book, an explicit attempt has been made to make it suitable for use in three modes:

1. As courseware for an introductory one-quarter or one-semester graduate or advanced undergraduate course on confabulation theory
2. As a vehicle for efficient professional self-study
3. For use in connection with a concentrated introductory short course on confabulation theory and confabulation architectures

In all modes, it is recommended that the first step be reading Chap. 1. Chapter 1 introduces confabulation theory and explains the organization of the book.

Next, the video presentation on the book's two DVDs should be viewed[1]. For a course, this will require four or five class sessions. For self-study or a concentrated course, this video presentation should be viewed in two sessions; with a brief intermission between Disk 1 and Disk 2 (i.e., after the end of the Confabulation Neuroscience section of the presentation). As the video is viewed, the relevant viewcell of Chap. 2 should be referred to at the same time. Glancing at a printed version of each viewcell while viewing its video presentation can significantly enhance the understanding and retention of the material being discussed.

Following the initial viewing of the DVD video in a university course, the remainder of the class sessions of the semester or quarter can be used to go into

[1] The video presentation, which is provided on the two DVDs attached to the cover of this book, can be played on a standard home DVD machine, or using a computer. The two PDF files on the DVDs (both files are included on both DVDs) can only be accessed using a computer.

the material of the presentation in more depth. A suggested approach is to proceed through the viewcells of the presentation again in sequence – a few in each subsequent class period. For example, for a course with a total of 24 class sessions, each class session would cover an average of four or five viewcells. At the outset of each class period, the portion of the video covering the target viewcells is played on the classroom screen. Then the notes for those viewcells (from the *Presentation Notes* PDF file on the book's DVDs) are projected onto the classroom screen and discussed, one at a time, in sequence. During discussion of each viewcell, course participants are directed by the instructor to specific passages in the chapters of the book for more in-depth material that the instructor then explains on the classroom blackboard. For convenience in instructor preparation, many references to pertinent chapters are provided in the Presentation Notes. This same approach can be used for self-study, but with a larger number of viewcells covered during each study session.

Instructors who wish to include an experimental assignment for course participants can assign replication of the "add-one-word-to-a-text-string" confabulation experiment of presentation viewcells 32 through 36 (see Chap. 2). This will require access to a large (tens of millions of words), clean, text corpus (or a word list and one-thru-five-grams). One source for such a corpus (and for English n-grams) is the Linguistic Data Consortium (http://www.ldc.upenn.edu/). Chapters 4 and 7 explain the details of the "add-one-word-to-a-text-string" confabulation experiment. There is quite a bit of work involved in this experiment, so it should be assigned before class meeting 10. Experience has shown that course participants get more out of the assignment if they must do the whole thing themselves. However, it is also possible to divide up the work among different groups of students (some create the four knowledge bases, others build the experiment GUI, and others build the confabulation system). Experience has shown that students get the most out of the assignment if they are required to present the background, the experiment definition, the experiment implementation, and the experimental results in a formal "conference-presentation style" PowerPoint presentation to the class (the author does not agree with the currently fashionable objections to PowerPoint). Each participant presents the entire story, from scratch. This high redundancy might seem unacceptable; however, experience has shown that this actually works very well. The participants acutely sense, and benefit from, the differences in presentation effectiveness.

Other possible course assignments might involve course participants giving formal presentations to the class on expanded course topics. For example, students might research and then present on:

- The formal proofs of Theorem 4.1 or 4.2 from Chap. 4
- The benefits of the *bandgap* formalism of input intensity calculation used in Chap. 6 and described in detail in Chap. 7

- A survey of existing attractor neural networks for carrying out confabulation based upon references in Chaps. 5 and 8 – with comments explaining the significant deficiencies of each approach (all known approaches have major deficiencies that make them incapable of carrying out confabulation)
- A discussion of direct experimental exploration of the function of different portions of human cerebral cortex based upon references in Chaps. 5 and 8

Alternatively, such assignments might be made optional; allowing the more capable participants to be fully challenged by the course.

Neuroscience-oriented courses can augment the book's material with instructor discussions of how radically confabulation theory challenges some strongly embedded traditional views such as:

- Continuous, rather then categorical (i.e., symbolic), cortical "feature detector" neuron response
- Persistent qualitative notions that cortical information-processing involves ongoing, intrinsic analog "dynamical processes" and not discrete, repeatable, externally directed convergence to "attractor states"
- Related ideas that cortical processing "just happens naturally" (what might be termed a *data flow* viewpoint) whenever the necessary inputs to a local portion of cortex happen to arrive

Class participants are then invited to research, and present to the class, hard experimental neuroscience facts that challenge traditional views. For example:

- The reliable and repeatable nature of a host of stimulus–response experiments in which both movements and thoughts must be executed at precise stimulus-specified times (e.g., music and dance)
- Reaction-time experiments (wherein human subjects often carry out complex decision-making and reaction-generation processes in time intervals only slightly longer than the involved axonal propagation delay)
- The ubiquity of "gestalt switching" (winning symbol changing) in both perceptual and behavioral realms
- The clear EEG, MEG, and fMRI correlates of the activation of cortically localized thought processes

Replication of the multiconfabulation experiments of Chap. 6 would involve more effort than a one-semester or one-quarter course could reasonably accommodate. Such a project would require a number of capable researchers and take a year or more to complete. This might be an appropriate first project for a research team that wishes to inaugurate work on confabulation theory-based neuroscience or neurotechnology. Research sponsors might be willing to support such work as a sensible lead-in to original research.

This book marks the end of the initial discovery process of confabulation theory, which began in 1968. There are many people who have contributed significantly to this quest: my wife Judi and our son Marcus, domestic cat Zeus Hecht-Nielsen, other family members, friends, collaborators, TAs, students, colleagues, UCSD, advisors, supporters, research sponsors, opponents, and the millions of scholars whose accumulated research I have benefited from. Special thanks to my current research collaborators Robert W. Means, Kate Mark, Syrus C. Nemat-Nasser, and Andrew Smith; and to those key figures who helped make the last 20 years of this research possible, especially: Todd Gutschow, Robert L. North, Paul Glenn, A. T. La Prade, Gordon Davidson, Ken Linhares, Simon Ramo, James Martin, Carol Bonomo, Tom Grudnowski, Larry Rosenberger, Eric Educate, and Andrea Fike. Finally, thanks to Doug Ramsey and Alexander Matthews of UCSD Jacobs School of Engineering and Calit2 for filming and editing the presentation and creating the DVDs. Research support from Fair Isaac Corporation and ONR is gratefully acknowledged.

San Diego, May 2007 Robert Hecht-Nielsen

Contents

1 **Introduction** **1**
1.1 In the Beginning 1
1.2 Cerebral Cortex and Thalamus: The Seat of Cognition 5
1.3 The Four Key Elements of Confabulation Theory 7
1.4 Cognitive Brain “Hardware” and “Software” 14
1.5 Implications of Confabulation Theory 15
1.6 Content of the Book 17

2 **Video Presentation Viewcells** **21**

3 **The Mathematics of Thought** **73**
3.1 The Constructs of Confabulation Mathematics 73
3.2 A Single Confabulation 74
3.3 Multiconfabulation 78
3.4 Symbols: The Universal Language of Cognition 79
3.5 Knowledge Links 81
3.6 Neuronal Implementation of Knowledge Links 83
3.7 Neuronal Implementation of Confabulation 87
3.8 The Origins of Confabulation Theory 88
3.9 Discussion 90

4 **Cogent Confabulation** **91**
4.1 Introduction 91
4.2 Cogency and Confabulation 92
4.3 Confabulation Examples 94
4.4 Discussion 96

5 **Confabulation Neuroscience I** **99**
5.1 Introduction 99
5.2 Confabulation Theory Element 1 100
5.3 Confabulation Theory Element 2: Knowledge Links 102
5.4 Confabulation Theory Element 3: Confabulation 104
5.5 Confabulation Theory Element 4: The Origin of Behavior 106
5.6 Confabulation Theory Mathematics 107
5.7 Multiconfabulation 110
5.8 Discussion 115

6 The Mechanism of Thought 117
6.1 Introduction 117
6.2 Confabulation Architecture 119
6.3 Learning 121
6.4 Thought Process 121
6.5 Discussion 123
Appendix 6.A: Methods 125
Appendix 6.B: Complete List of Experiments and Results 130

7 Mechanization of Confabulation 139
7.1 Introduction 139
7.2 Training and Education 143
7.3 Language Cognition 149
7.4 Sound Cognition 167
7.5 Visual Cognition 177
7.6 Discussion 191

8 Confabulation Neuroscience II 193
8.1 Introduction 193
8.2 Summary of the Theory 194
8.3 Implementation of Modules 197
8.4 Implementation of Knowledge 208
8.5 Implementation of Confabulation 212
8.6 Action Commands 219
8.7 Discussion 225

References 229

Index 235

1 Introduction

The books, papers, and lectures which I appreciate most start by giving the punch lines of the presentation in a simplified and immediately understandable form. The first four sections of this chapter are intended to provide a summary of this type for confabulation theory. Section 1.1 provides background perspective and a nutshell description of confabulation theory. The following three sections then provide a progressively more detailed overview of the human case (with deliberate repetition to aid learning these new concepts). Section 1.5 discusses some of confabulation theory's implications. Finally, Sect. 1.6 provides a brief overview of the book's content.

1.1 In the Beginning

There is strong neuroscience evidence of many kinds suggesting that the initial phase of the story of life on Earth ended about 580 million years ago with a large, rapid, and sustained (to the present) increase in atmospheric oxygen concentration (Canfield et al. 2007, Fike et al. 2006, Kerr 2006). Immediately thereafter, a profusion of macroscopic moving animals emerged (the "Cambrian explosion" of species). The fitness advantages of complex, purposeful movement rapidly drove the evolutionary development of articulated bodies, muscle complements, and the brains and sensory systems needed to purposefully run them.

Movement involves smooth, coordinated control of ensembles of discrete muscles by the animal's brain. Each muscle is supplied with a single neuronal input signal controlling its "analog," continuously variable, level of contraction. Shortly after the emergence of animals capable of sophisticated movement, a new design possibility arose: The extensive neuronal machinery developed to control animal movement could easily be expanded and these additions could be used to control brain *modules*: discrete bodies of neuronal tissue specifically evolved to exploit the pre-existing neuronal muscle-control mechanisms. Instead of conferring motility, these new brain module "movement" processes would carry out a type of information processing called *cognition* or *thinking*. The enormous success of this evolutionary adaptational "redeployment" of movement control led to today's ubiquity of cognition in macroscopic animals (trout, bees, ravens, humans, octopi, et al.). Further, the neuronal mechanisms of cognition were subsequently further adopted as the starting basis for additional brain functions that subsequently evolved, such as the cognitive learning control system (entorhinal cortex, hippocampus, amygdala, etc.) of mammals. This book concerns itself with explaining the mechanism of thought in detail – with primary focus on the human example.

The purpose of each cognitive module (of which humans have about 4,000 – in contrast with our 700 individual muscles) is to describe one *attribute* that an *object* of the animal's mental universe may possess. This description usually takes the form of *activating* one of a large number of *symbols* (each represented by a small collection of specialized neurons) that are contained within the module. The vast majority of symbols within each module develop during childhood and then remain stable throughout life. Symbols are the fundamental, fixed *terms of reference* that must exist if knowledge is to be accumulated and used over long periods of time.

An individual axonal *knowledge link* (of which the average human adult possesses billions) unidirectionally connects one *source symbol* in one module with one *target symbol* in a second module. These links arise as a result of meaningful causal co-occurrence of the involved pair of symbols (*a la* Donald Hebb). [NOTE: Besides symbol co-occurrence, most animals also impose (e.g., via a centralized cognitive learning control system; as in mammals) the requirement that a new knowledge link also be associated with a reduction in a drive or goal state. Imposition of this requirement has many important advantages – not least of which is the avoidance of a vast buildup of low-value knowledge. Because it is tangential to understanding the mechanism of thought, this "knowledge relevance" requirement and its formidable implementation machinery (it needs to be formidable; because hours often elapse between the temporary establishment of a knowledge link – which the neurons directly involved in implementing the link carry out via instantaneous temporary synapse strengthening – and the realization that this candidate link was involved in a drive or goal reduction) will be ignored in this book. When we need to actually construct knowledge links (e.g., for conducting computer experiments with confabulation), we simply require that all of the knowledge links that are allowed are "of significant value" using some simple criterion. This approach works well for a number of applications – further reinforcing the decision to skip detailed discussion of animal cognitive learning control systems.]

The set of all knowledge links connecting the symbols of one module with the symbols of a second module are collectively termed a *knowledge base* or cortical knowledge *fascicle*. In humans, the set of all cortical knowledge fascicles is, by far, the most massive single brain structure. The capacity for accumulating a vast number of knowledge links is the single most important attribute of the human brain (at an average rate, for most people, exceeding one new knowledge link per second of life); followed by the large symbol capacities of human modules.

Besides implementing symbols, modules also carry out one, and only one, cognitive information processing operation: *confabulation*. Confabulation is the analog of contraction in a skeletal muscle. It occurs only upon receipt of a deliberate *thought command* input to the module. Thought command signals originate in subcortical structures. Both because not much is known, and to keep the story of confabulation theory focused on cognition, the exact origin of thought commands, and the details of the neuronal processes involved (which involve many subcortical brain nuclei – mostly exactly the same ones as in movement) will be ignored in this book. The origin of the *action commands* that ultimately launch all movement and thought processes (i.e., *behaviors*) **will** be briefly discussed, because they arise as a direct product of cognition (see below).

Strangely, as with a motorneuron signal to a muscle, a thought command is a graded, analog, signal. This is one of several aspects of cognitive information processing that make it starkly alien in comparison with existing concepts such as algorithmic and rule-based computing.

In the milliseconds leading up to a particular target module being commanded to begin (or intensify) a confabulation "contraction," axonal knowledge links from source symbols which are currently *excited* on other selected source modules deliver *input excitation* to neurons representing each knowledge link's target symbol. [The ensemble of modules transmitting excitation are deliberately selected by the overall thought process being executed (thought processes are learned, stored, and recalled in the same manner as movement processes).]

Confabulation is the process of selecting that one symbol (termed the *conclusion* of the confabulation) whose representing neurons happen to be receiving the highest level of excitation. In the case of a single target module undergoing confabulation, this is a simple "winner takes all" competition among the symbols of the target module. At the end of a confabulation all of the neurons which represent the winning symbol are transmitting at high efficacy through any knowledge links that have the conclusion symbol as their source. Through the use of a *neuronal attractor network* circuit contained within the module, a simple confabulation can often be completed in under 100 ms, even if the module implements hundreds of thousands of symbols. Conclusions reached by confabulations in the recent past can be used as the sources of knowledge link input to subsequent confabulations. Conclusion symbols subsequently selected to supply such input are often referred to as *assumed facts* of those subsequent confabulations.

In cognition, single confabulations are rare (much as movements involving contraction of only a single muscle are rare). Usually, thought processes involve an ensemble of tens to hundreds of modules being confabulated contemporaneously during overlapping time intervals – with intercommunication between the symbols of the modules at various points during the gradual, expertly controlled, "contraction" to a single "winning" symbol on each module. This is *multiconfabulation*. A multiconfabulation is typically much more powerful than a single confabulation because it facilitates a process of gradual convergence to a set of "mutually consistent" conclusions; reached by means of mutual communication between the ever-shrinking intermediate sets of candidate conclusions. Multiconfabulation facilitates the application of massive numbers of *relevant* knowledge links (each emanating from a symbol which, at least at that stage of the contraction process, is a viable candidate to be the final conclusion of that module). Properly executed, a multiconfabulation allows multiple opportunities to "cross-check" the lists of not-yet-eliminated candidate symbol conclusions to ensure that the final conclusions reached (collectively termed the *confabulation consensus*) are mutually consistent with respect to the available knowledge. Thus, the slowed convergence process of multiconfabulation (with the rising "contraction" thought command signal corresponding to the gradual shrinking of the list of remaining candidate conclusions from which the final single winning symbol will be selected) is an essential aspect of cognition. An information processing system employing carefully and skillfully coordinated smooth information processing (thought) commands to the involved processors (modules) is starkly alien in comparison with all existing concepts of information processing.

How can confabulation – a simple competition process between the symbols of a module on the basis of which symbols are receiving the most axonal excitation – be the complete and final explanation for all aspects of cognition? This would seem to imply that, in some sense, confabulation is a powerful, general purpose, universal decision-making procedure. Surely there must be some new and powerful mathematics underlying it. And there is. Describing and characterizing this surprising and strange cognitive mathematics is a main focus of this book.

Finally, a key unanswered neuroscience question is the origin of *behavior* (thought processes and movement processes). Obviously, animals launch many behaviors every minute – often many per second. There must be a unified source of these actions. The shockingly simple answer is that every time a confabulation is completed, action commands, uniquely associated with the winning conclusion, are instantly launched and sent to subcortical structures (e.g., the basal ganglia) for evaluation and, perhaps, execution. All non-reflexive and non-autonomic behavior originates in this manner.

The axonal associations between each symbol in a module and its fixed set of action commands are termed *skill knowledge*. While skill knowledge is stored in cerebral cortex, it is established and modified by subcortical brain nuclei. Skill knowledge is very different from cognitive knowledge – e.g., far from being very long lasting like cognitive knowledge, skill knowledge, if unused, fades rapidly – often within a few weeks. Skill knowledge is "use it or lose it." Also, skill knowledge is inherently "overwritable," allowing more recent skill practice session performances to "overwrite" older, presumably less competent, skill knowledge. In order to remain focused on cognition, very little is said in this book about skill knowledge.

A major advantage of cognition is that all cognitive knowledge is *interoperable*. The knowledge links delivering excitation to a particular thought process might emanate from symbols representing auditory, visual, linguistic, or even movement process attributes of mental world objects. The type of attribute that their source symbols encode makes no difference: the knowledge link excitation input to the symbols of the involved target modules are simply approximately summed up. To appreciate the power of this capability, consider the difficult challenge faced by an algorithmic information processing researcher who is attempting to combine image and sound data from a theatrical motion picture to accurately recognize specific movie actors.

Cognition is a "core competence" of macroscopic multicellular Earth animals. In each taxonomic category, species with particularly high cognitive skill stand out: bees among insects, humans among primates, jays and ravens among birds, cetaceans among aquatic mammals, etc. Anthropocentrism puts humans on the highest pedestal; but all cognitive "champions" have their distinctive relative superiorities. I leave it to philosophers, SETI researchers, future interstellar explorers and theologians to incorporate the insights of confabulation theory into larger points of view and to address sweeping universal questions (such as: Is confabulation the unique extant approach to natural intelligence in our universe, or are there others?). This book concentrates on the confabulation theory explanation for human cognitive function and on the use of confabulation theory as the basis for building intelligent machines.

1.2 Cerebral Cortex and Thalamus: The Seat of Cognition

There is strong neuroscience evidence of many kinds suggesting that the "information-processing" involved in all aspects of cognition (seeing, hearing, planning, language, reasoning, control of movement and thought, etc.) is carried out by the cerebral cortex and thalamus. There is also strong evidence that the "cognitive knowledge" used in this processing is stored in the cerebral cortex. Beyond vague statements of this sort, at present essentially nothing is known about how cognition (which will also be referred to in this book as *thinking*) works, or about what cognitive *knowledge* is.

This book presents the first concrete and detailed (and thus falsifiable) scientific theory of how thinking works. This *confabulation theory* proposes the specific neuroanatomical structures, and their functions, that are involved in human cognition.

The two main human neuroanatomical structures postulated by confabulation theory to be involved in the implementation of thought are *thalamocortical modules* (Fig. 1.1) and *knowledge bases* (Fig. 1.2). These structures, which constitute the "information-processing hardware" used to carry out thought, exist within the cerebral cortex and thalamus. The human brain possesses roughly 4,000 thalamocortical modules and roughly 40,000 knowledge bases[2]. All vertebrates (and even invertebrates such as bees and octopi) are postulated to possess functionally analogous structures, albeit in smaller quantities.

[2] For concreteness, confabulation theory specifies many numeric values quantifying aspects of the theory's postulated human neuroanatomical structures. These can be thought of as crude, rough order of magnitude, estimates of means; with most quantities also having significant variance. For simplicity, value accuracy and variability are not discussed.

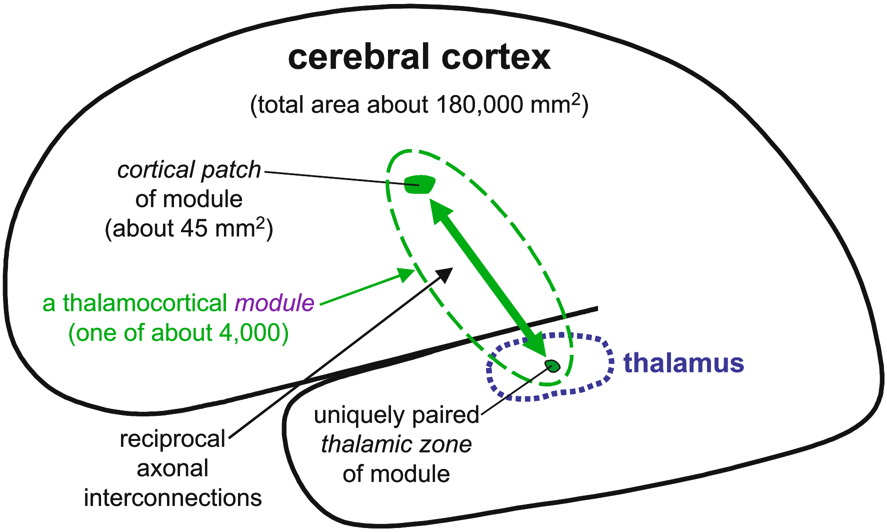

cerebral cortex
(total area about 180,000 mm²)
cortical patch of module (about 45 mm²)
a thalamocortical module (one of about 4,000)
thalamus
uniquely paired thalamic zone of module
reciprocal axonal interconnections

Fig. 1.1. A *thalamocortical module* (one of roughly 4,000 in the human brain). Each thalamocortical module is comprised of a small *patch* of cerebral cortex and a uniquely paired small *zone* of thalamus. The cortical patch of each module is reciprocally axonally connected with the thalamic zone of the module. The cortical patches of different modules are largely disjoint (partial overlaps do likely occur). Similarly for their thalamic zones. The union of the cortical patches of all thalamocortical modules comprise the entire area of cerebral cortex. However, the union of the thalamic zones of all modules do not comprise all of the thalamus

The cortical neural tissue encompassed by each thalamocortical module bears resemblance to that of the "cortical columns" proposed decades ago (Mountcastle 1988; Paxinos and Mai 2004), except that the cortical component of a module is roughly 200 times larger in volume than a cortical column. The postulated functions of thalamocortical modules are also completely different from those envisioned for columns.

Knowledge bases are related to the axonal links between pairs of cortical "neuron populations," as postulated vaguely by Hebb 57 years ago (Hebb 1949) and more concretely and recently by Abeles (Abeles 1991).

The level of description of function offered by confabulation theory is one level up from that of the individual neurons. The study of how these functions are implemented at the neuron and molecular levels is termed *confabulation neuroscience*. Since very little is known, the discussion of confabulation neuroscience in this book (principally Chaps. 2, 3, 5, and 8) is mostly speculation, and will likely require significant revision as more is learned.

As noted in bibliographic citations throughout the book, and discussed explicitly in Chaps. 3, 5, and 8, confabulation theory is strongly related to many bodies of past research.

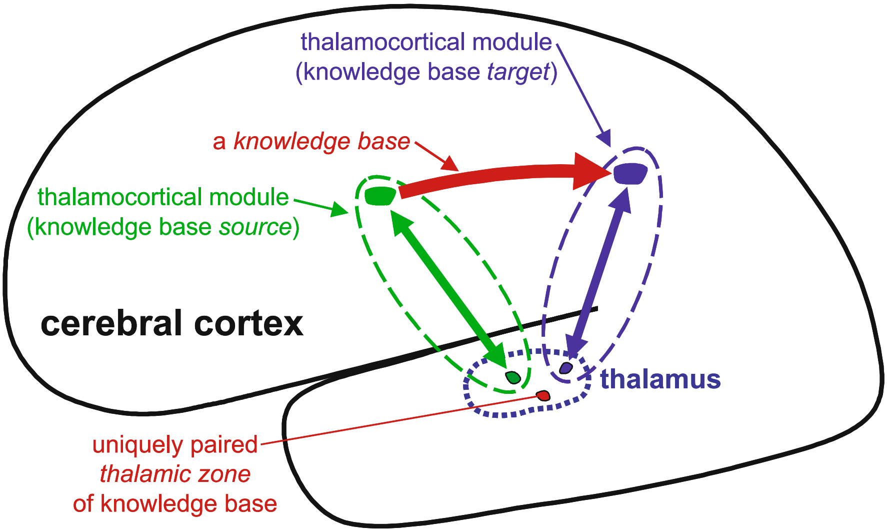


Fig. 1.2. A cognitive *knowledge base* (one of roughly 40,000 in the human brain). Roughly 40,000 ordered pairs of thalamocortical modules (*source* and *target* modules) are selected (by genetically specified developmental processes carried out in childhood) to each have their cortical patches unidirectionally linked by a knowledge base. Each knowledge base is comprised of a large number (often millions) of individual *knowledge links*. Much like a thalamocortical module, each knowledge base is postulated to be paired with a unique, dedicated zone of thalamus which is postulated to be involved in that knowledge base's functional *enablement*. The combination of the thalamic zones of the modules and knowledge bases make up the vast majority of the thalamus

1.3 The Four Key Elements of Confabulation Theory

Today, the cognitive information-processing and cognitive knowledge acquisition, storage, and use functions of cerebral cortex and thalamus are completely unknown. Confabulation theory specifies them completely. In particular, confabulation theory postulates four key functional elements (#s 1, 3, and 4 implemented by thalamocortical modules and #2 implemented by knowledge bases) which together comprise *the neuronal information-processing "hardware" of thought*. These four key elements, and the manner in which thalamocortical modules and knowledge bases implement them, are each individually sketched in the four sub-sections of this section. The manner in which these functional hardware elements are used to implement thought is explored in detail in the book's video presentation (and the associated presentation notes) and in Chaps. 3, 4, 6, and 7.

1.3.1 Confabulation Theory Key Element #1: Each Thalamocortical Module Describes One Mental Object Attribute

Each thalamocortical module (Fig. 1.3) is used for describing one *attribute* which an *object* (sensory, language, abstract, movement process, thought process, plan, etc.) of the mental universe may possess. To describe its attribute, the module is equipped with a large collection of *symbols*. When utilized for describing an object, a module typically *expresses* one symbol chosen from its collection. The

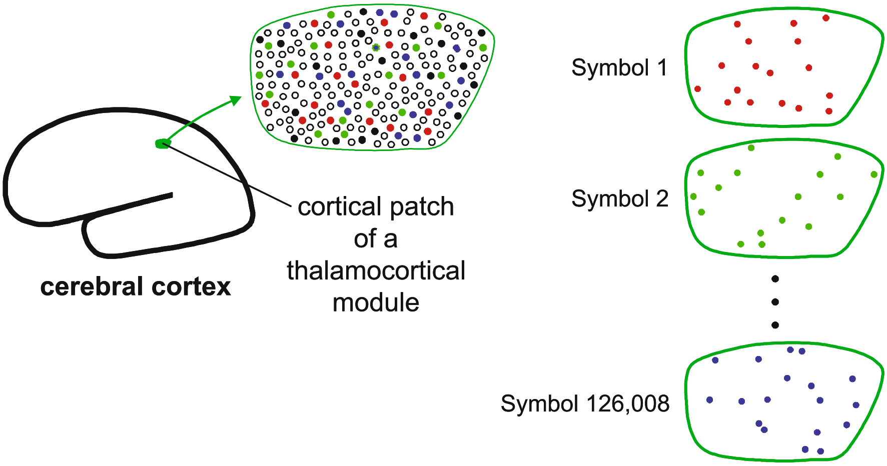


Fig. 1.3. A primary function of each thalamocortical module is to describe exactly one *attribute* that an *object* of the mental universe (a sensory object, a motor process object, a thought process object, a plan object, a language object, etc.) may possess. To carry out this object – attribute – description function, each module implements a large collection of *symbols*. When utilized for describing an object, a module typically *expresses* one symbol chosen from its collection (primary sensory and motor modules usually express multiple symbols). Each symbol is represented by roughly 60 neurons selected (approximately uniformly at random) from a special population of *symbol-representing neurons* (shown as colored dots within the enlarged depiction of the module's cortical patch) that reside within the cortical patch of the module. Here, a module with 126,008 symbols is depicted. Each symbol's subset of 60 neurons is shown schematically. Symbols are mostly formed in childhood and then remain stable throughout life – they are the *stable terms of reference* that must exist if knowledge is to be accumulated across decades. The famous *binding problem* (von der Malsburg 1981) does not apply to confabulation theory because each of the attribute description symbols of an object is typically linked to many of the others pairwise by *knowledge links* (see Sect. 1.2.2). In effect, a mental world object *is* any reasonably large subset of its pairwise-linked attribute description symbols. Thalamocortical module symbol sets (the collection of different descriptive terms for representing the object attribute that the module is responsible for encoding) are the first of the four key functional elements of confabulation theory

symbols of a module are mostly created in childhood and are stable over decades. Symbols are the *stable terms of reference* which must exist if knowledge is to be accumulated over long periods of time. For example, in a human a particular thalamocortical module might be responsible for representing the *name* of an object. This module might possess 128,008 symbols, representing words, phrases, and punctuations such as: `mother`, `father`, `President Kennedy`, `Bunsen burner`, `lunar regolith`, `candy`, and `Candy`.

1.3.2 Confabulation Theory Key Element #2: Knowledge Links Connect Pairs of Co-occurring Symbols

Although the concept of cognitive human knowledge – something which is acquired, stored, and then used – has been in widespread use for millennia, even today there is no understanding of the mechanisms involved (other than the persistent suspicion that Hebbian synaptic modification might somehow be involved) or of the nature of knowledge. Confabulation theory (see Figs. 1.4 and 1.5) specifies precisely what cognitive knowledge is, how it is acquired, how it is stored, and how it is used in thinking (Sect. 1.3.3).

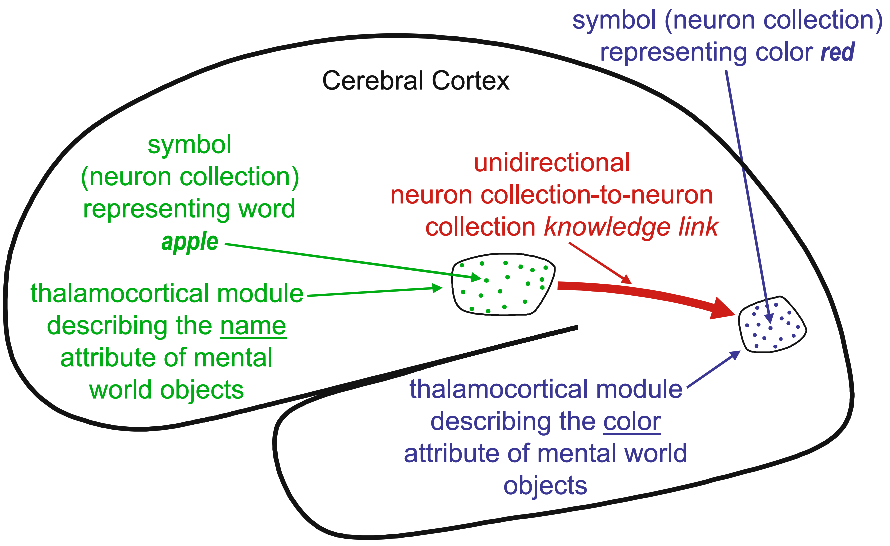


Fig. 1.4. A cognitive *knowledge link*. Here, a human subject is viewing and considering a red apple. A visual thalamocortical module is expressing a symbol for the color of the apple. At the same time, a language thalamocortical module is expressing a symbol for the name of the apple. Pairs of symbols which *meaningfully co-occur* in this manner have unidirectional axonal links, termed *knowledge links* (each considered a single *item of knowledge*), established between them via synaptic strengthening (assuming that the required axons are actually present – this is determined by genetics). The average adult human has billions of knowledge links, most of which are established in childhood. The rate of human knowledge acquisition often exceeds one link per second of life

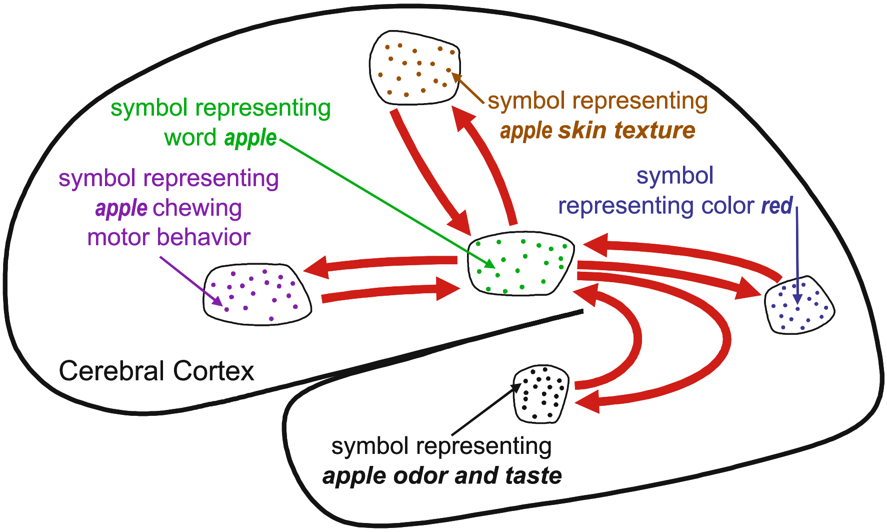

symbol representing word *apple*
symbol representing *apple skin texture*
symbol representing *apple* chewing motor behavior
symbol representing color *red*
Cerebral Cortex
symbol representing *apple odor and taste*

Fig. 1.5. Billions of pairs of symbols are connected via knowledge links. The set of all knowledge links joining symbols belonging to one specific *source* module to symbols belonging to one specific *target* module is termed a *knowledge base*. In the human brain, knowledge bases take the form of huge bundles of axons termed *fascicles*, which together make up a large portion of each cerebral hemisphere's ipsilateral white matter. Each module also typically has a knowledge base to its contralateral "twin" module (and perhaps to a few others near its twin) – which together constitute the *corpus callosum* fascicle linking the two cerebral hemispheres. Here, reciprocal knowledge links (red arrows), only some of which are shown, connect each expressed symbol representing an attribute of an apple pairwise with other such symbols. When an apple is currently present in the mental world, it *is* its collection of knowledge-link-connected symbols which are currently being expressed. There is no binding problem because all of these symbols are mutually "bound" by their previously established pairwise knowledge links. Shockingly, confabulation theory contends that such knowledge links – formed exclusively on the basis of meaningful symbol pair co-occurrence – are the only type of knowledge used (or needed) in cognition! Knowledge links are the second of the four key elements of confabulation theory

1.3.3 Confabulation Theory Key Element #3: Confabulation – The Information-Processing Operation of Thought

The vague notion that cognition employs some sort of "information-processing" has been around for millennia. Today, the understanding of the exact nature of this "cognitive information-processing" is roughly the same as it was in 350 B.C. – the time of Aristotle (arguably the first neuroscientist). Confabulation theory states explicitly and exactly that cognition involves only one information-processing operation – *confabulation* (see Fig. 1.6): a simple winners-take-all

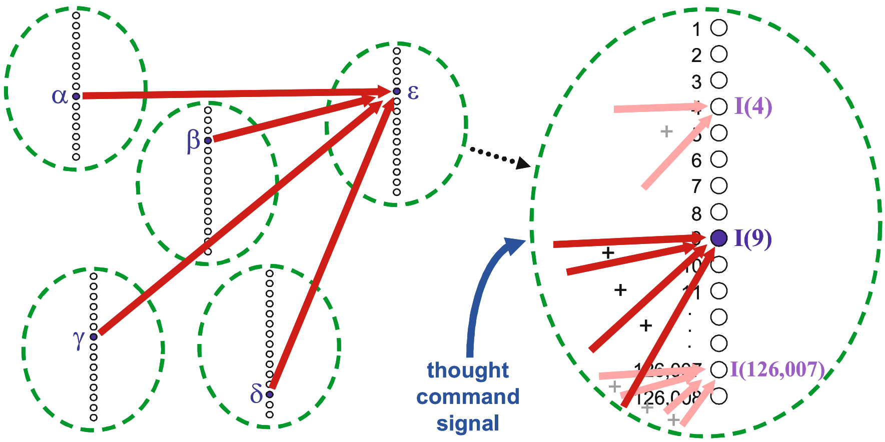


Fig. 1.6. *Confabulation* – the only information-processing operation used in cognition. Here, a concrete example involving five thalamocortical modules is shown (for simplicity, each module is illustrated as a dashed green oval with a list of that module's symbols inside it). See text for details. Confabulation is the third of the four key elements of confabulation theory

competition between symbols on the basis of the total input excitation they are receiving from knowledge links.

As seen in Fig. 1.6, the four modules on the left are each describing the attributes of one or more mental world objects by each expressing a single symbol: α, β, γ, and δ. Each of these four expressed symbols has a large number of knowledge links connecting it with symbols of the fifth module (of which four knowledge links, linking each expressed symbol to symbol ε of the fifth module, are shown). The situation within this fifth module, which is about to undergo *confabulation*, is shown enlarged on the right. For illustration, symbol 4 of this module is receiving two knowledge links (one from symbol α, and one from symbol γ), whereas symbols 9 and 126,007 are receiving knowledge links from all of α, β, γ, and δ. Each knowledge link is delivering a certain quantity of *input excitation* to the neurons of its target symbol.

The input excitations arriving at symbol k from different knowledge links are summed to yield the *total input excitation for symbol k*: I(k) (this summation is noted by the plus signs between the knowledge links in the enlarged illustration of module five). [As discussed extensively in this book, this *additive knowledge combination* property is one of the paramount reasons for the enormous information-processing power and flexibility of thought.].

Upon being commanded to do so (by a deliberate externally supplied *thought-command signal* – analogous to the motorneuron input to a muscle – illustrated by a blue arrow in Fig. 1.6), the symbols of the fifth module compete with one another (via a highly parallel, fast, *neuronal attractor network* function), yielding a final state in which all of the neurons representing the symbol with the largest input intensity I (in this example, symbol 9) are highly activated and all other symbol-representing neurons are not. This "winners-take-all" information-processing operation is called *confabulation*, and the winning symbol is termed the *conclusion*.

Confabulation is hypothesized to be the only information-processing operation involved in thought. In the Fig. 1.6 example, there is only one confabulation taking place. Ordinarily, confabulations on multiple modules take place together, with convergence to the winning symbol slowed somewhat to allow mutual interaction during convergence ("comparing notes" in order to arrive at a *confabulation consensus* of final conclusions). In such a *multiconfabulation*, often millions of items of knowledge, each emanating from a viable candidate conclusion, are employed in parallel in a "swirling" convergence process. (As discussed extensively in this book, this is another paramount reason for the enormous information-processing power and flexibility of thought.) Confabulation is the third of the four key elements of confabulation theory.

Confabulation is starkly alien in comparison with existing concepts in neuroscience, computational intelligence, neural networks, computer science, AI, and philosophy in general. For example, computer CPUs all follow the Turing paradigm: when commanded via a specific, digital, instruction code they execute a pre-defined logical or arithmetic instruction on specified variables. Thalamocortical modules, on the other hand, have only one information-processing "instruction" – confabulation. Further, the command to confabulate (termed the *thought-control command* – which is delivered to the confabulating module from outside cerebral cortex and thalamus) is not digital; rather, it is *analog*. Yet the result of a completed confabulation is digital: a single symbol. Very weird.

The ultimate challenge is to show that it is possible to explain Newton, Mozart, Einstein, and Crick using confabulation. That will probably take a while. Yet, the evidence presented in this book is intended to build confidence that this challenge will someday be met.

1.3.4 Confabulation Theory Key Element #4: The Conclusion → Action Principle – The Origin of Behavior

One of the most obvious aspects of brain function (and therefore one of the most consistently ignored) is that animals typically launch many behaviors every second they are awake. Most of these are *microbehaviors* (small corrective modifications to ongoing behaviors), but, typically, many times per hour major new behaviors are launched, predicated on newly emerged events. Beyond simple reflexes (e.g., knee jerk) and autonomic reactions (e.g., digestion), no understanding of how and why behaviors originate currently exists.

Confabulation theory proposes the *conclusion → action principle* (Fig. 1.7), which states that every time a confabulation operation on a thalamocortical module reaches a conclusion, an associated set of *action commands* are launched from the cortical patch of the module via axons which proceed towards sub-cortical

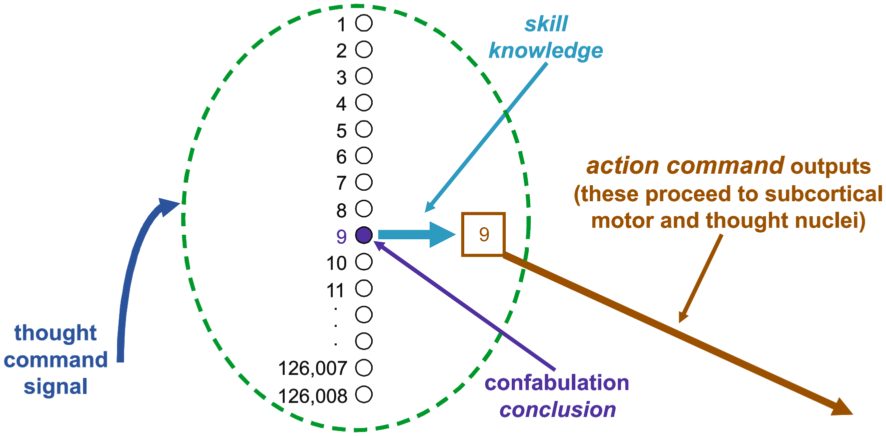


Fig. 1.7. The *conclusion → action principle*: hypothesized to be the origin of all non-reflexive and non-autonomic behavior. Here, a thalamocortical module (illustrated, in consonance with Fig. 1.6, as an abstract "oval" structure containing a list of the module's symbols) has successfully completed a confabulation operation (under control of its externally supplied thought-command signal) and reached a conclusion (symbol number 9 as in Fig. 1.6). Whenever a module completes a confabulation and reaches a conclusion it immediately causes a set of *action command* outputs to be launched (these outputs proceeding to sub-cortical nuclei). The specific action command outputs that are launched are those which have been previously *associated from* this specific conclusion symbol via a completely separate, sub-cortically managed, *skill-learning* process. These action command outputs can cause behaviors to occur. The conclusion → action principle is the fourth and last of the key elements of confabulation theory

structures. Often, these action commands lead to the initiation of *behaviors* (either immediately or after further evaluation). All non-reflexive and non-autonomic behavior arises in this manner.

Action commands can be regarded as *suggested behaviors* – which sub-cortical structures either immediately execute, consider further for future execution, or (e.g., if the suggested behavior is not consistent with past successful reductions in currently elevated goal or drive states) discard.

The *associations* between each symbol of a module and the specific action commands which are to be issued when that symbol wins a confabulation competition are termed *skill knowledge*. Skill knowledge is formed via selective strengthening of special synapses within cerebral cortex; but the involved *skill-learning* process is controlled by sub-cortical structures.

Skill knowledge, although implemented by synapses in cortex, differs greatly in neuroanatomical location and physiological properties from cognitive knowledge links. For example, unlike a cognitive knowledge link (which, if solidified over the 100 hours following the initial symbol pair co-occurrence, is extremely durable), skill knowledge is often fragile and short-lived (this is important for *rehearsal* *learning* of skills, where later, more competent skill knowledge needs to "supersede" and supplant earlier, less perfected skill knowledge).

Behavioral triggering, skill knowledge, and skill learning are not parts of thinking (they come into play only after thinking has completed its job of reaching conclusions) and so they are not discussed much in this book. Of course, this decision is subject to the criticism that thinking itself is utterly dependent upon the *thought-command sequences* which control the operation of the thalamocortical modules involved in a particular thought process. These thought-command sequences are learned, stored, and recalled in exactly the same manner (using knowledge links) as the movement command sequences (actually, *postural goal* sequences) employed in movement. So, thought begets movement and thought (both termed *actions*) in an endless action – thought – action – thought – action – thought – … sequence during wakefulness (thereby exorcising the homunculus hiding behind a curtain pulling the control levers of the brain and body). Actually there is quite a bit that could be said about all this; but this topic is deferred to a future edition. In this book, the focus is on the basic mechanism of thought.

1.4 Cognitive Brain "Hardware" and "Software"

The four key functional elements of confabulation theory described in Sect. 1.3 constitute the "information-processing hardware" upon which confabulation theory contends thinking is implemented. But what about the "software" of thought (the procedures, called *thought processes*, for using the hardware)?

A central hypothesis of confabulation theory is that thinking is a phylogenetic outgrowth of movement. Animals began moving over a billion years ago. The mechanisms for flexible, adaptive control of movement emerged early and expanded rapidly. As animal movement complexity and capability grew, a new design possibility emerged: the elaborate machinery already developed for controlling movement could be applied to brain tissue. In particular, discrete brain structures, *modules*, emerged that could be controlled exactly like individual muscles. By manipulating these modules in properly coordinated "movements" (thought processes), information-processing could be carried out – thereby further enhancing competitive success.

As discussed in Sect. 1.3.3, each human thalamocortical module has a single thought-command input signal that tells it when to "contract." This is analogous to the roughly 700 muscles of the human body, each of which has a single input signal (motorneuron input) that commands it to contract. Just as with a muscle, the thought-command input to a module is an *analog* signal: it can range from a low level ("contract a little") to a highest level ("contract with maximum force"); where "contraction" corresponds roughly to the rate of convergence, from multiple candidate conclusions to a single conclusion, of a module's confabulation competition.

In effect, the human brain thinks by maneuvering subsets of 4,000 digital processors (the thalamocortical modules) through smooth, graceful, thought maneuvers. These thought processes are learned, stored, and recalled just as movement processes are learned, stored, and recalled. At higher hierarchical levels, closely related movement processes and thought processes are often stored mixed together in the same knowledge links.

Just as the repertoire of human movement can be vast (walking, writing, running, cartwheels, uneven parallel bar routines, pole vaulting, etc.), so the repertoire of thought can contain a vast variety of different ways of using thalamocortical modules. However, at the present time, confabulation theory has only identified a few of these ways. And only two of these, a single isolated confabulation (crudely analogous to flexing of a single muscle) and *multiconfabulation swirling* (crudely analogous to walking – the most basic and useful of human movements), have received significant study. All of the remaining chapters of this book discuss these two types of basic thought process.

As is discussed in detail in the video presentation, brains carry out a multitude of functions in addition to cognition. Quite a few of these interact intimately with, and are required to implement, thought processes. However, these other brain functions are poorly understood and are only briefly mentioned in this book. The thought processes considered here (single confabulations and multiconfabulations) are implemented using an *external thought controller* executing a crude, contrived thought process. The only feedback that a *thought controller* gets from the thought process being executed on the involved collection of modules is knowledge of when a module has reached a conclusion (in effect, an action command output, as in Sect. 1.3.4). This feedback can be used to trigger recall and playback of a different "canned" thought process. While this approach only implements a tiny subset of the capabilities of real brain thought and movement control, as the reader will see, it is still possible to achieve interesting results.

1.5 Implications of Confabulation Theory

Confabulation theory has a variety of implications. A few examples are discussed here.

Since all of cognition is "categorical" (i.e., based upon the symbol sets of the thalamocortical modules), the total number of modules, and the number of symbols in each of those modules, provides a reasonable estimate for the "descriptive power" of a brain. A trout may have only a few tens of modules, each with a few hundred symbols. A raven might have hundreds of modules, each with many hundreds of symbols. A human probably has thousands of modules, each with thousands to hundreds of thousands of symbols. Similarly, the total number of knowledge links that an animal possesses gives a crude quantification of how "smart" that animal is (although, clearly, the distribution of those knowledge links also matters: idiot savants may have huge numbers of knowledge links).

The experiments of Chap. 6 imply that the average human possesses billions of items of knowledge, of which the majority are often obtained in childhood. Some humans may possess tens, or perhaps even hundreds, of billions of items of knowledge. Clearly, since there are only about 32 million seconds in a year, the average rate of knowledge acquisition often exceeds one item per second and might sometimes exceed 100 items per second. It is thus not surprising that we need to sleep a third of the time in order to catch up with evaluating and selectively solidifying each day's new cognitive knowledge links (i.e., implement cognitive learning control decision-making for recently established, and intrinsically rapidly fading, temporary knowledge links – which is probably the main activity of sleep).

Humans (and animals in general) are almost certainly much "smarter" than has been generally appreciated. Assuming such findings are confirmed, fields as diverse as psychology, education, philosophy, psychiatry, medicine (both human and veterinary), law, and theology will need to be extensively overhauled.

With one relatively small exception, the axonal connectivity between the thalamocortical modules in the human brain seems to roughly resemble that of other great apes. That one exception is the modules of the human language faculty – which seem to connect widely to modules in many other faculties. In this sense, language is the *hub* of human cognition. It seems likely that this (along with having a brain which is, overall, over three times larger) can explain some of the commanding power of human thought in comparison with that of other apes. As we learn more about cetaceans, it may well be that some of them (and perhaps other species as well, such as jays, ravens, and parrots) also have this *language hub cognitive architecture* characteristic to some degree.

The near-term implications of confabulation theory for neuroscience are uncertain. Neuroscience is dominated by bottom-up thinking and by "methods." To succeed, neuroscientists must often spend the decade after completing their Ph.D. developing their own effective experimental methods. The subset of aspirants who successfully complete this process must then, in general, inaugurate and manage a large lab that quickly acquires enormous built-in inertia. After completing this arduous initiation at about age 40, few of these newly established neuroscientists are going to be interested in abandoning, or significantly altering, their research direction in order to begin to follow up on the hypotheses of confabulation theory. Thus, integration of confabulation theory into neuroscience is likely to be largely confined to new investigators who decide to pursue experimental exploration of confabulation theory's neuroscience implications (probably mainly using human subjects carrying out controlled thought processes while being monitored by brain activity scanners with greatly improved spatial and temporal resolution). Assuming this established social pattern continues to hold, it seems unlikely that confabulation neuroscience can join the mainstream of the subject until the next decade.

Notwithstanding the above, members of the small community of mathematical neuroscientists may soon realize that, given the hard constraints provided by confabulation theory, it may be possible to tackle large-scale understanding of brain function. For example, it may be possible within a few years to build an integrated functional mathematical model of cerebral cortex, thalamus, basal ganglia, subthalamus, red nucleus, substantia nigra, hippocampus, amygdala, hypothalamus, spinal cord, locus coeruleus, pons, and cerebellum. This model may well answer most of the large questions of neuroscience that remain after confabulation theory.

A large-scale human brain modeling project of this sort will surely require a widely knowledgeable and exceptionally well educated team of hundreds of mathematical neurobiologists and computer scientists operating as willing and compliant subordinates under the hierarchical command of a master genius. The usual "herd of cats" sort of scientific research program would probably not work effectively in this instance. I personally know at least five people who could each probably successfully lead such an effort. Such an integrated brain modeling project is, in my opinion, one of the most important tasks that the human species should now carry out. It will be expensive (probably exceeding $200,000,000 per year for a decade; along with another $400,000,000 for a proper building to house the project and the budget for the required equipment). A single, open, international project of this type would seem ideal. However, given the potential economic and national security implications, multiple projects of this type seem more likely. With respect to these practical implications of confabulation theory, I leave it to you, the reader, to form your own opinion as you absorb the book's content.

1.6 Content of the Book

The content of the eight chapters and two DVDs of this book is briefly surveyed below:

- **Chapter 1: Introduction**
  An introductory overview of confabulation theory: comments on some of the theory's possible implications and presentation of this overview of the book's contents.

- **Chapter 2: Video Presentation Viewcells**
  The viewcells used in the book's DVD video presentation are presented. To help with understanding and retention of the material, each of these should be referred to while it is being presented during the video.

- **Chapter 3: The Mathematics of Cognition**
  An introduction to the mathematics of confabulation theory. Comments on the relationship between cogency maximization and Bayesian analysis. An extensive discussion of the status of confabulation neuroscience. Comments on the origins of confabulation theory.

This chapter is based on the original publication

> Hecht-Nielsen R (2006) The mathematics of thought. In: Yen GY, Fogel DB (eds) Computational intelligence: Principles and practice. IEEE Computational Intelligence Society, Piscataway, NJ, pp 1–16

and is adapted here in accordance with IEEE copyrights.

- **Chapter 4: Cogent Confabulation**
  Mathematical foundations of confabulation theory are presented, including statement and proof of the Fundamental Theorem of Cognition and the theorem showing that cogent confabulation within a logical information environment yields Aristotelian logic. Computer experiments with a single confabulation are presented, with all details provided. Replication of these single confabulation experiments is the logical starting point for those wanting to gain hands-on experience with confabulation architectures.
  This chapter is a reformatted reprint of the original publication

  > Hecht-Nielsen R (2005) Cogent confabulation. Neural Networks 18:111–115, Copyright (2005)

  used with permission from Elsevier.

- **Chapter 5: Confabulation Neuroscience I**
  A concise overview of confabulation neuroscience. This material is prerequisite for Chap. 6.
  This chapter is based on the original publication

  > Hecht-Nielsen R (2006) The mechanism of thought. In: Proceedings of the World Congress on Computational Intelligence. 16–21 July, Vancouver, BC, Canada. IEEE Press, Piscataway, NJ

  and is adapted here in accordance with IEEE copyrights.

- **Chapter 6: The Mechanism of Thought**
  Computer experiments with multiconfabulation are presented, with all details. These *sentence continuation* experiments illustrate that thinking is exactly like moving. Replication of these multiconfabulation experiments is the second logical step for those wishing to gain hands-on experience with confabulation architectures.

- **Chapter 7: Mechanization of Confabulation**
  Further details of confabulation architecture design and implementation are presented. Approaches for application of confabulation architectures to language, vision, and hearing are discussed in some detail.
  This chapter is based on the original publication

  > Hecht-Nielsen R (2006) The mechanization of cognition. In: Bar-Cohen Y (ed) Biomimetics. CRC Press, Boca Raton, FL, pp 57–128

  and is adapted here from the original with kind permission of the publisher.

- **Chapter 8: Confabulation Neuroscience II**
  An expanded discussion of confabulation neuroscience.
  This chapter is based on the Appendix of the original publication

  > Hecht-Nielsen R (2006) The mechanization of cognition. In: Bar-Cohen Y (ed.) Biomimetics. CRC Press, Boca Raton, FL, pp 57–128

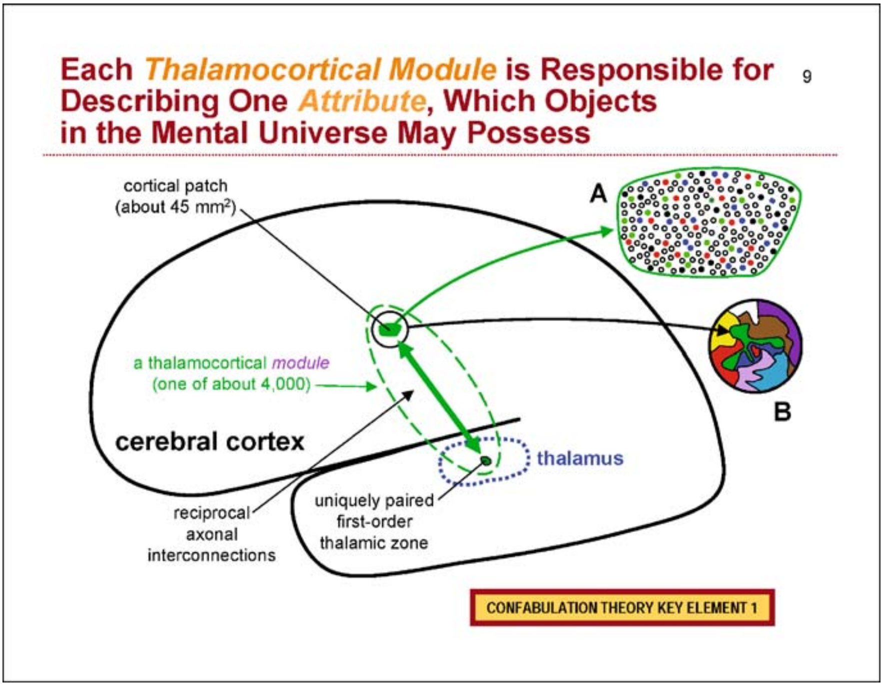
  and is adapted here from the original with kind permission of the publisher.

- **DVDs**
  The book's two DVDs (attached to this book) contain the following material:

  1. *The Mechanism of Thought* video presentation (Part I on DVD Disk 1 and Part II on DVD Disk 2).
  2. PDF file of the *Viewcells* used in *The Mechanism of Thought* video presentation. This computer-readable file is included on both Disk 1 and Disk 2.
  3. PDF file of the *Presentation Notes* for *The Mechanism of Thought* video presentation. This computer-readable file is included on both Disk 1 and Disk 2. [Note: These *Presentation Notes*, intended for use as courseware, are probably the most important component of the book.]

2 Video Presentation Viewcells

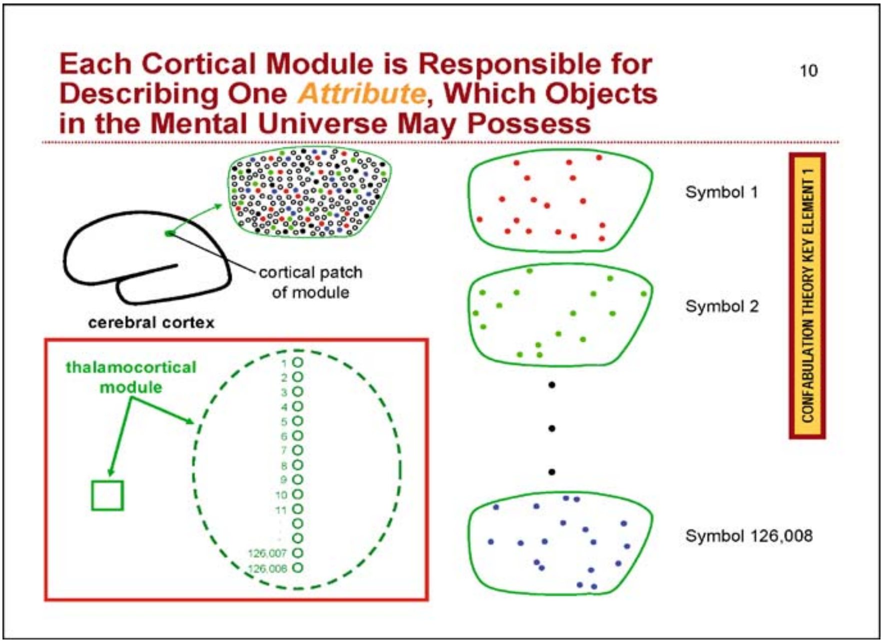
The viewcells used in the book's DVD video presentation are presented. To help with understanding and retention of the material, each of these should be referred to while it is being presented during the video. In this chapter we start with viewcell 9, while viewcells 1 through 110 are shown in the DVD presentation.



Each *Thalamocortical Module* is Responsible for Describing One *Attribute*, Which Objects in the Mental Universe May Possess

9

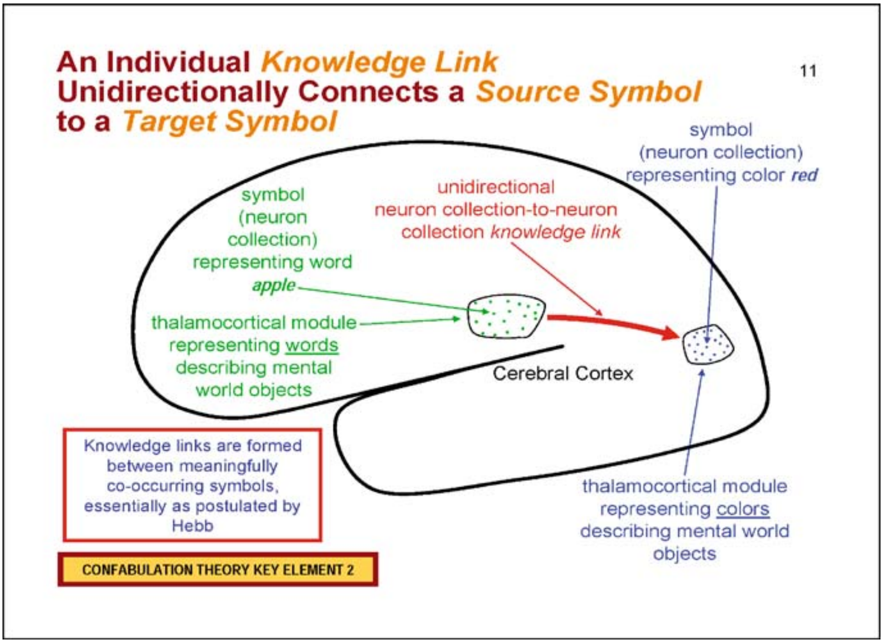
CONFABULATION THEORY KEY ELEMENT 1



Each Cortical Module is Responsible for Describing One *Attribute*, Which Objects in the Mental Universe May Possess

10

CONFABULATION THEORY KEY ELEMENT 1



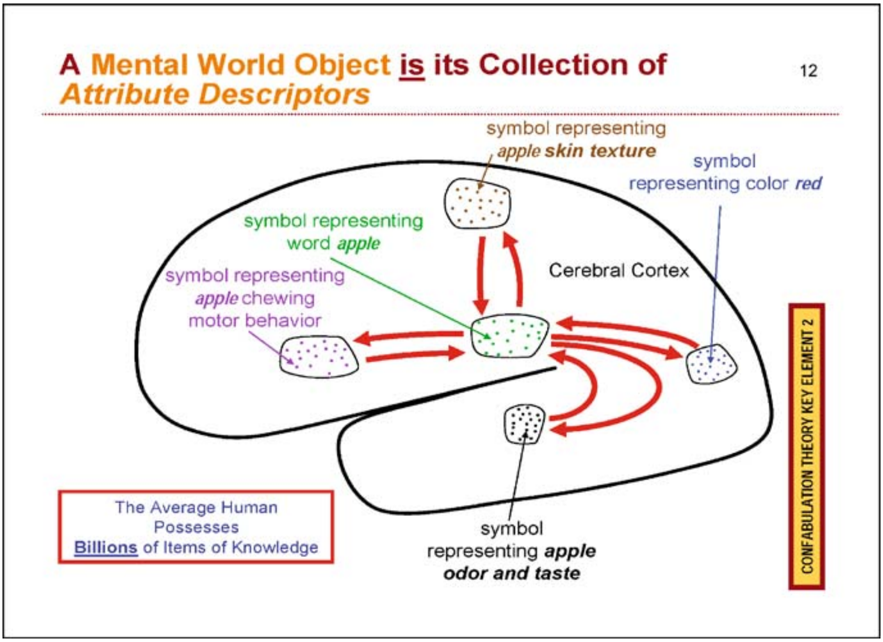
An Individual *Knowledge Link* Unidirectionally Connects a *Source Symbol* to a *Target Symbol*

11

symbol (neuron collection) representing color *red*
symbol (neuron collection) representing word *apple*
unidirectional neuron collection-to-neuron collection *knowledge link*
thalamocortical module representing words describing mental world objects
Cerebral Cortex
thalamocortical module representing colors describing mental world objects

Knowledge links are formed between meaningfully co-occurring symbols, essentially as postulated by Hebb

CONFABULATION THEORY KEY ELEMENT 2



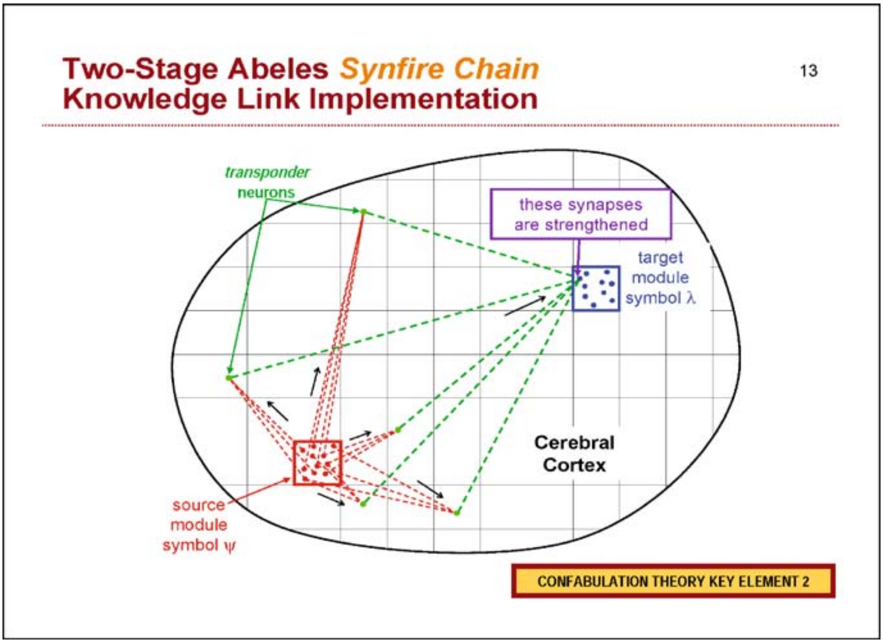
A Mental World Object is its Collection of *Attribute Descriptors*

12

symbol representing *apple skin texture*
symbol representing color *red*
symbol representing word *apple*
symbol representing *apple* chewing motor behavior
Cerebral Cortex
symbol representing *apple odor and taste*

The Average Human Possesses Billions of Items of Knowledge

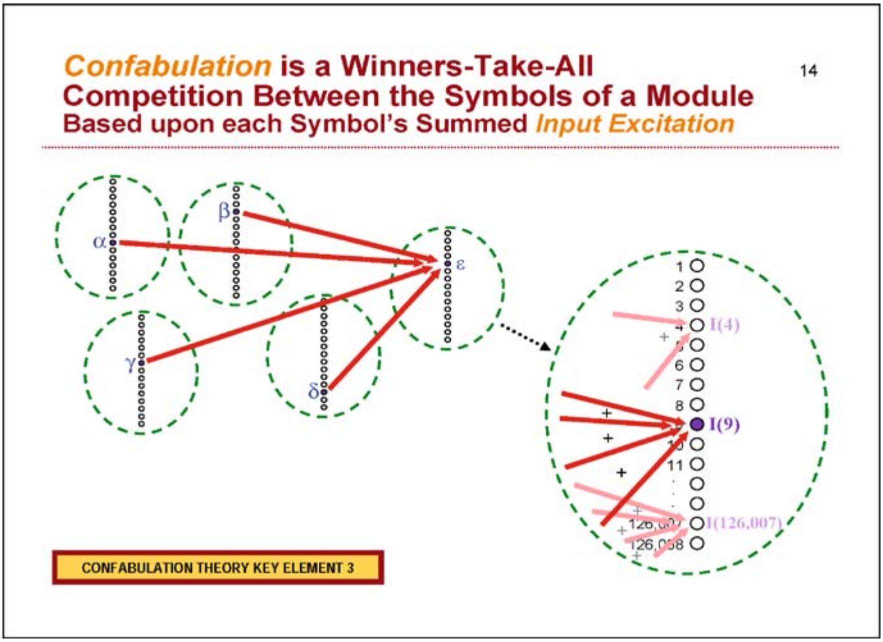
CONFABULATION THEORY KEY ELEMENT 2



Two-Stage Abeles *Synfire Chain* Knowledge Link Implementation

13

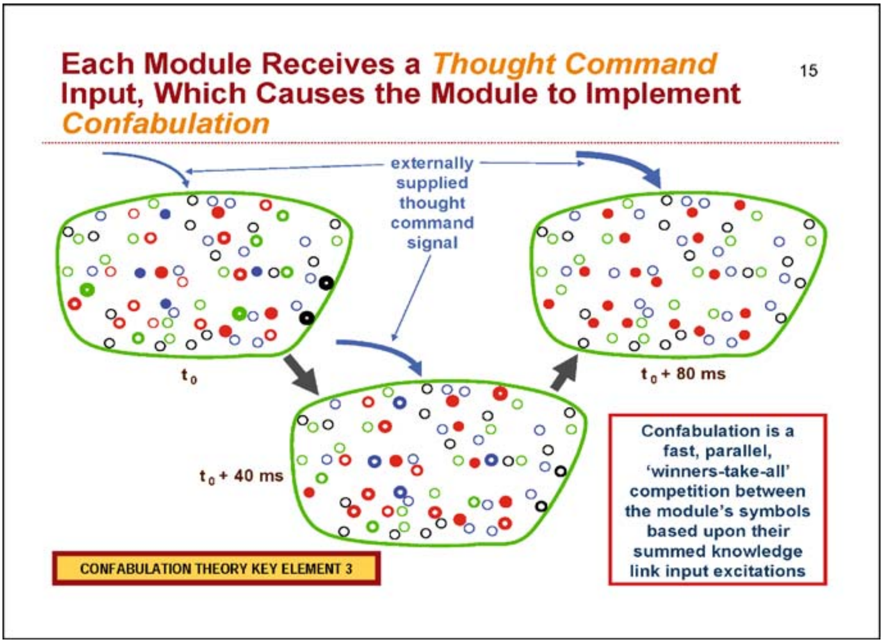
CONFABULATION THEORY KEY ELEMENT 2



Confabulation is a Winners-Take-All Competition Between the Symbols of a Module Based upon each Symbol's Summed *Input Excitation*

14

CONFABULATION THEORY KEY ELEMENT 3



Each Module Receives a *Thought Command* Input, Which Causes the Module to Implement *Confabulation*

15

externally supplied thought command signal

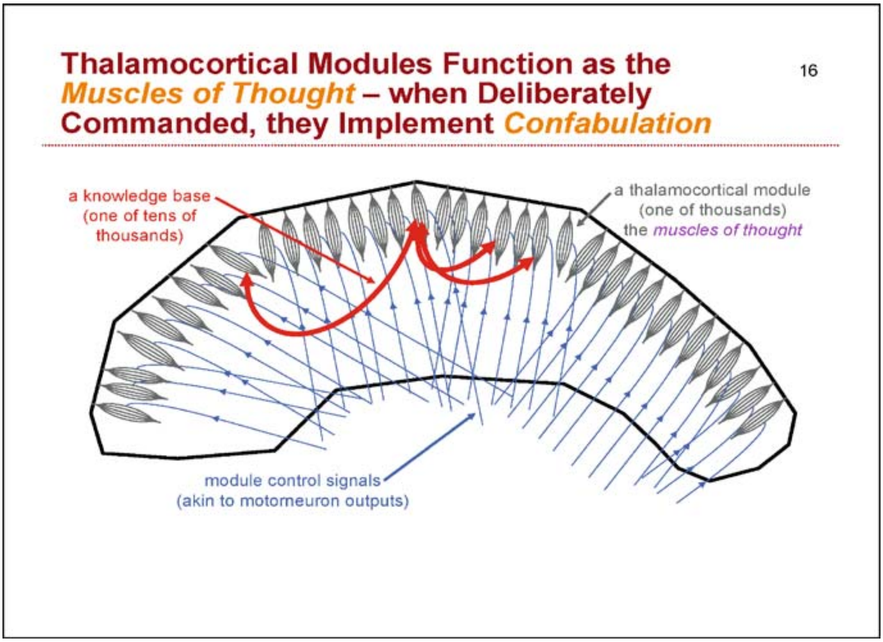
t_0

t_0 + 40 ms

t_0 + 80 ms

Confabulation is a fast, parallel, 'winners-take-all' competition between the module's symbols based upon their summed knowledge link input excitations

CONFABULATION THEORY KEY ELEMENT 3



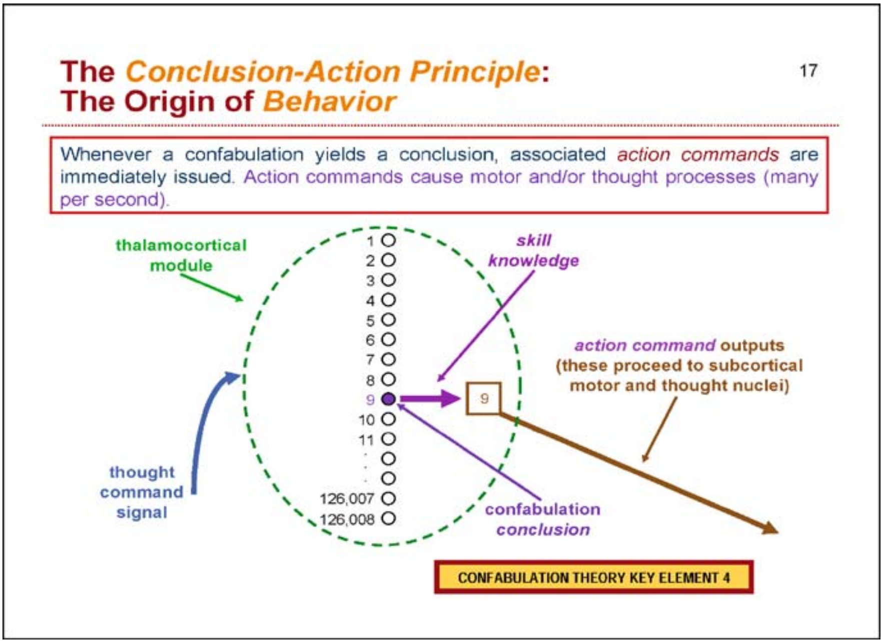
Thalamocortical Modules Function as the *Muscles of Thought* – when Deliberately Commanded, they Implement *Confabulation*

16

a knowledge base (one of tens of thousands)

a thalamocortical module (one of thousands) the *muscles of thought*

module control signals (akin to motorneuron outputs)

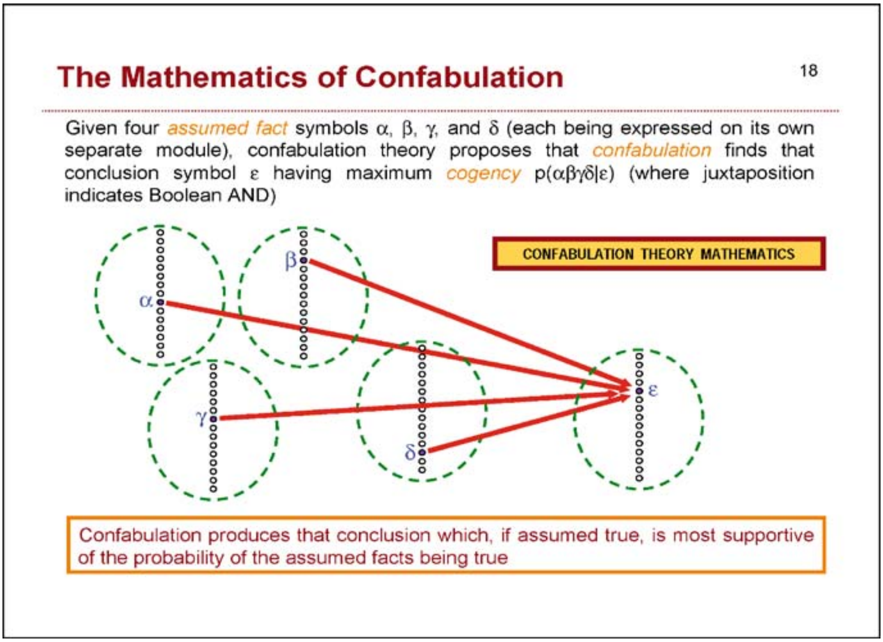


The *Conclusion-Action Principle*: The Origin of *Behavior*

17

Whenever a confabulation yields a conclusion, associated *action commands* are immediately issued. Action commands cause motor and/or thought processes (many per second).

CONFABULATION THEORY KEY ELEMENT 4



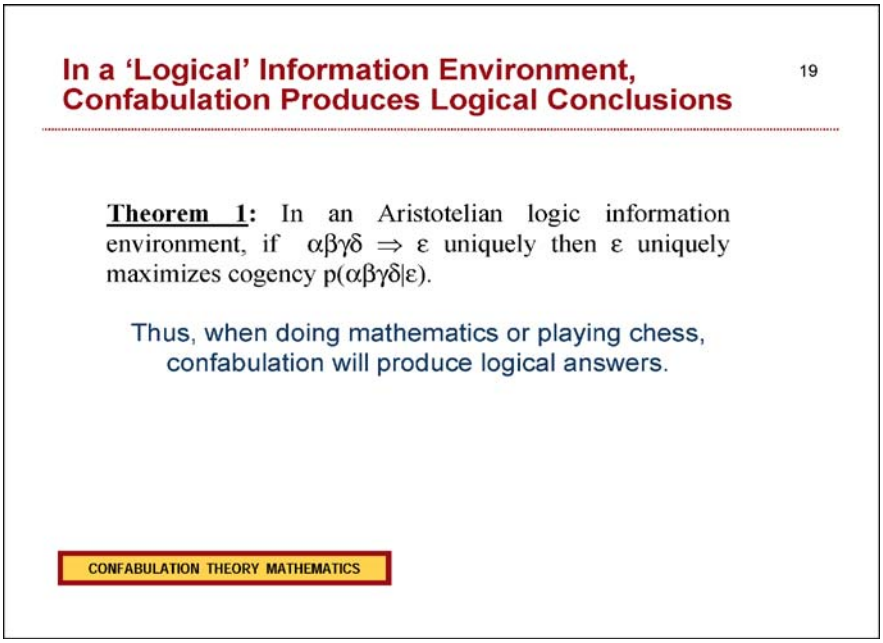
The Mathematics of Confabulation

18

Given four *assumed fact* symbols α, β, γ, and δ (each being expressed on its own separate module), confabulation theory proposes that *confabulation* finds that conclusion symbol ε having maximum *cogency* p(αβγδ|ε) (where juxtaposition indicates Boolean AND)

CONFABULATION THEORY MATHEMATICS

Confabulation produces that conclusion which, if assumed true, is most supportive of the probability of the assumed facts being true



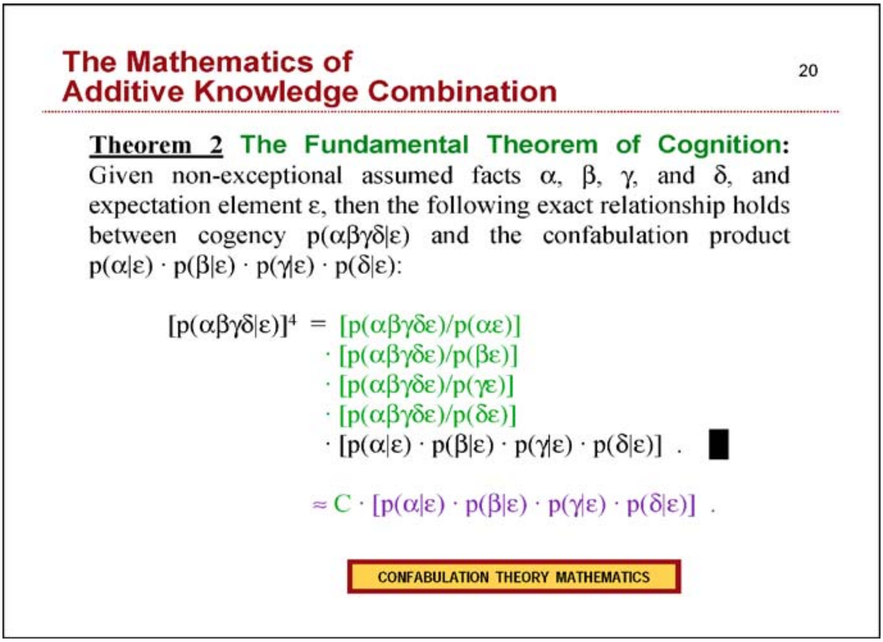
In a 'Logical' Information Environment, Confabulation Produces Logical Conclusions

19

Theorem 1: In an Aristotelian logic information environment, if $\alpha\beta\gamma\delta \Rightarrow \varepsilon$ uniquely then ε uniquely maximizes cogency $p(\alpha\beta\gamma\delta|\varepsilon)$.

Thus, when doing mathematics or playing chess, confabulation will produce logical answers.

CONFABULATION THEORY MATHEMATICS



The Mathematics of Additive Knowledge Combination

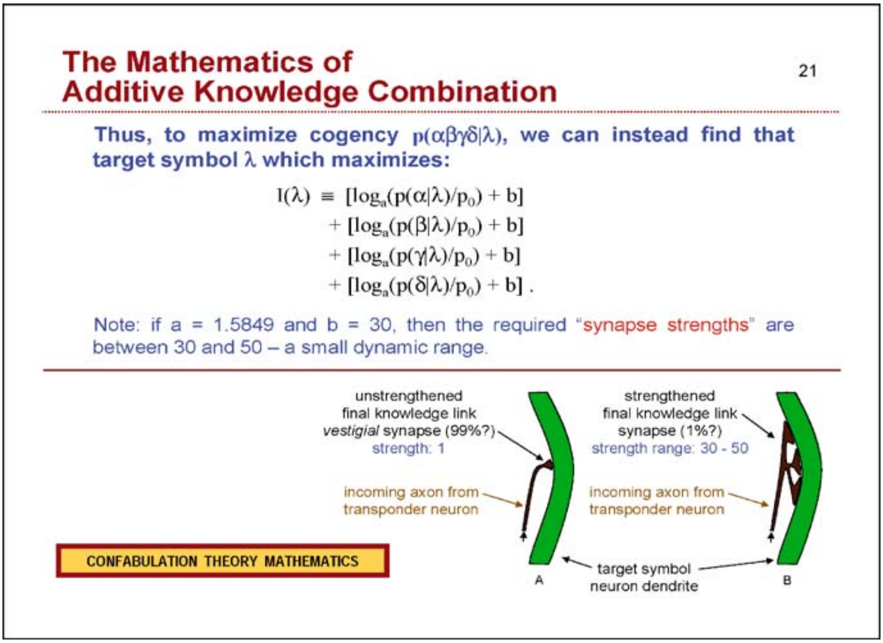
20

Theorem 2 The Fundamental Theorem of Cognition**:** Given non-exceptional assumed facts α, β, γ, and δ, and expectation element ε, then the following exact relationship holds between cogency $p(\alpha\beta\gamma\delta|\varepsilon)$ and the confabulation product $p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)$:

$$
\begin{aligned}
[p(\alpha\beta\gamma\delta|\varepsilon)]^4 = {} & [p(\alpha\beta\gamma\delta\varepsilon)/p(\alpha\varepsilon)] \\
& \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\beta\varepsilon)] \\
& \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\gamma\varepsilon)] \\
& \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\delta\varepsilon)] \\
& \cdot [p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)] \ . \quad \blacksquare
\end{aligned}
$$

$$\approx C \cdot [p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)] \ .$$

CONFABULATION THEORY MATHEMATICS



The Mathematics of Additive Knowledge Combination

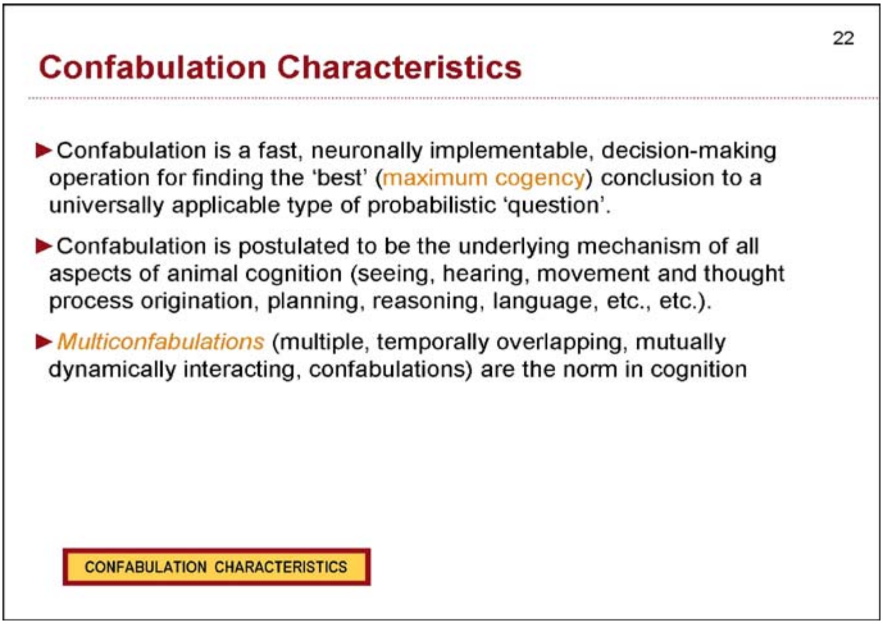
21

Thus, to maximize cogency $p(\alpha\beta\gamma\delta|\lambda)$, we can instead find that target symbol λ which maximizes:

$$\begin{aligned} I(\lambda) \equiv\ & [\log_a(p(\alpha|\lambda)/p_0) + b] \\ & + [\log_a(p(\beta|\lambda)/p_0) + b] \\ & + [\log_a(p(\gamma|\lambda)/p_0) + b] \\ & + [\log_a(p(\delta|\lambda)/p_0) + b]\ . \end{aligned}$$

Note: if a = 1.5849 and b = 30, then the required "synapse strengths" are between 30 and 50 – a small dynamic range.

CONFABULATION THEORY MATHEMATICS

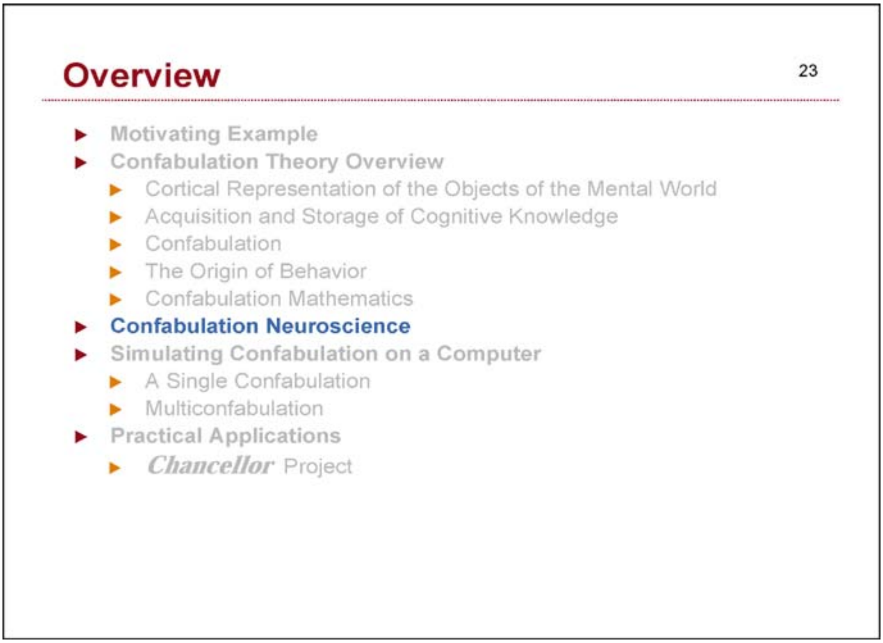


Confabulation Characteristics

22

- ►Confabulation is a fast, neuronally implementable, decision-making operation for finding the 'best' (maximum cogency) conclusion to a universally applicable type of probabilistic 'question'.
- ►Confabulation is postulated to be the underlying mechanism of all aspects of animal cognition (seeing, hearing, movement and thought process origination, planning, reasoning, language, etc., etc.).
- ►*Multiconfabulations* (multiple, temporally overlapping, mutually dynamically interacting, confabulations) are the norm in cognition

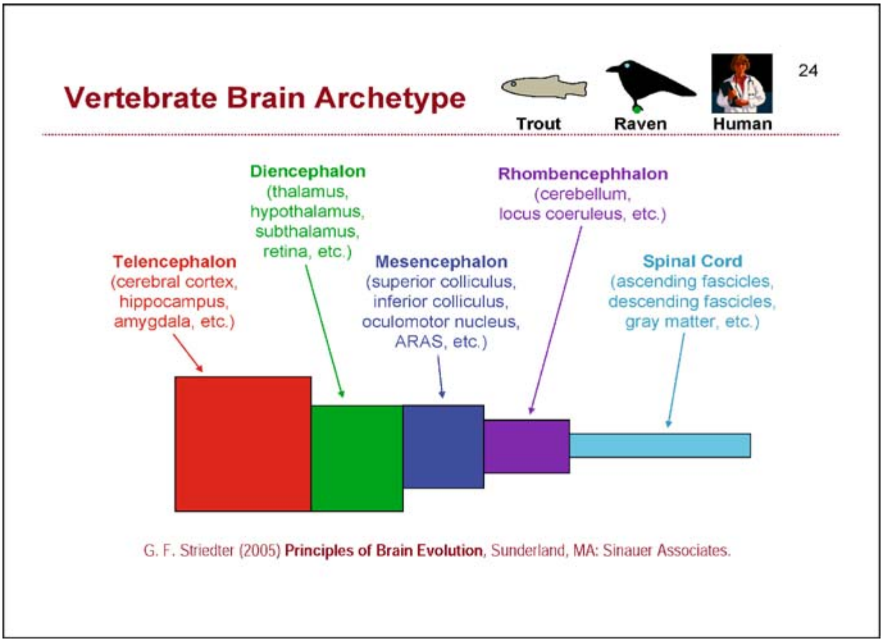
CONFABULATION CHARACTERISTICS



Overview

23

- Motivating Example
- Confabulation Theory Overview
  - Cortical Representation of the Objects of the Mental World
  - Acquisition and Storage of Cognitive Knowledge
  - Confabulation
  - The Origin of Behavior
  - Confabulation Mathematics
- Confabulation Neuroscience
- Simulating Confabulation on a Computer
  - A Single Confabulation
  - Multiconfabulation
- Practical Applications
  - *Chancellor* Project



Vertebrate Brain Archetype

24

Telencephalon (cerebral cortex, hippocampus, amygdala, etc.)

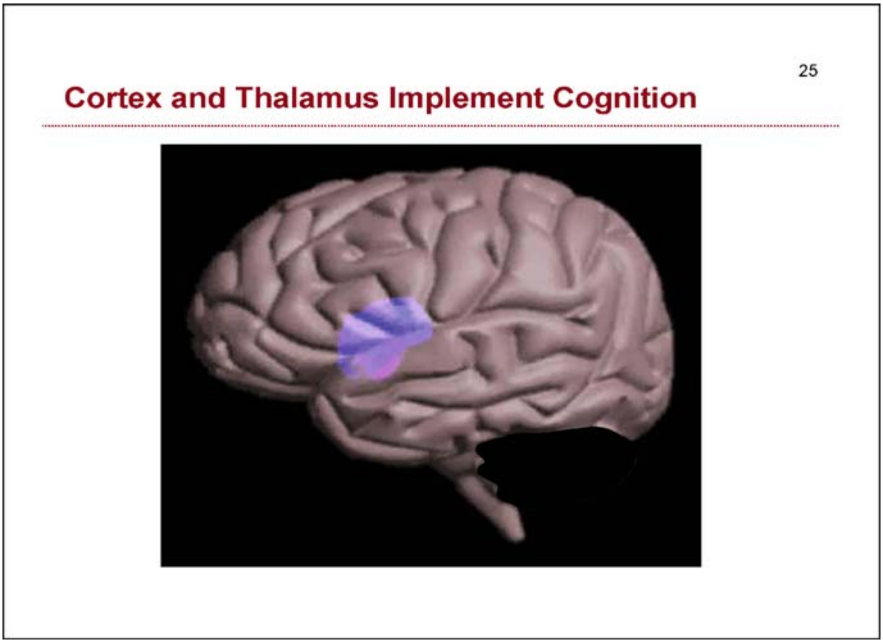
Diencephalon (thalamus, hypothalamus, subthalamus, retina, etc.)

Mesencephalon (superior colliculus, inferior colliculus, oculomotor nucleus, ARAS, etc.)

Rhombencephhalon (cerebellum, locus coeruleus, etc.)

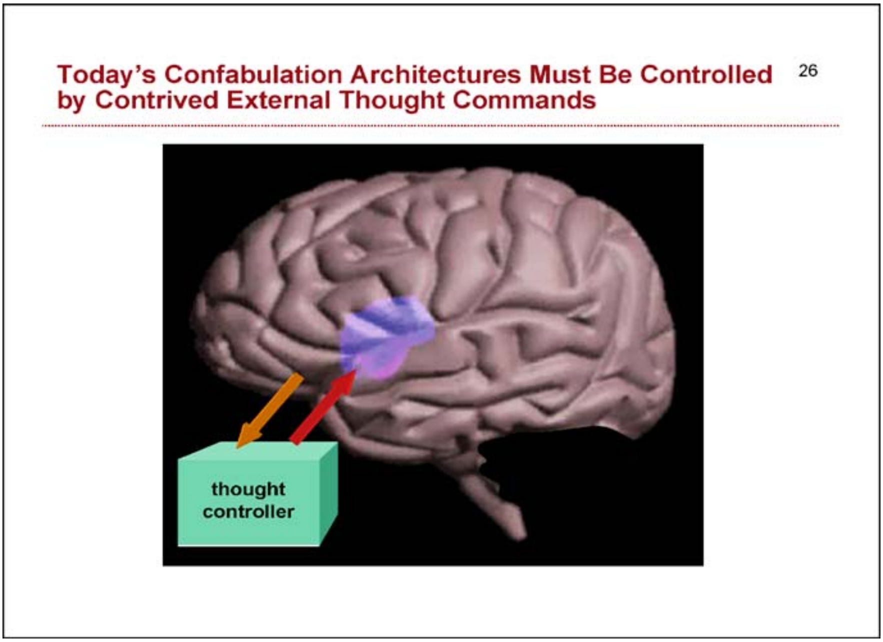
Spinal Cord (ascending fascicles, descending fascicles, gray matter, etc.)

G. F. Striedter (2005) **Principles of Brain Evolution**, Sunderland, MA: Sinauer Associates.



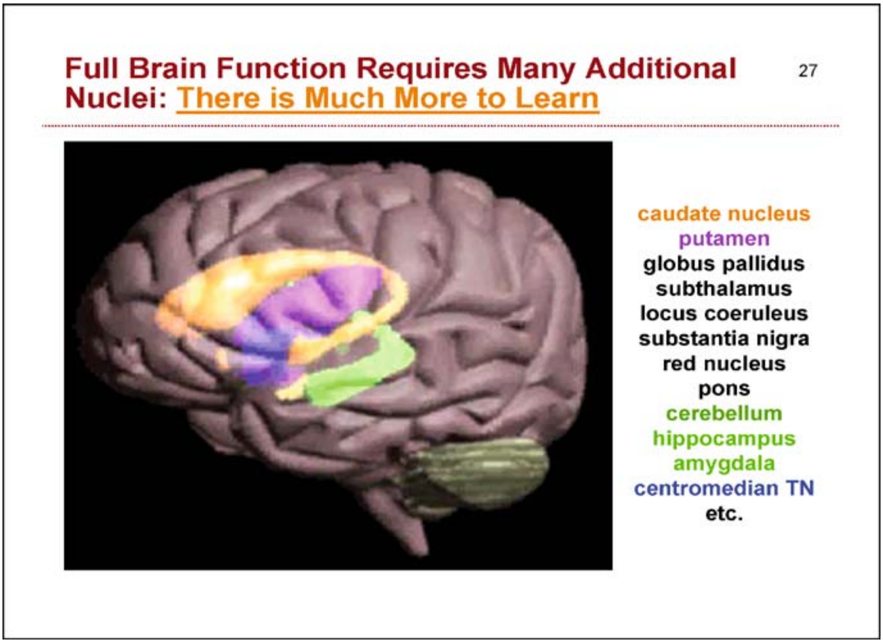
25

Cortex and Thalamus Implement Cognition



26

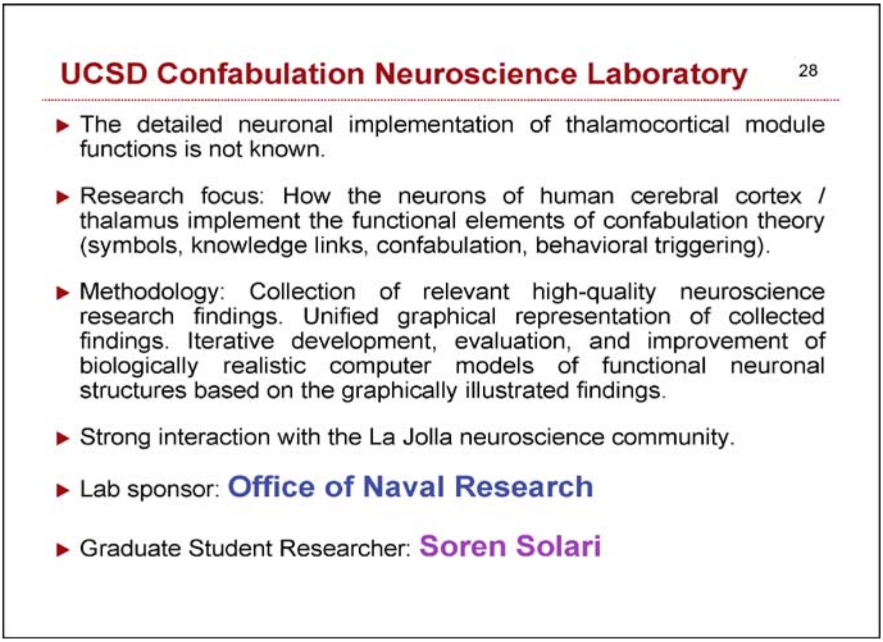
Today's Confabulation Architectures Must Be Controlled by Contrived External Thought Commands



Full Brain Function Requires Many Additional Nuclei: There is Much More to Learn

27

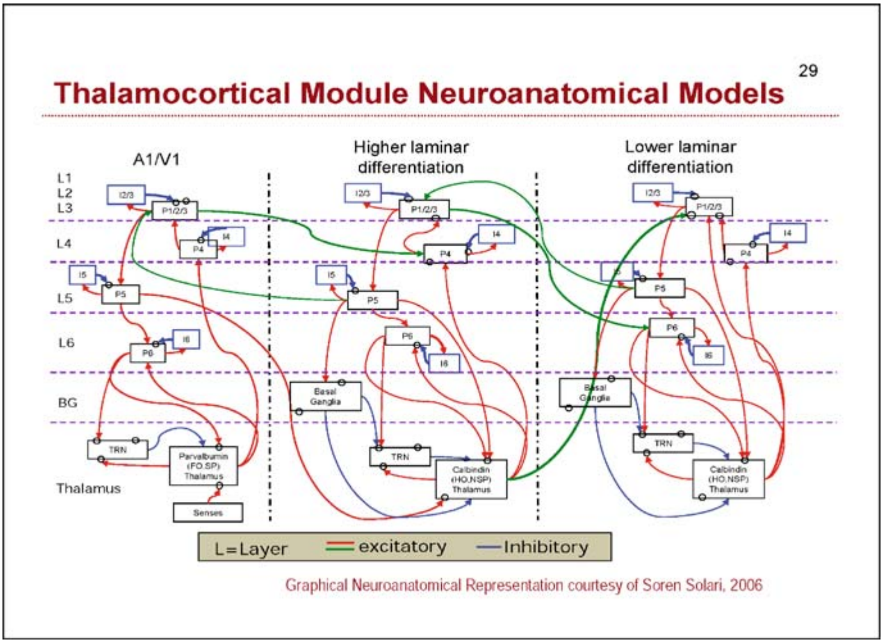
caudate nucleus
putamen
globus pallidus
subthalamus
locus coeruleus
substantia nigra
red nucleus
pons
cerebellum
hippocampus
amygdala
centromedian TN
etc.



UCSD Confabulation Neuroscience Laboratory

28

- The detailed neuronal implementation of thalamocortical module functions is not known.
- Research focus: How the neurons of human cerebral cortex / thalamus implement the functional elements of confabulation theory (symbols, knowledge links, confabulation, behavioral triggering).
- Methodology: Collection of relevant high-quality neuroscience research findings. Unified graphical representation of collected findings. Iterative development, evaluation, and improvement of biologically realistic computer models of functional neuronal structures based on the graphically illustrated findings.
- Strong interaction with the La Jolla neuroscience community.
- Lab sponsor: **Office of Naval Research**
- Graduate Student Researcher: **Soren Solari**

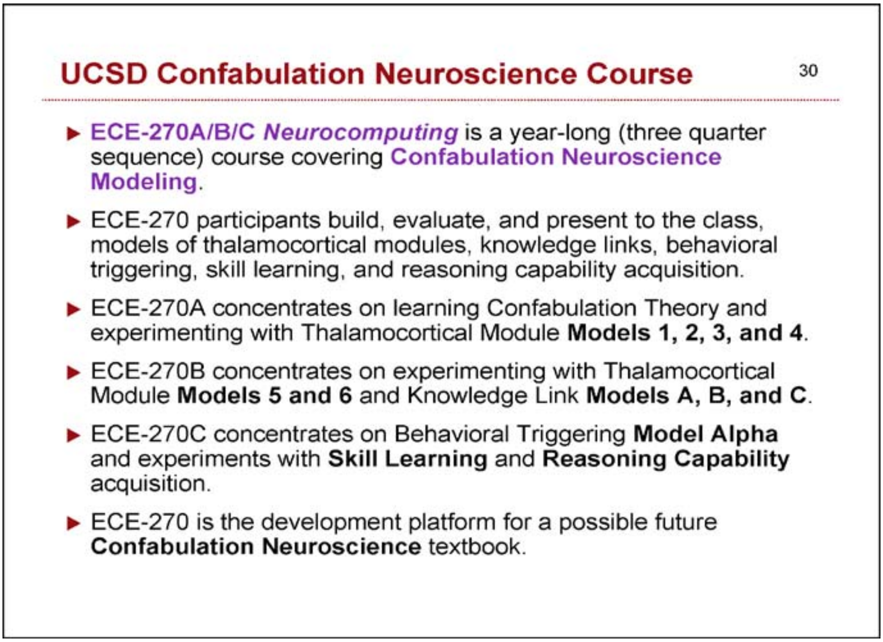


Thalamocortical Module Neuroanatomical Models

29

A1/V1
Higher laminar differentiation
Lower laminar differentiation
L1
L2
L3
L4
L5
L6
BG
Thalamus
I2/3
P1/2/3
I4
P4
I5
P5
P6
I6
Basal Ganglia
TRN
Parvalbumin (FO,SP) Thalamus
Calbindin (HO,NSP) Thalamus
Senses
L=Layer
excitatory
Inhibitory

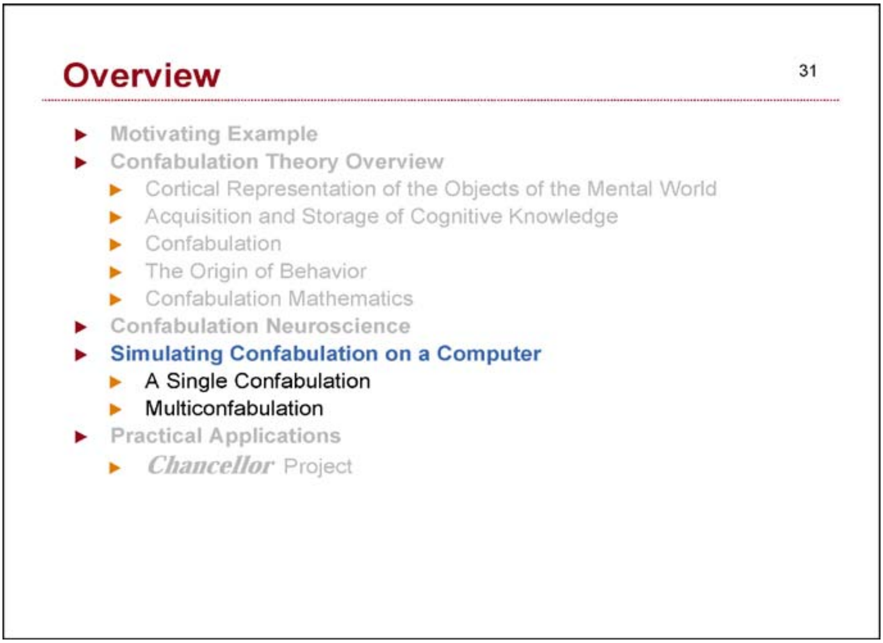
Graphical Neuroanatomical Representation courtesy of Soren Solari, 2006



UCSD Confabulation Neuroscience Course

30

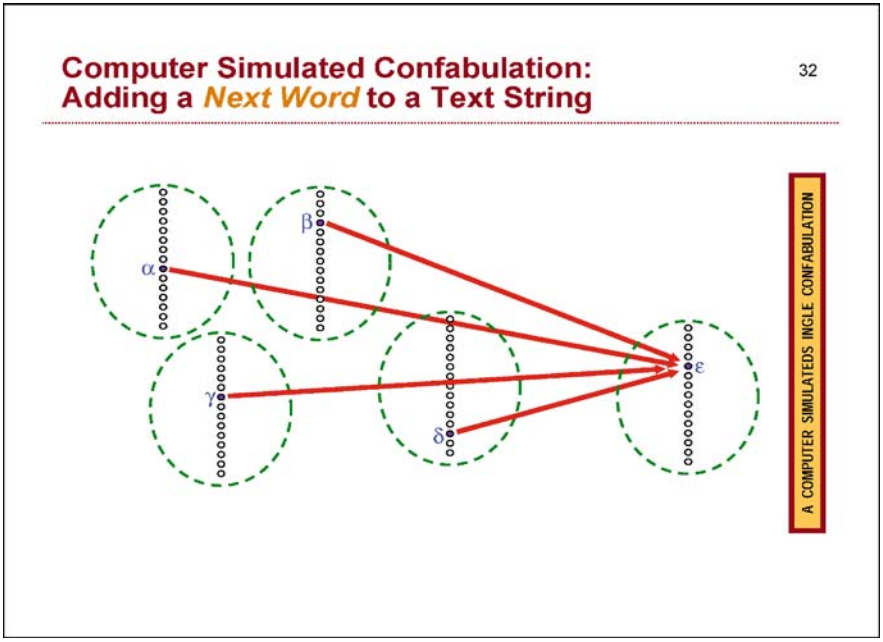
- **ECE-270A/B/C *Neurocomputing*** is a year-long (three quarter sequence) course covering **Confabulation Neuroscience Modeling**.
- ECE-270 participants build, evaluate, and present to the class, models of thalamocortical modules, knowledge links, behavioral triggering, skill learning, and reasoning capability acquisition.
- ECE-270A concentrates on learning Confabulation Theory and experimenting with Thalamocortical Module **Models 1, 2, 3, and 4**.
- ECE-270B concentrates on experimenting with Thalamocortical Module **Models 5 and 6** and Knowledge Link **Models A, B, and C**.
- ECE-270C concentrates on Behavioral Triggering **Model Alpha** and experiments with **Skill Learning** and **Reasoning Capability** acquisition.
- ECE-270 is the development platform for a possible future **Confabulation Neuroscience** textbook.



Overview

31

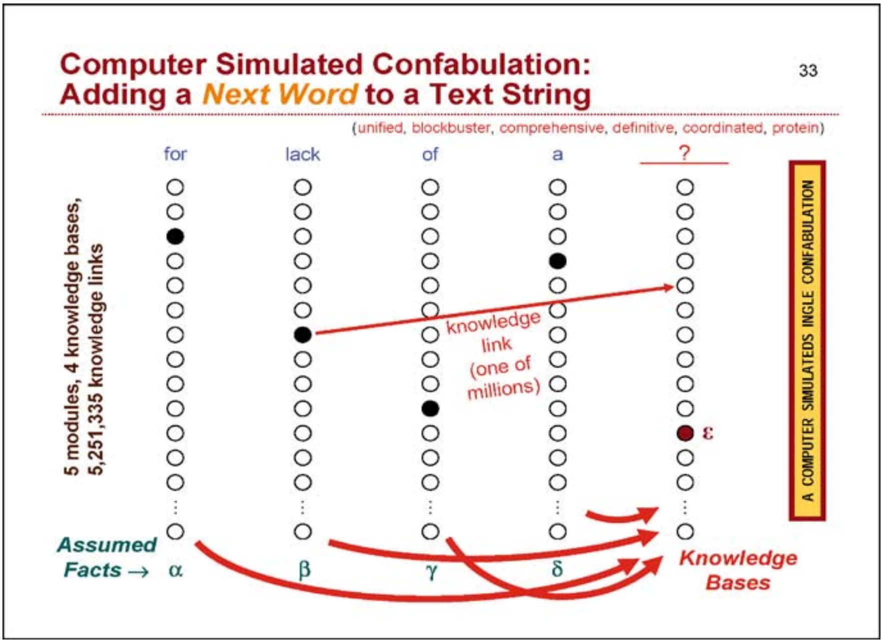
- Motivating Example
- Confabulation Theory Overview
  - Cortical Representation of the Objects of the Mental World
  - Acquisition and Storage of Cognitive Knowledge
  - Confabulation
  - The Origin of Behavior
  - Confabulation Mathematics
- Confabulation Neuroscience
- Simulating Confabulation on a Computer
  - A Single Confabulation
  - Multiconfabulation
- Practical Applications
  - *Chancellor* Project



Computer Simulated Confabulation: Adding a *Next Word* to a Text String

32

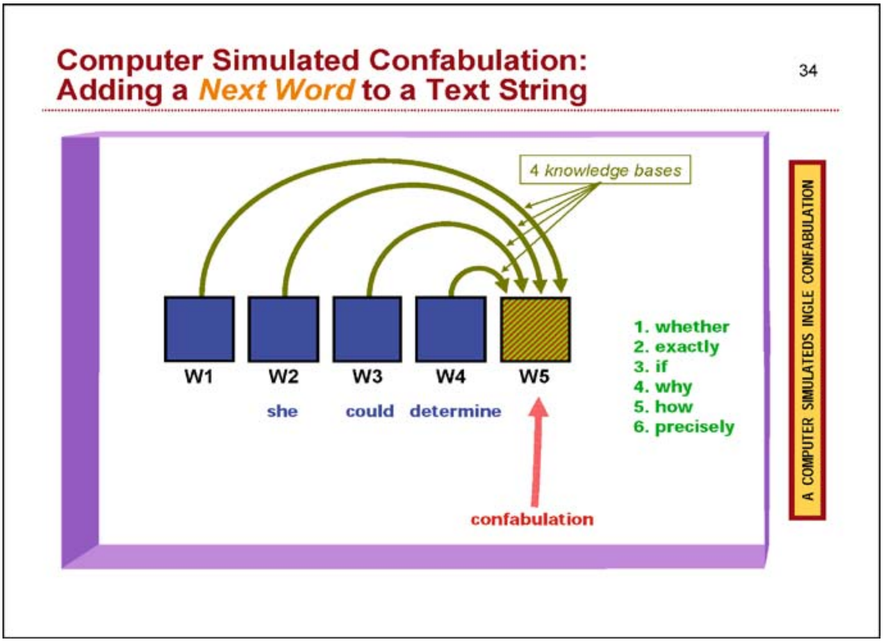
A COMPUTER SIMULATEDS INGLE CONFABULATION



Computer Simulated Confabulation:
Adding a *Next Word* to a Text String

33

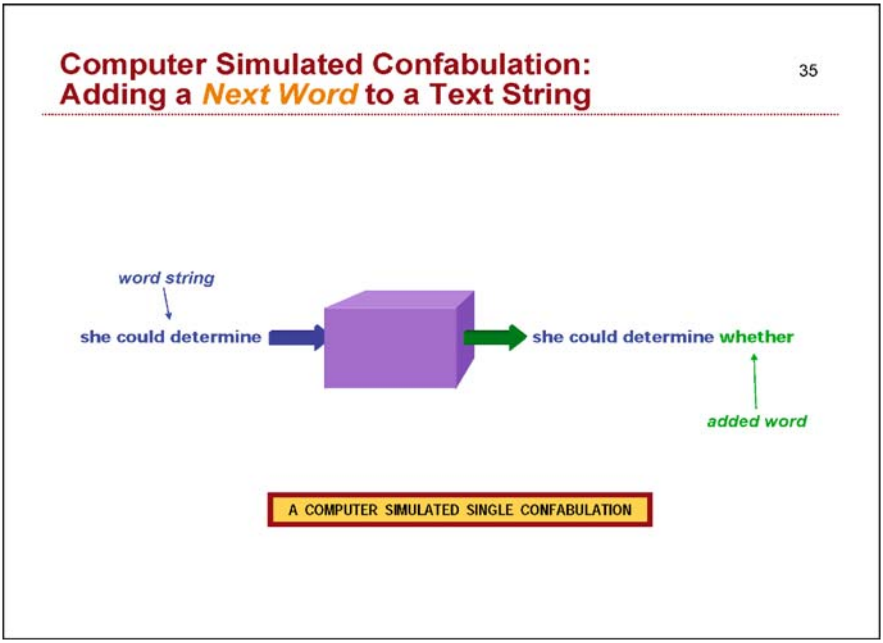
(unified, blockbuster, comprehensive, definitive, coordinated, protein)
for
lack
of
a
?
5 modules, 4 knowledge bases,
5,251,335 knowledge links
knowledge
link
(one of
millions)
ε
Assumed
Facts →
α
β
γ
δ
Knowledge
Bases
A COMPUTER SIMULATEDS INGLE CONFABULATION



Computer Simulated Confabulation:
Adding a *Next Word* to a Text String

34

1. whether
2. exactly
3. if
4. why
5. how
6. precisely



Computer Simulated Confabulation: Adding a *Next Word* to a Text String

35

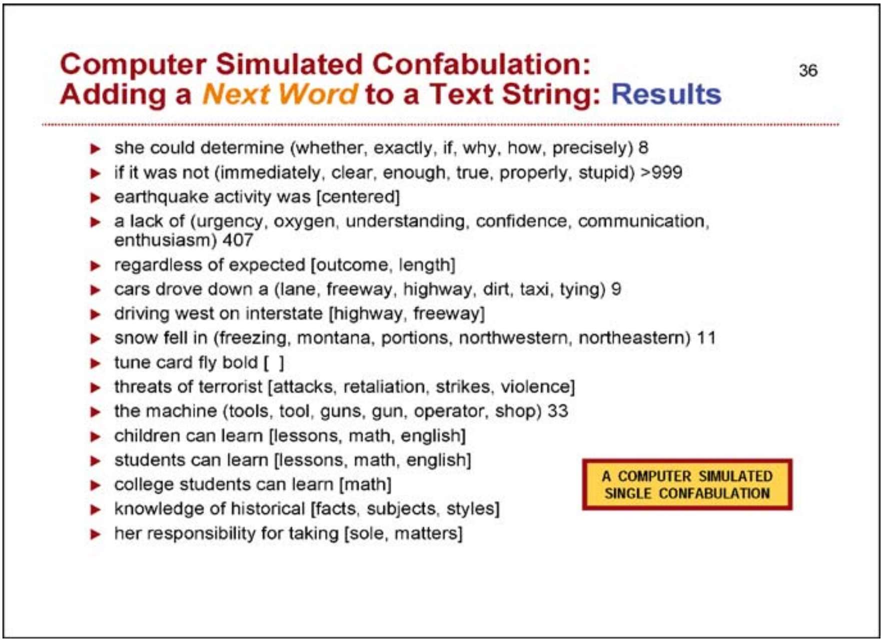
word string

she could determine

she could determine whether

added word

A COMPUTER SIMULATED SINGLE CONFABULATION

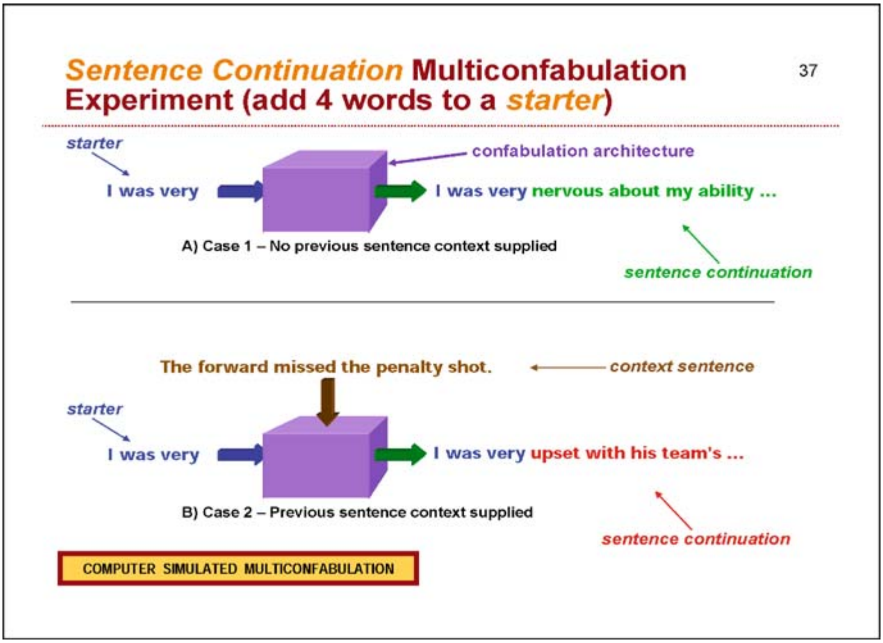


Computer Simulated Confabulation: Adding a *Next Word* to a Text String: Results

36

- she could determine (whether, exactly, if, why, how, precisely) 8
- if it was not (immediately, clear, enough, true, properly, stupid) >999
- earthquake activity was [centered]
- a lack of (urgency, oxygen, understanding, confidence, communication, enthusiasm) 407
- regardless of expected [outcome, length]
- cars drove down a (lane, freeway, highway, dirt, taxi, tying) 9
- driving west on interstate [highway, freeway]
- snow fell in (freezing, montana, portions, northwestern, northeastern) 11
- tune card fly bold [ ]
- threats of terrorist [attacks, retaliation, strikes, violence]
- the machine (tools, tool, guns, gun, operator, shop) 33
- children can learn [lessons, math, english]
- students can learn [lessons, math, english]
- college students can learn [math]
- knowledge of historical [facts, subjects, styles]
- her responsibility for taking [sole, matters]

A COMPUTER SIMULATED SINGLE CONFABULATION



Sentence Continuation Multiconfabulation Experiment (add 4 words to a *starter*)

37

starter

I was very

confabulation architecture

I was very nervous about my ability ...

A) Case 1 – No previous sentence context supplied

sentence continuation

The forward missed the penalty shot.

context sentence

starter

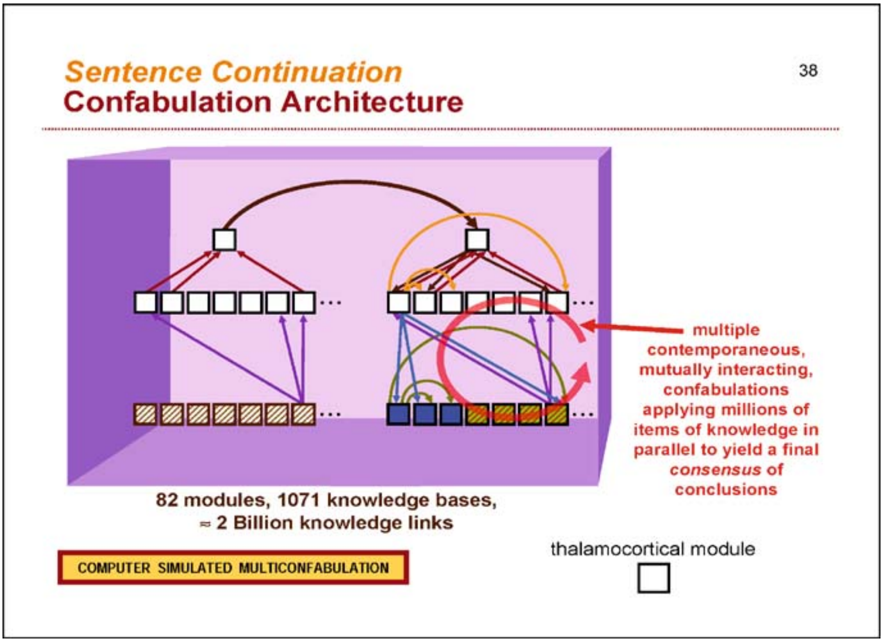
I was very

I was very upset with his team's ...

B) Case 2 – Previous sentence context supplied

sentence continuation

COMPUTER SIMULATED MULTICONFABULATION



Sentence Continuation Confabulation Architecture

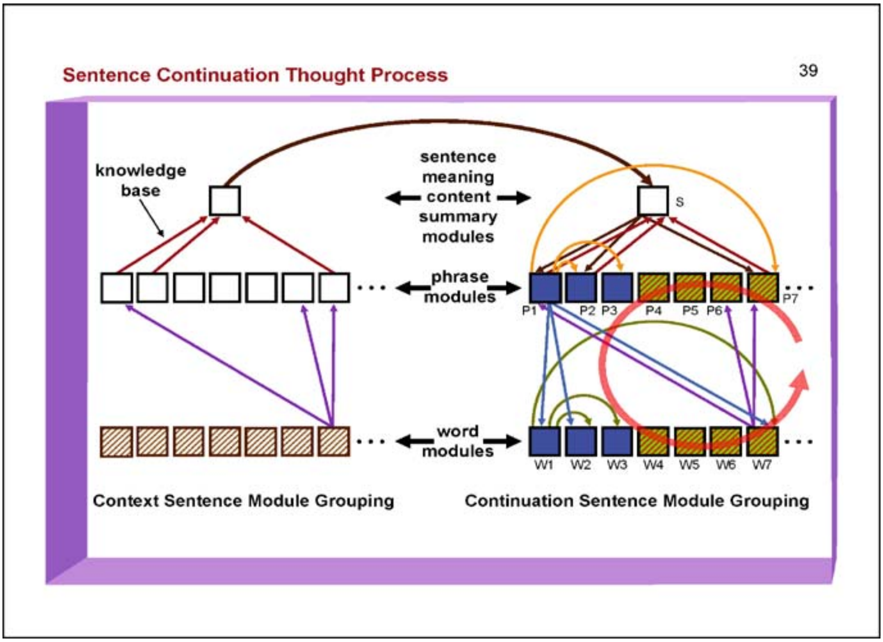
38

multiple contemporaneous, mutually interacting, confabulations applying millions of items of knowledge in parallel to yield a final *consensus* of conclusions

82 modules, 1071 knowledge bases, ≈ 2 Billion knowledge links

COMPUTER SIMULATED MULTICONFABULATION

thalamocortical module □



Sentence Continuation Thought Process

39

knowledge base

sentence meaning content summary modules

S

phrase modules

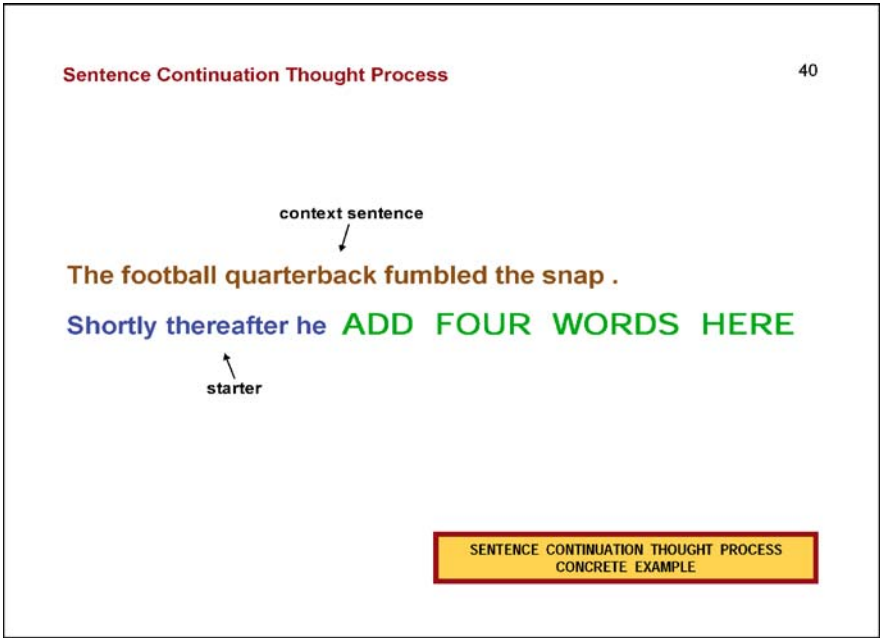
P1 P2 P3 P4 P5 P6 P7

word modules

W1 W2 W3 W4 W5 W6 W7

Context Sentence Module Grouping

Continuation Sentence Module Grouping



Sentence Continuation Thought Process

40

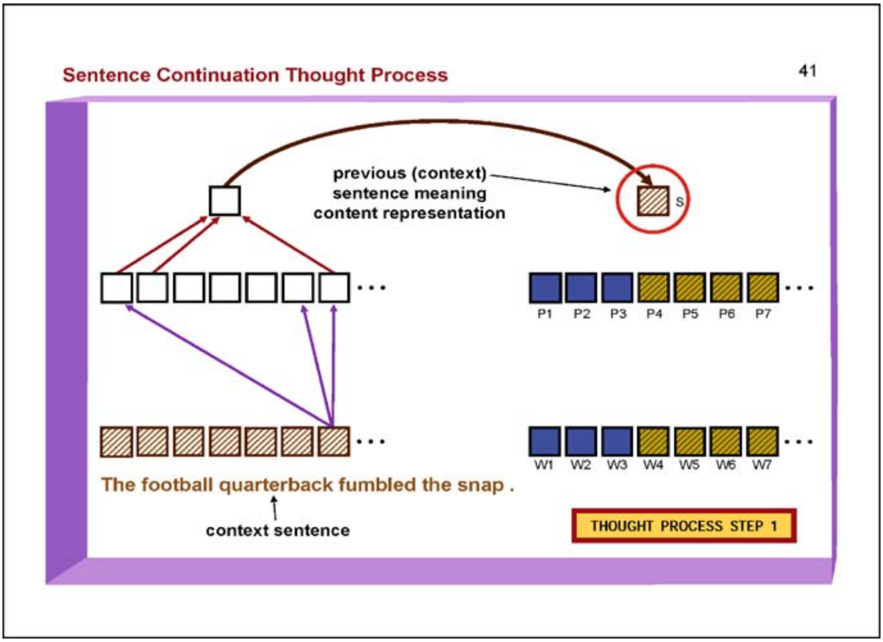
context sentence

The football quarterback fumbled the snap .

Shortly thereafter he ADD FOUR WORDS HERE

starter

SENTENCE CONTINUATION THOUGHT PROCESS
CONCRETE EXAMPLE



Sentence Continuation Thought Process

41

previous (context)
sentence meaning
content representation

S

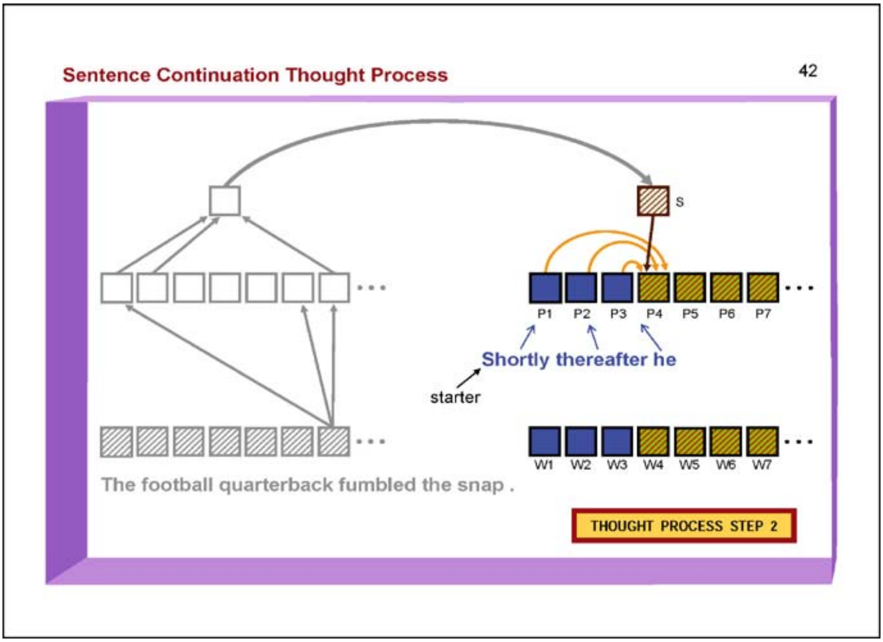
P1 P2 P3 P4 P5 P6 P7

W1 W2 W3 W4 W5 W6 W7

The football quarterback fumbled the snap .

context sentence

THOUGHT PROCESS STEP 1



Sentence Continuation Thought Process

42

S

P1 P2 P3 P4 P5 P6 P7

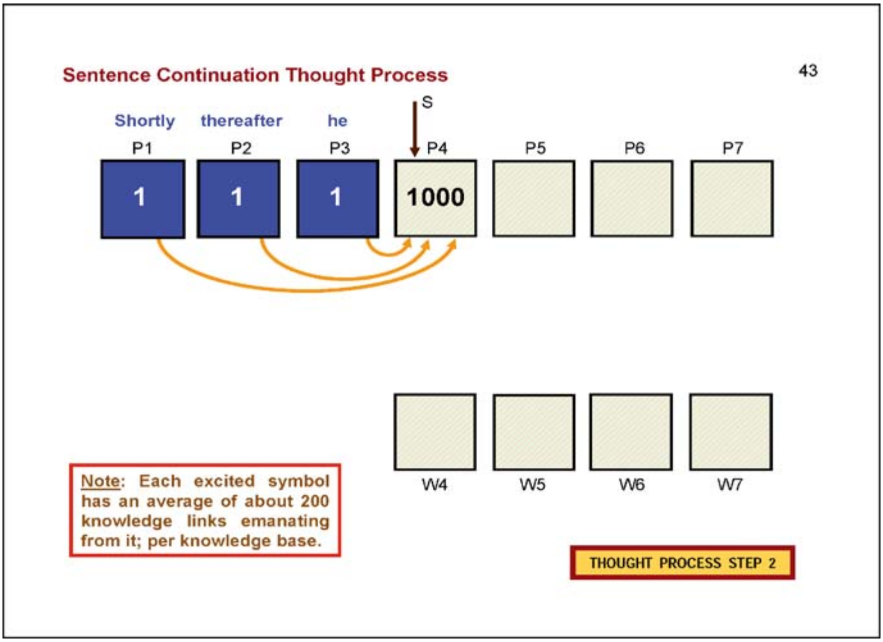
Shortly thereafter he

starter

W1 W2 W3 W4 W5 W6 W7

The football quarterback fumbled the snap .

THOUGHT PROCESS STEP 2



Sentence Continuation Thought Process

43

S

Shortly thereafter he

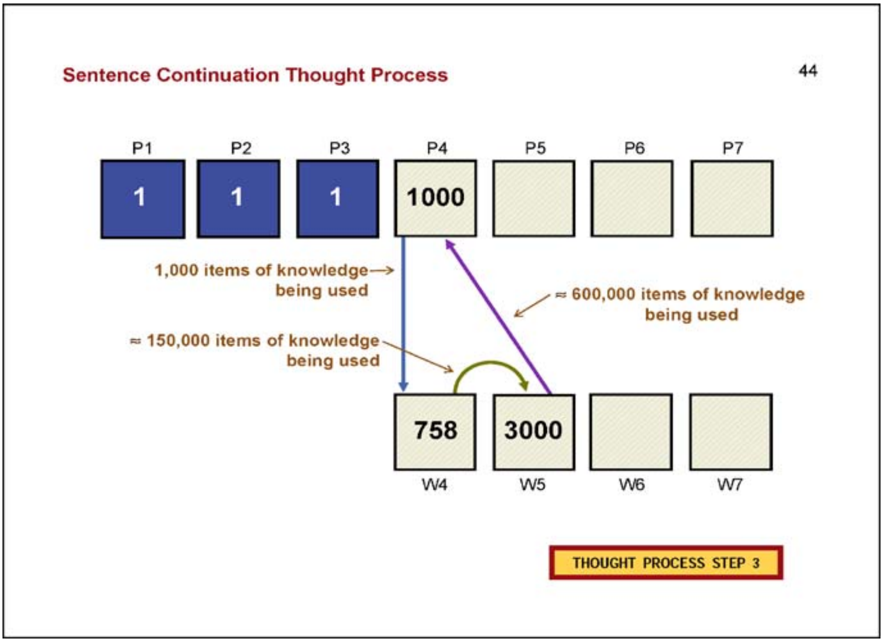
P1 P2 P3 P4 P5 P6 P7

1 1 1 1000

W4 W5 W6 W7

Note: Each excited symbol has an average of about 200 knowledge links emanating from it; per knowledge base.

THOUGHT PROCESS STEP 2



Sentence Continuation Thought Process

44

P1 P2 P3 P4 P5 P6 P7

1 1 1 1000

1,000 items of knowledge being used

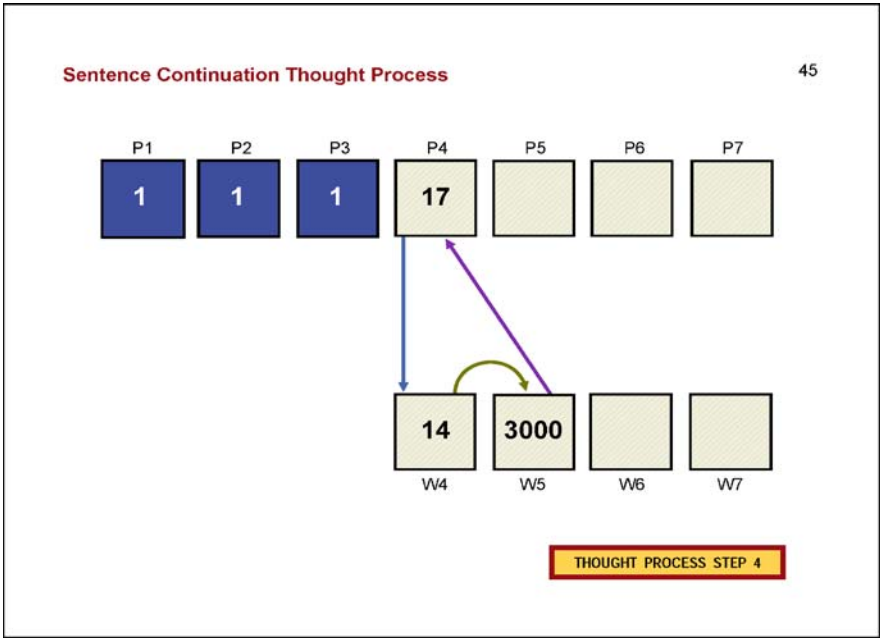
≈ 600,000 items of knowledge being used

≈ 150,000 items of knowledge being used

758 3000

W4 W5 W6 W7

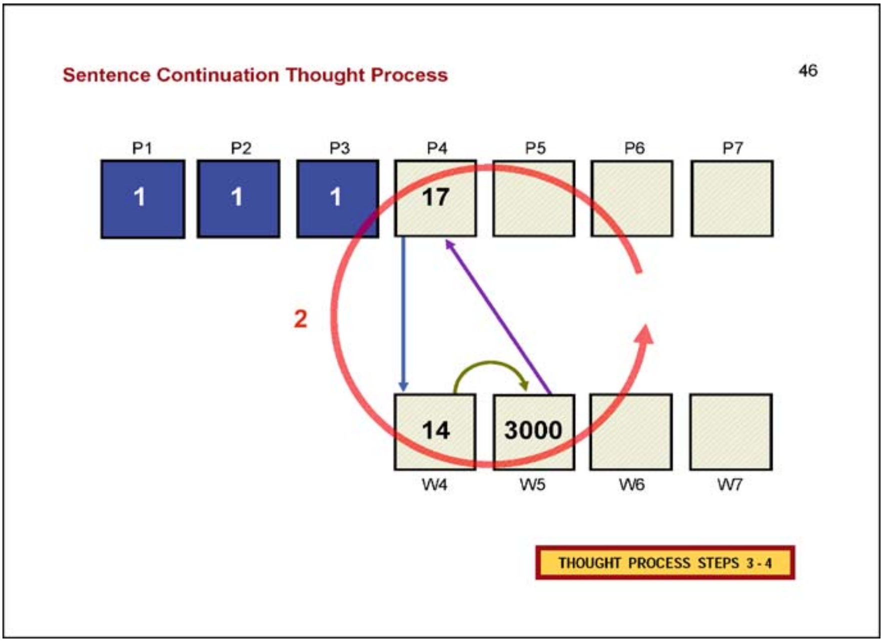
THOUGHT PROCESS STEP 3



Sentence Continuation Thought Process

45

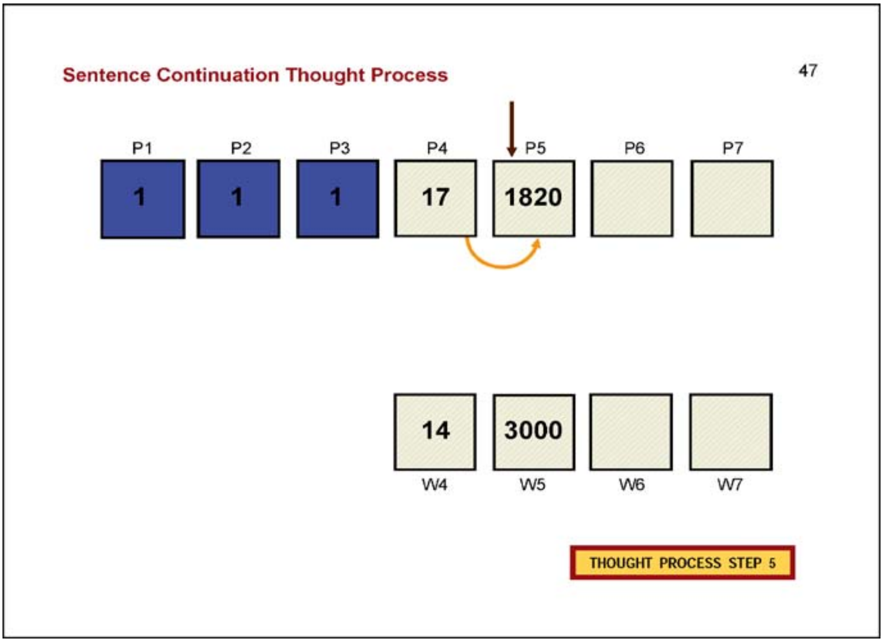
THOUGHT PROCESS STEP 4



Sentence Continuation Thought Process

46

THOUGHT PROCESS STEPS 3 - 4



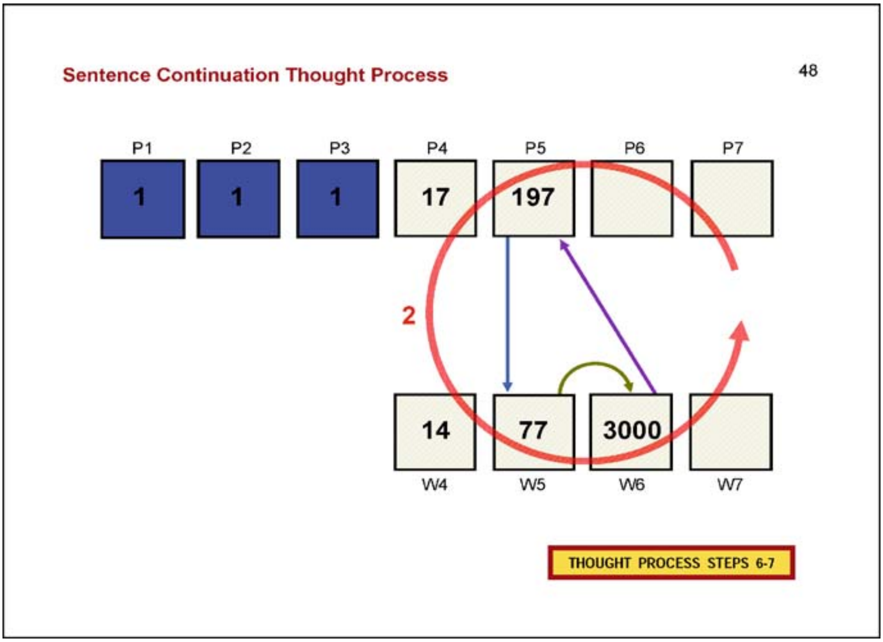
Sentence Continuation Thought Process

47

P1	P2	P3	P4	P5	P6	P7
1	1	1	17	1820		

14	3000		
W4	W5	W6	W7

THOUGHT PROCESS STEP 5



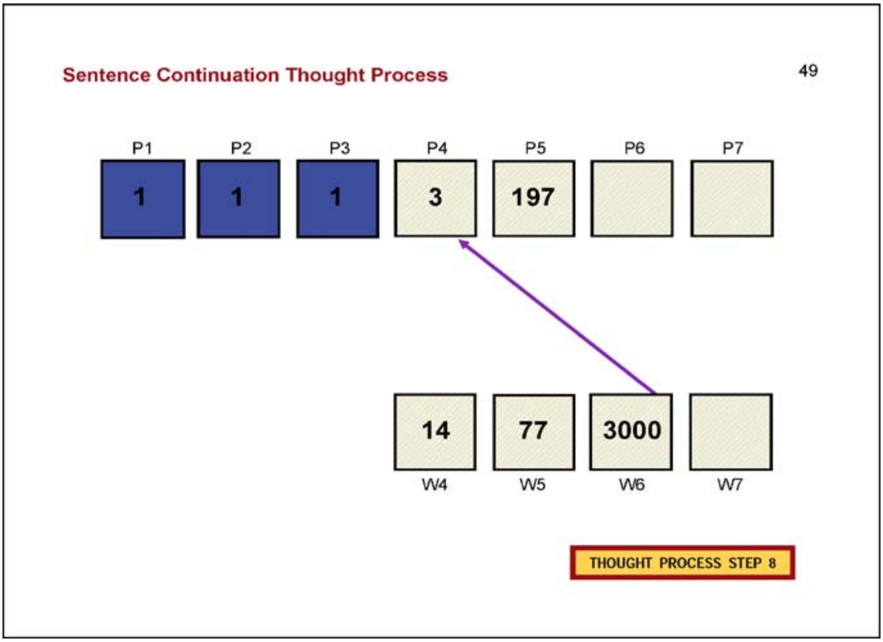
Sentence Continuation Thought Process

48

P1	P2	P3	P4	P5	P6	P7
1	1	1	17	197		

14	77	3000	
W4	W5	W6	W7

THOUGHT PROCESS STEPS 6-7



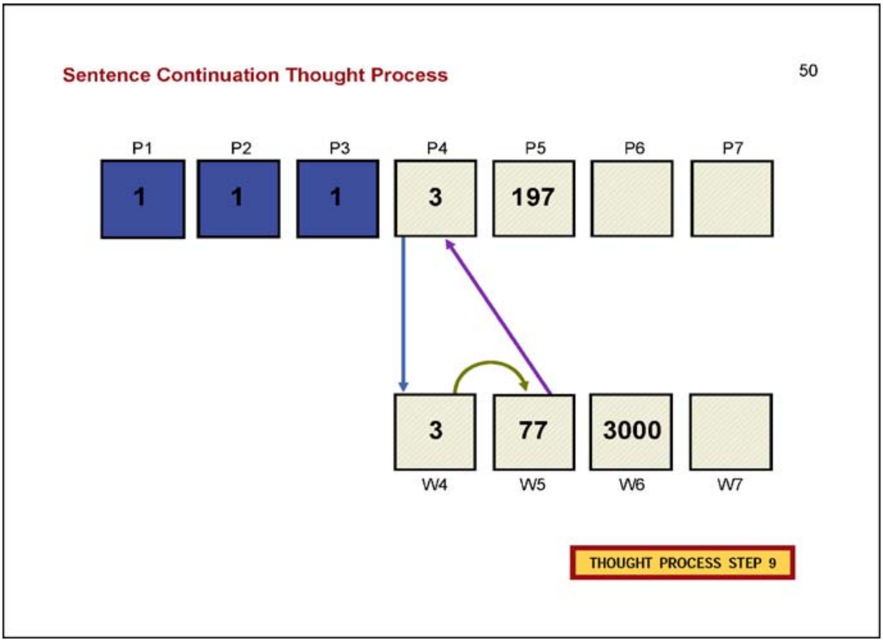
Sentence Continuation Thought Process

49

P1	P2	P3	P4	P5	P6	P7
1	1	1	3	197		

14	77	3000	
W4	W5	W6	W7

THOUGHT PROCESS STEP 8



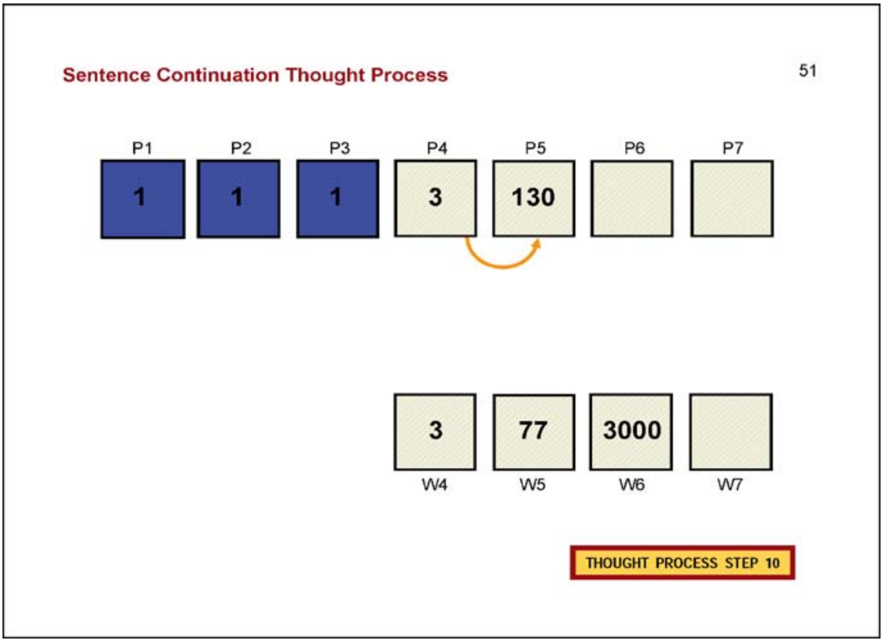
Sentence Continuation Thought Process

50

P1	P2	P3	P4	P5	P6	P7
1	1	1	3	197		

3	77	3000	
W4	W5	W6	W7

THOUGHT PROCESS STEP 9



Sentence Continuation Thought Process

51

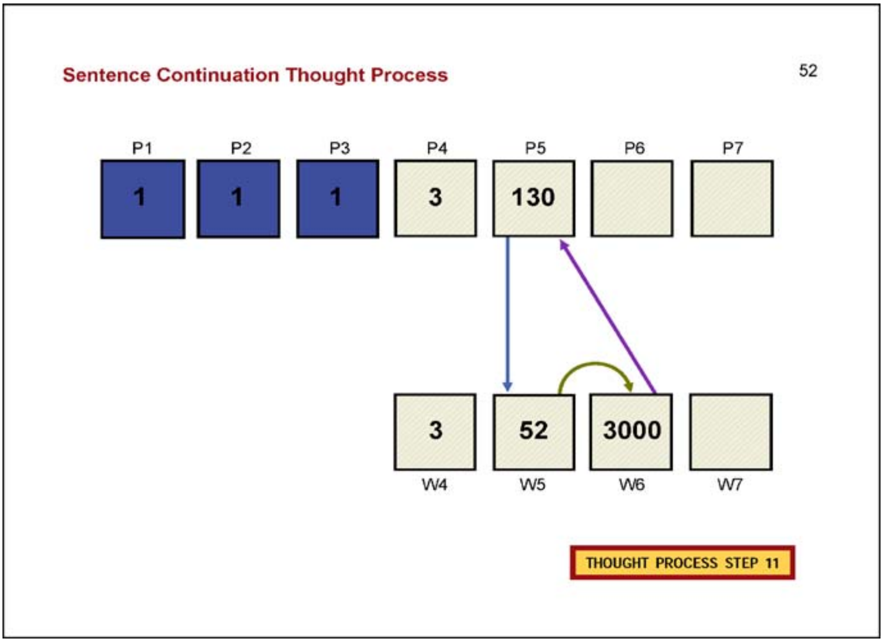
P1 P2 P3 P4 P5 P6 P7

1 1 1 3 130

3 77 3000

W4 W5 W6 W7

THOUGHT PROCESS STEP 10



Sentence Continuation Thought Process

52

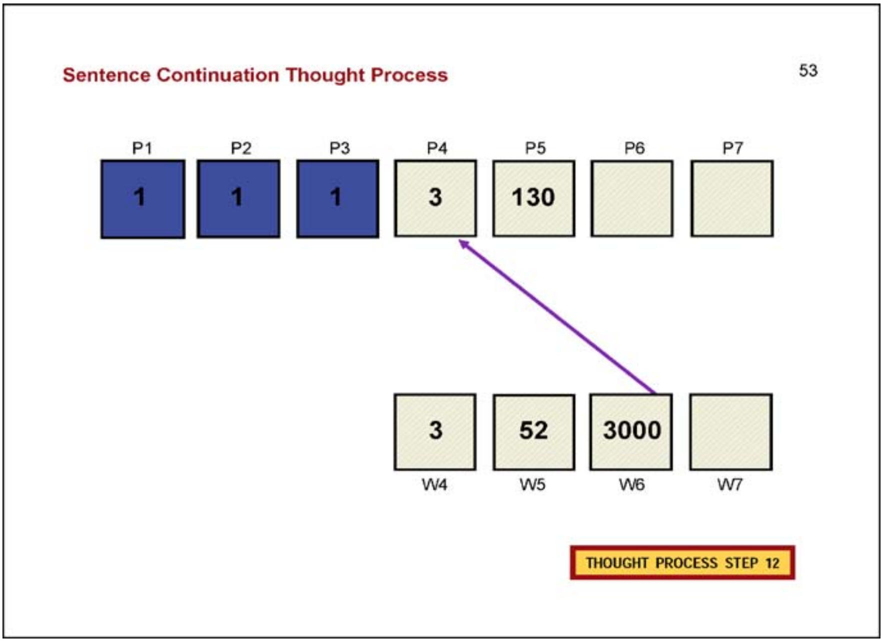
P1 P2 P3 P4 P5 P6 P7

1 1 1 3 130

3 52 3000

W4 W5 W6 W7

THOUGHT PROCESS STEP 11



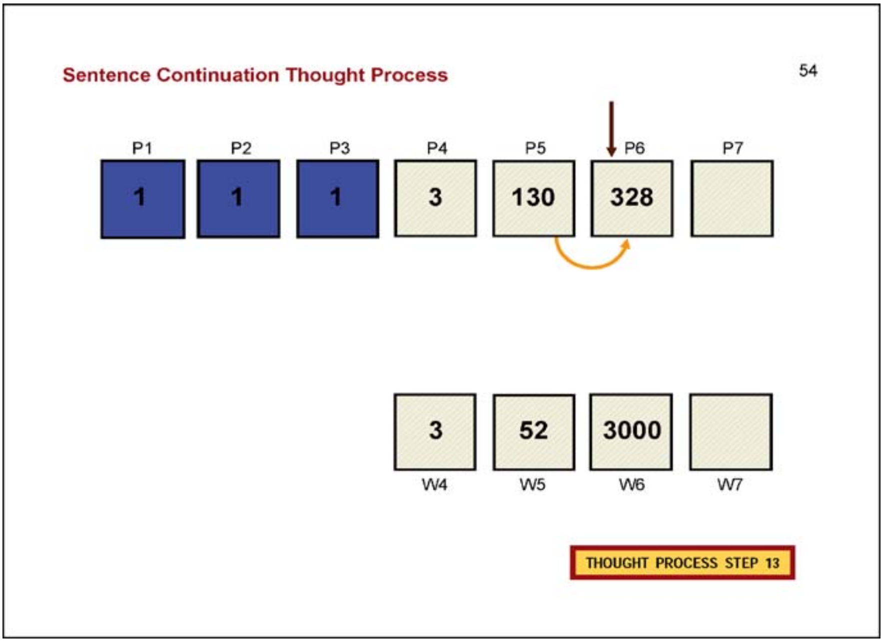
Sentence Continuation Thought Process

53

P1	P2	P3	P4	P5	P6	P7
1	1	1	3	130		

W4	W5	W6	W7
3	52	3000	

THOUGHT PROCESS STEP 12



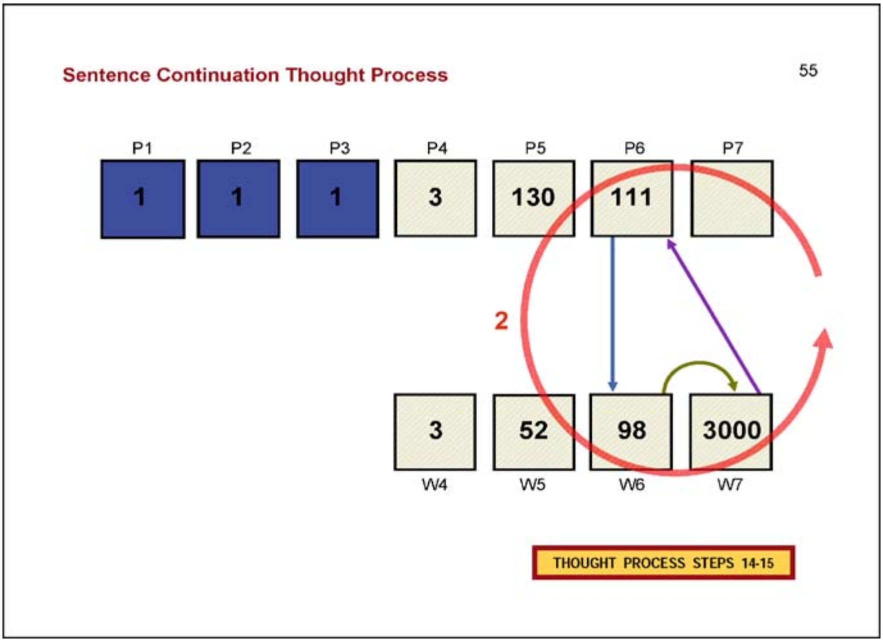
Sentence Continuation Thought Process

54

P1	P2	P3	P4	P5	P6	P7
1	1	1	3	130	328	

W4	W5	W6	W7
3	52	3000	

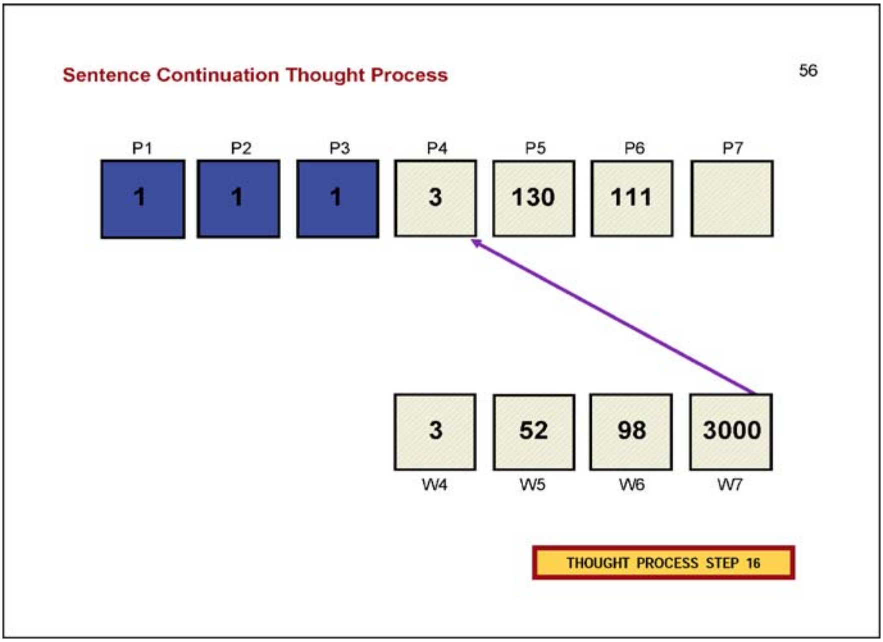
THOUGHT PROCESS STEP 13



Sentence Continuation Thought Process

55

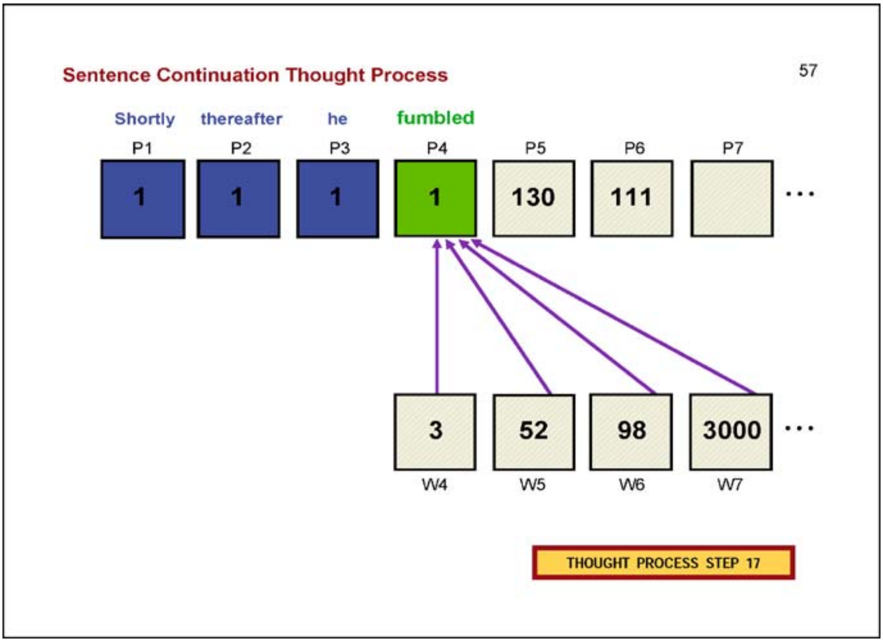
THOUGHT PROCESS STEPS 14-15



Sentence Continuation Thought Process

56

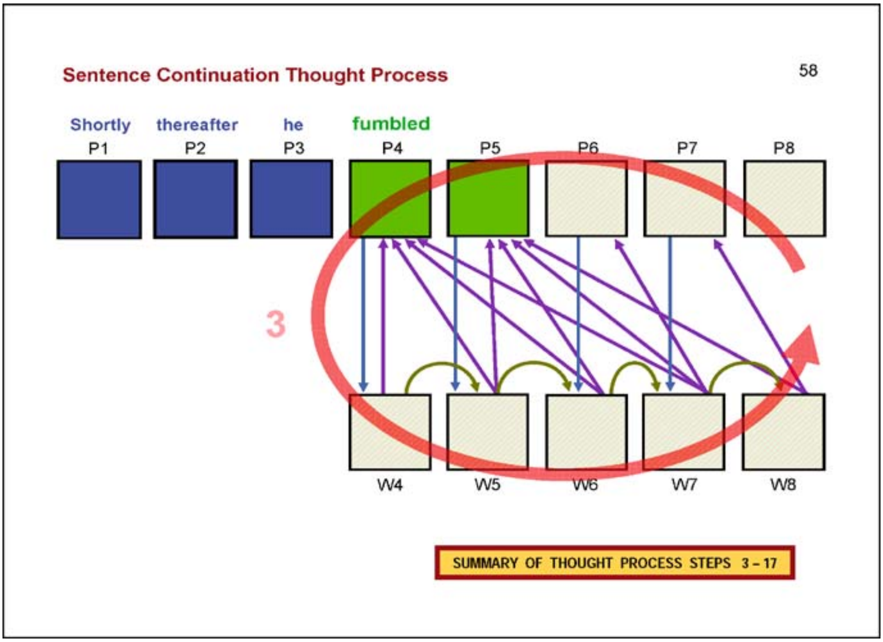
THOUGHT PROCESS STEP 16



Sentence Continuation Thought Process

57

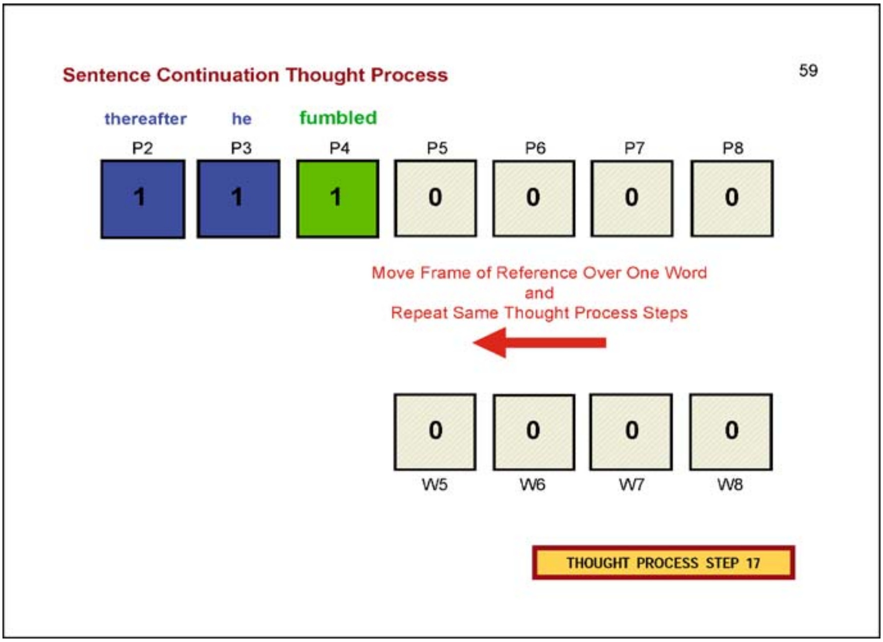
THOUGHT PROCESS STEP 17



Sentence Continuation Thought Process

58

SUMMARY OF THOUGHT PROCESS STEPS 3 – 17



Sentence Continuation Thought Process

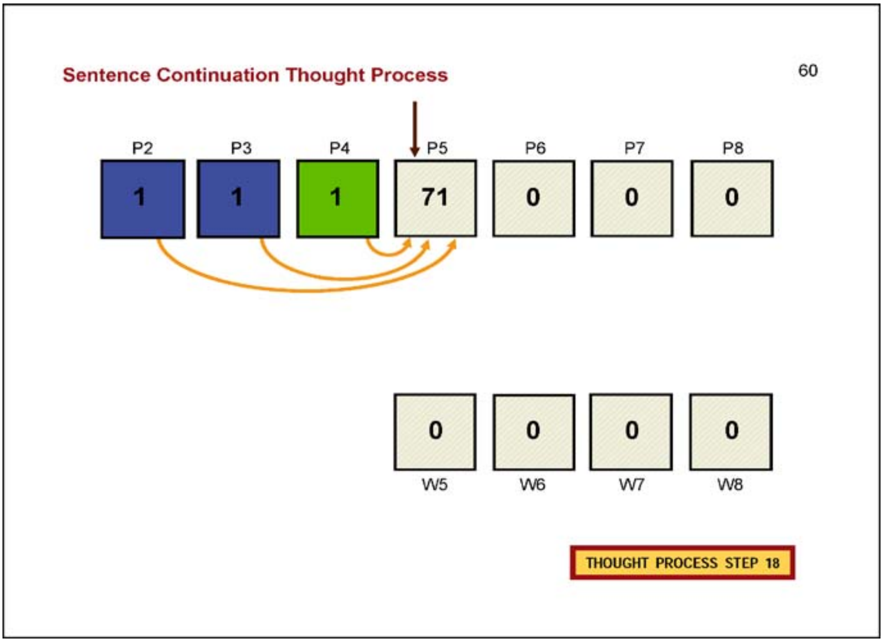
59

thereafter	he	fumbled				
P2	P3	P4	P5	P6	P7	P8
1	1	1	0	0	0	0

Move Frame of Reference Over One Word
and
Repeat Same Thought Process Steps

0	0	0	0
W5	W6	W7	W8

THOUGHT PROCESS STEP 17



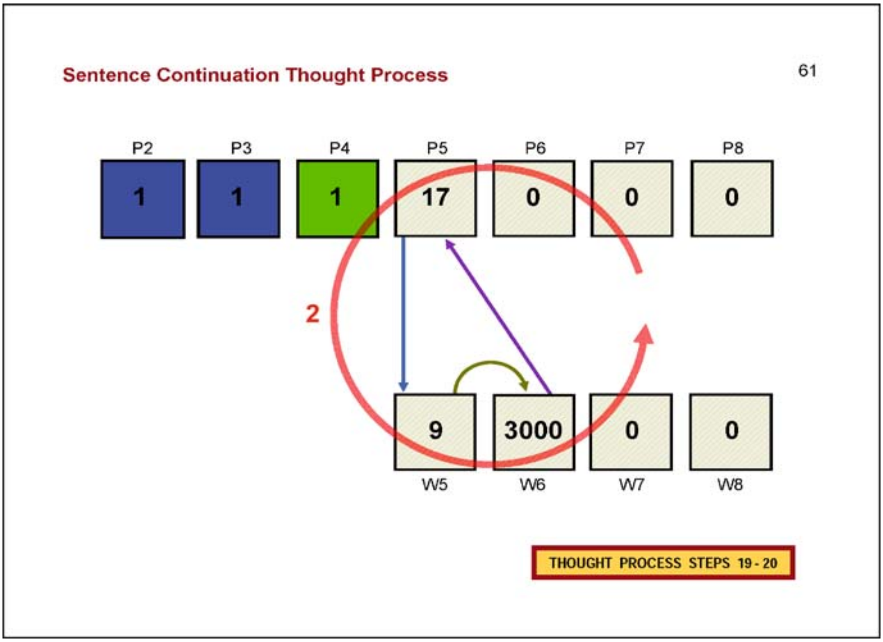
Sentence Continuation Thought Process

60

P2	P3	P4	P5	P6	P7	P8
1	1	1	71	0	0	0

0	0	0	0
W5	W6	W7	W8

THOUGHT PROCESS STEP 18



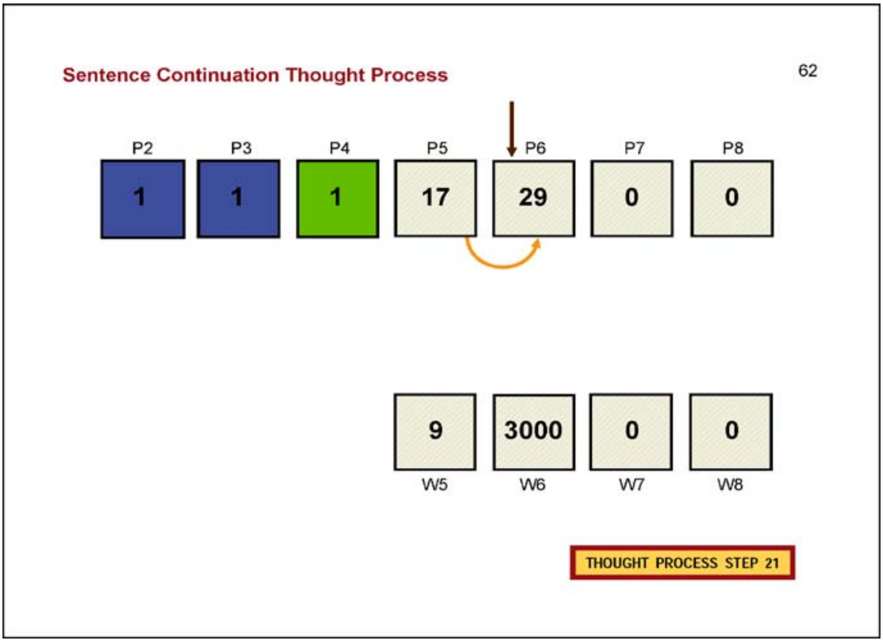
Sentence Continuation Thought Process

61

P2	P3	P4	P5	P6	P7	P8
1	1	1	17	0	0	0

9	3000	0	0
W5	W6	W7	W8

THOUGHT PROCESS STEPS 19 - 20



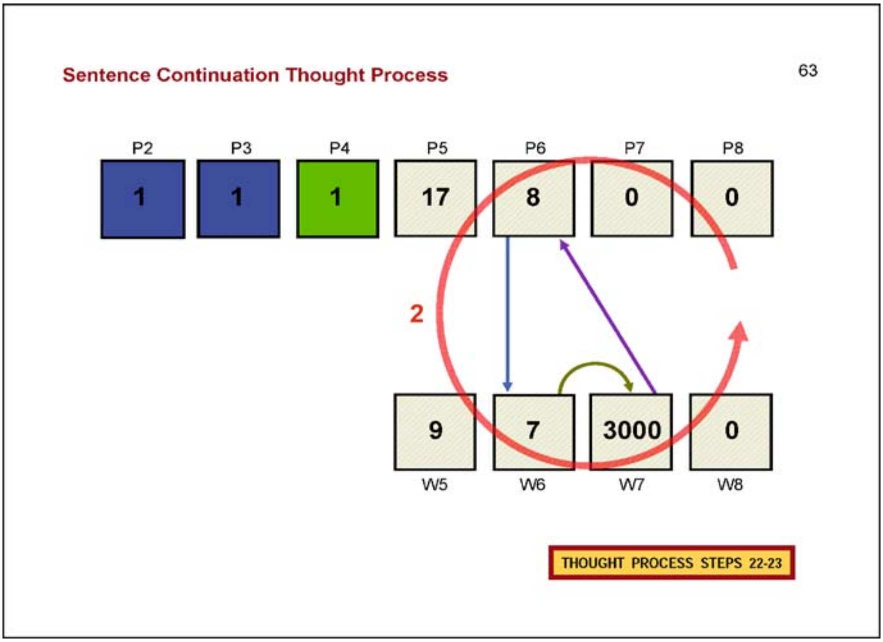
Sentence Continuation Thought Process

62

P2	P3	P4	P5	P6	P7	P8
1	1	1	17	29	0	0

9	3000	0	0
W5	W6	W7	W8

THOUGHT PROCESS STEP 21



Sentence Continuation Thought Process

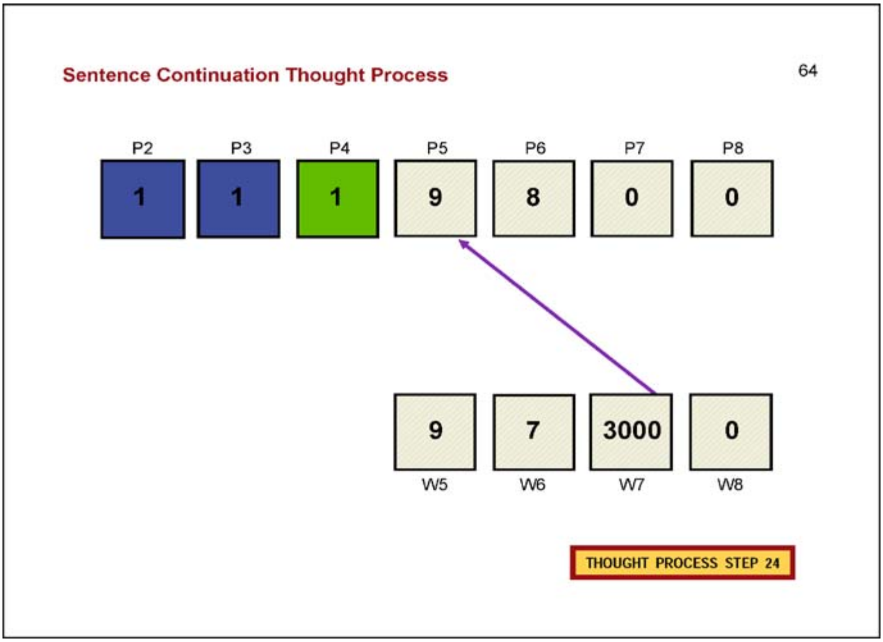
63

P2	P3	P4	P5	P6	P7	P8
1	1	1	17	8	0	0

2

9	7	3000	0
W5	W6	W7	W8

THOUGHT PROCESS STEPS 22-23



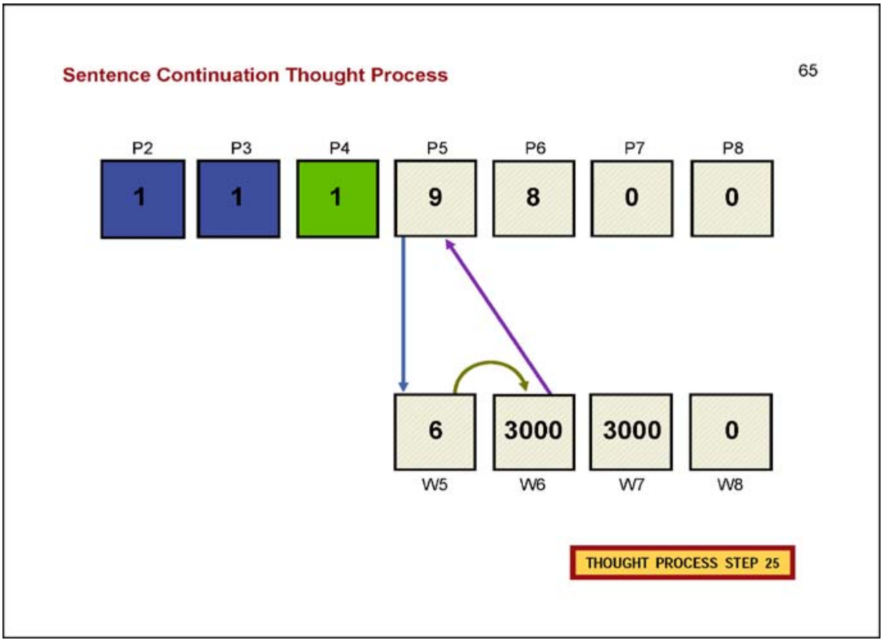
Sentence Continuation Thought Process

64

P2	P3	P4	P5	P6	P7	P8
1	1	1	9	8	0	0

9	7	3000	0
W5	W6	W7	W8

THOUGHT PROCESS STEP 24



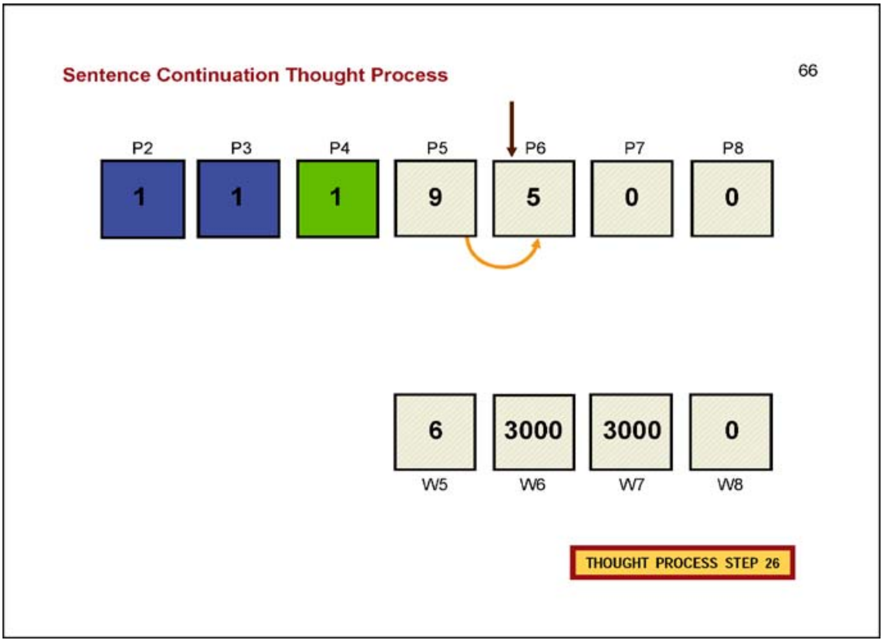
Sentence Continuation Thought Process

65

P2	P3	P4	P5	P6	P7	P8
1	1	1	9	8	0	0

W5	W6	W7	W8
6	3000	3000	0

THOUGHT PROCESS STEP 25



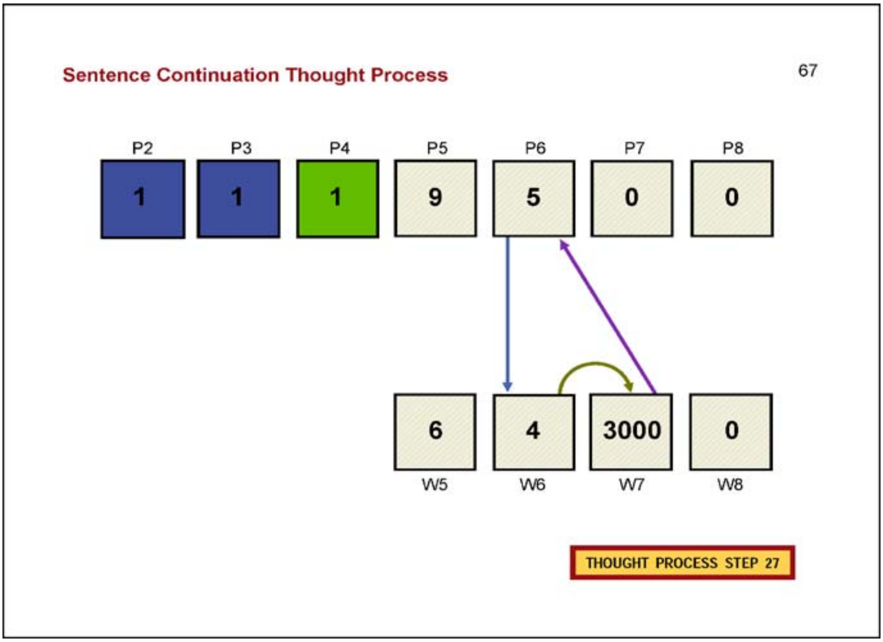
Sentence Continuation Thought Process

66

P2	P3	P4	P5	P6	P7	P8
1	1	1	9	5	0	0

W5	W6	W7	W8
6	3000	3000	0

THOUGHT PROCESS STEP 26



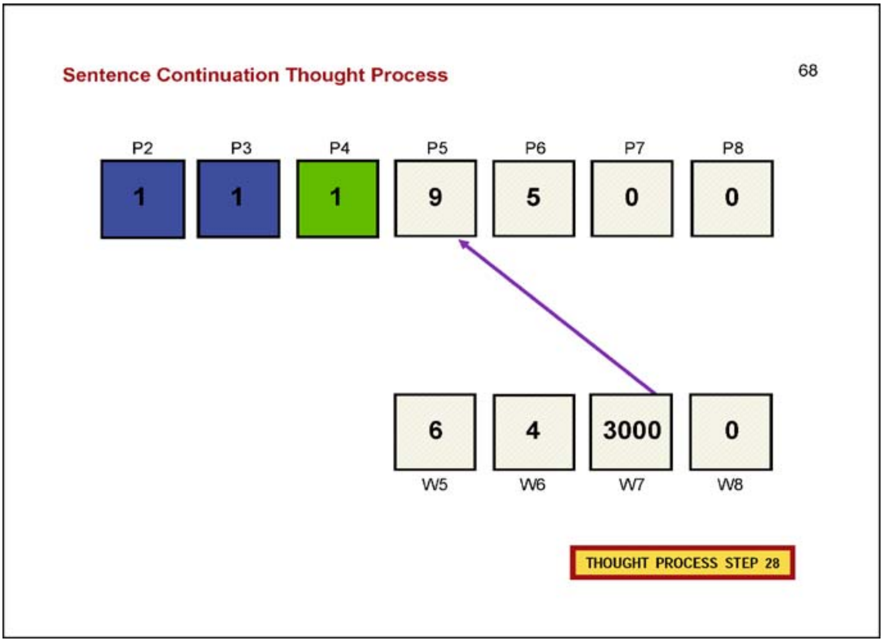
Sentence Continuation Thought Process

67

P2	P3	P4	P5	P6	P7	P8
1	1	1	9	5	0	0

W5	W6	W7	W8
6	4	3000	0

THOUGHT PROCESS STEP 27



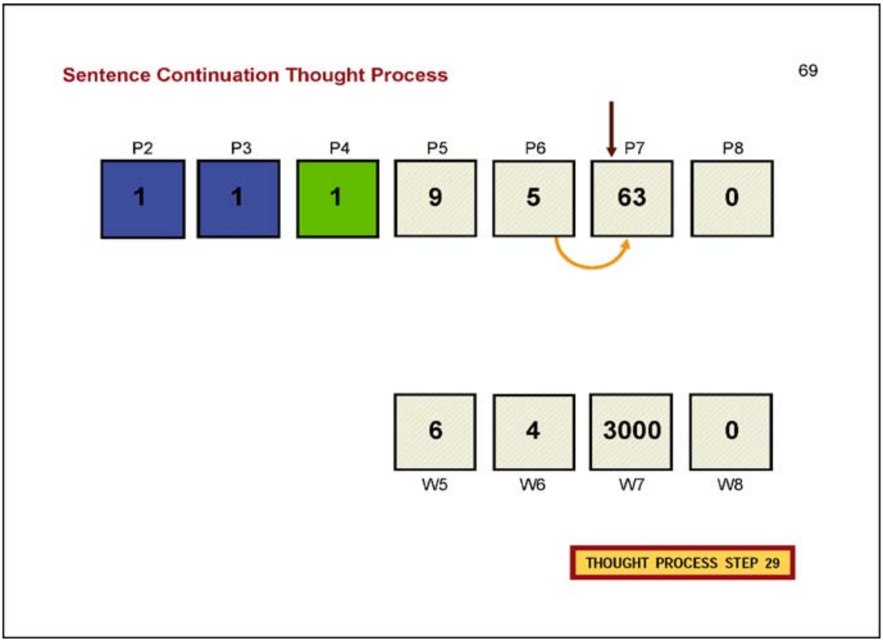
Sentence Continuation Thought Process

68

P2	P3	P4	P5	P6	P7	P8
1	1	1	9	5	0	0

W5	W6	W7	W8
6	4	3000	0

THOUGHT PROCESS STEP 28



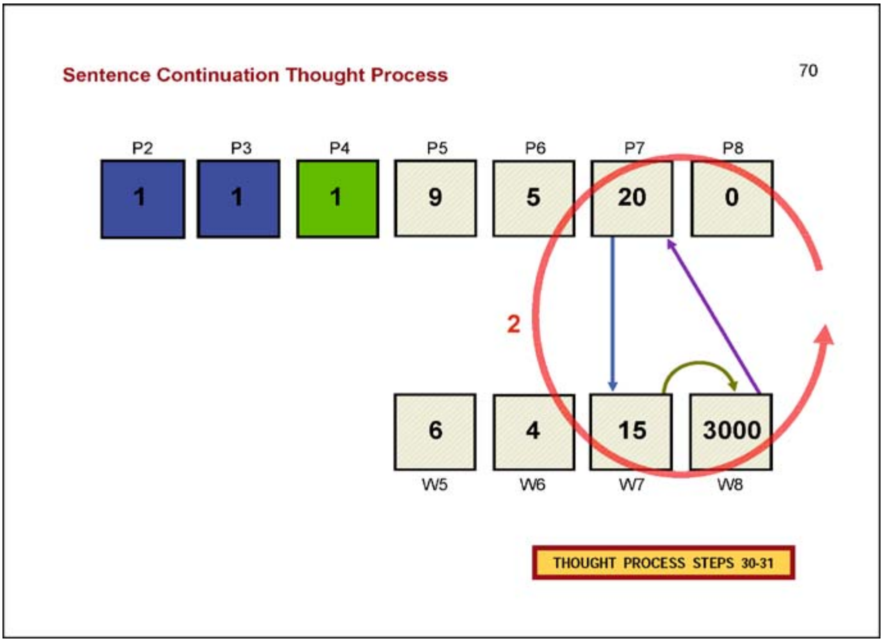
Sentence Continuation Thought Process

69

P2	P3	P4	P5	P6	P7	P8
1	1	1	9	5	63	0

W5	W6	W7	W8
6	4	3000	0

THOUGHT PROCESS STEP 29



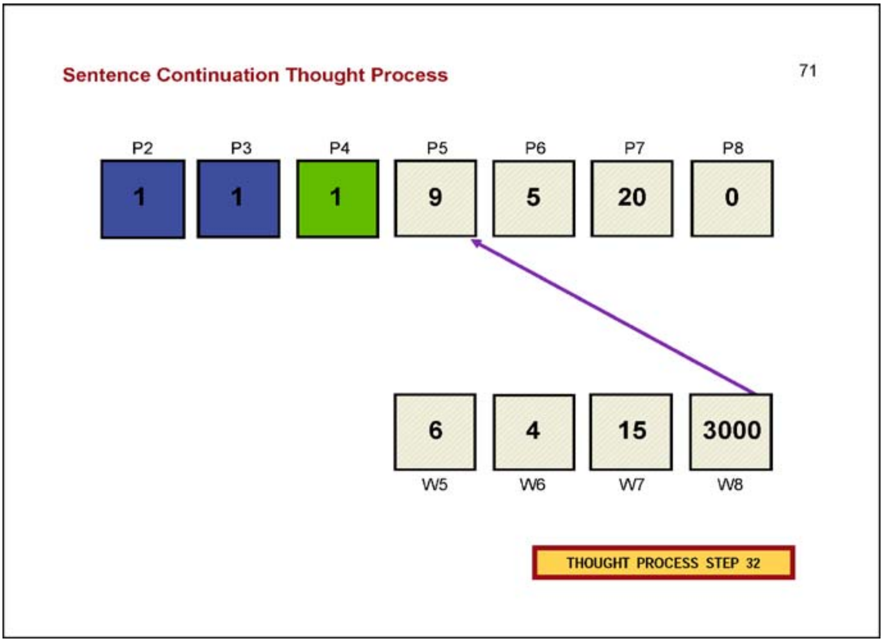
Sentence Continuation Thought Process

70

P2	P3	P4	P5	P6	P7	P8
1	1	1	9	5	20	0

W5	W6	W7	W8
6	4	15	3000

THOUGHT PROCESS STEPS 30-31



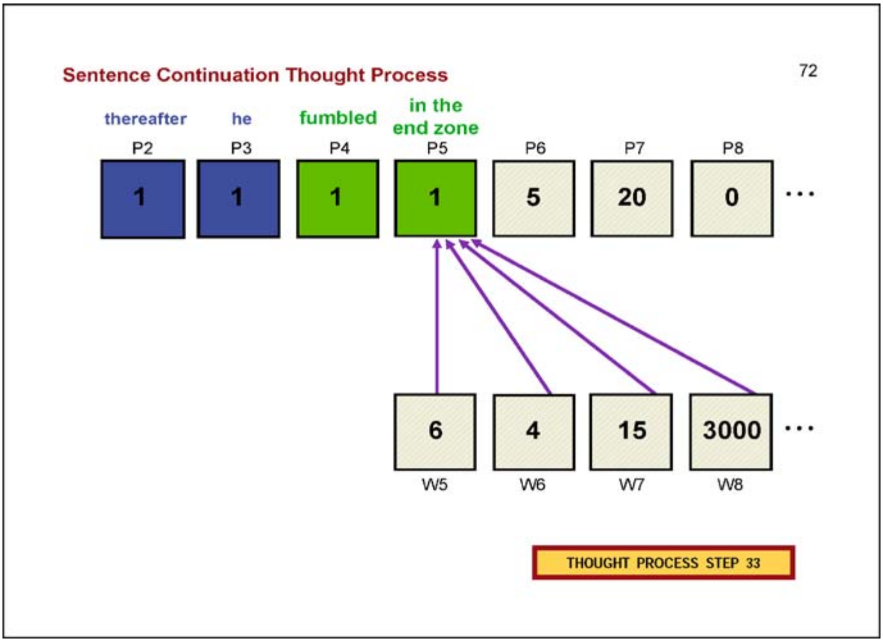
Sentence Continuation Thought Process

71

P2	P3	P4	P5	P6	P7	P8
1	1	1	9	5	20	0

6	4	15	3000
W5	W6	W7	W8

THOUGHT PROCESS STEP 32



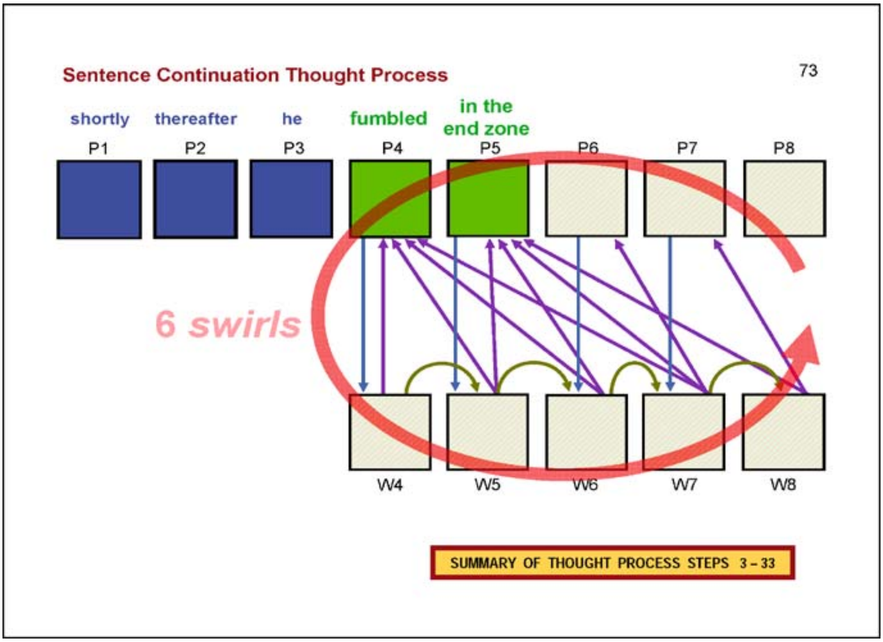
Sentence Continuation Thought Process

72

thereafter	he	fumbled	in the end zone				
P2	P3	P4	P5	P6	P7	P8	
1	1	1	1	5	20	0	···

6	4	15	3000	···
W5	W6	W7	W8	

THOUGHT PROCESS STEP 33



Sentence Continuation Thought Process

73

SUMMARY OF THOUGHT PROCESS STEPS 3 – 33

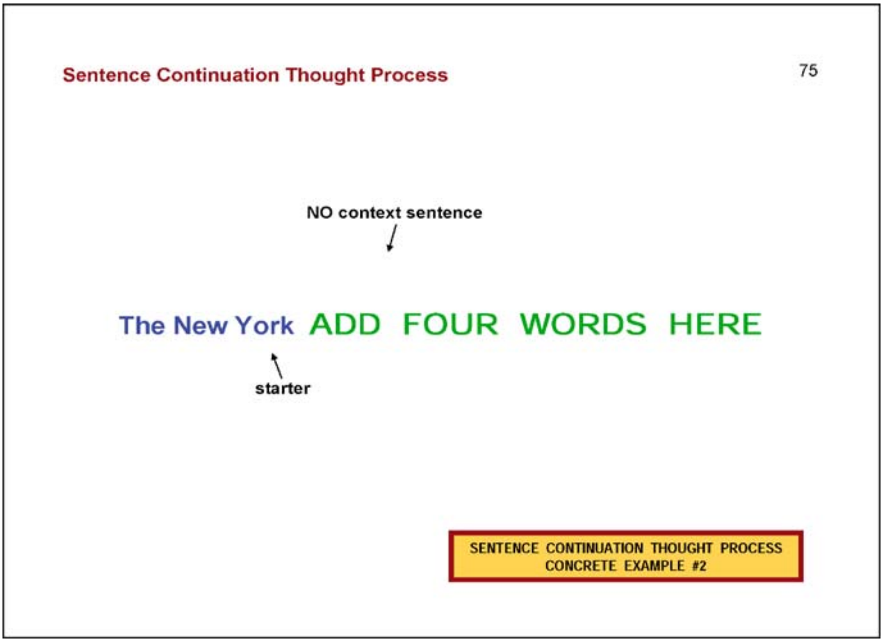
Sentence Continuation Thought Process

74

The football quarterback fumbled the snap .

Shortly thereafter he fumbled in the end zone …

FINAL ANSWER



Sentence Continuation Thought Process

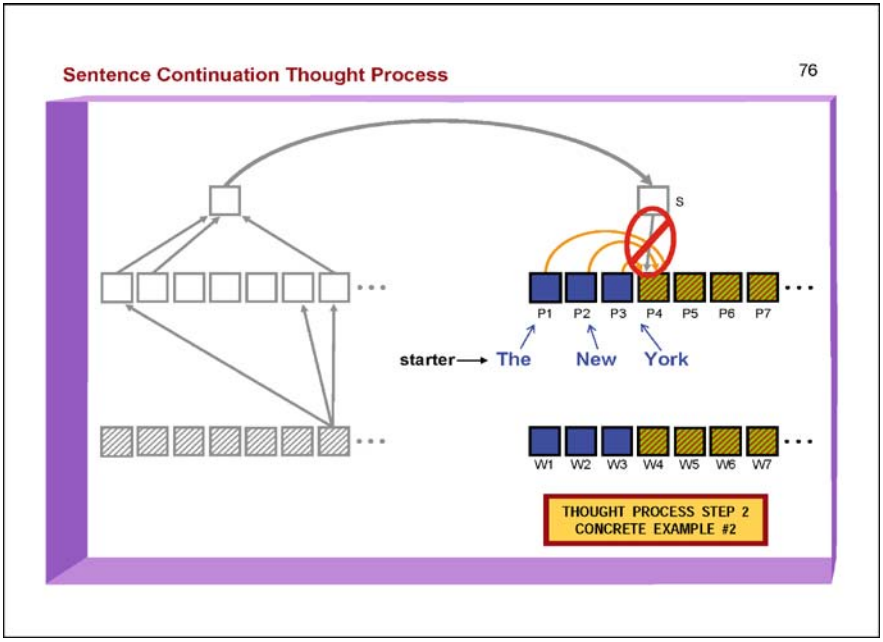
75

NO context sentence

The New York ADD FOUR WORDS HERE

starter

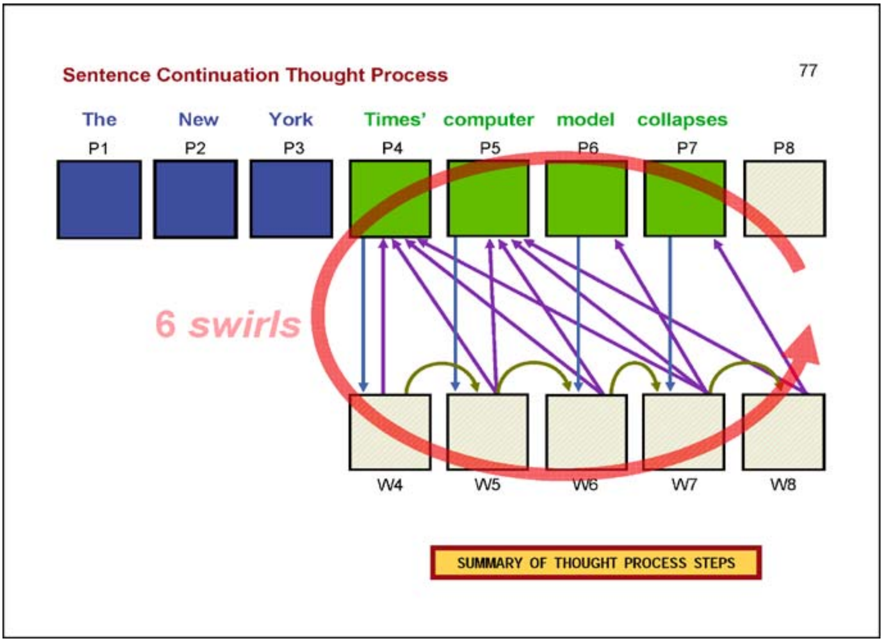
SENTENCE CONTINUATION THOUGHT PROCESS
CONCRETE EXAMPLE #2



Sentence Continuation Thought Process

76

THOUGHT PROCESS STEP 2
CONCRETE EXAMPLE #2



Sentence Continuation Thought Process

77

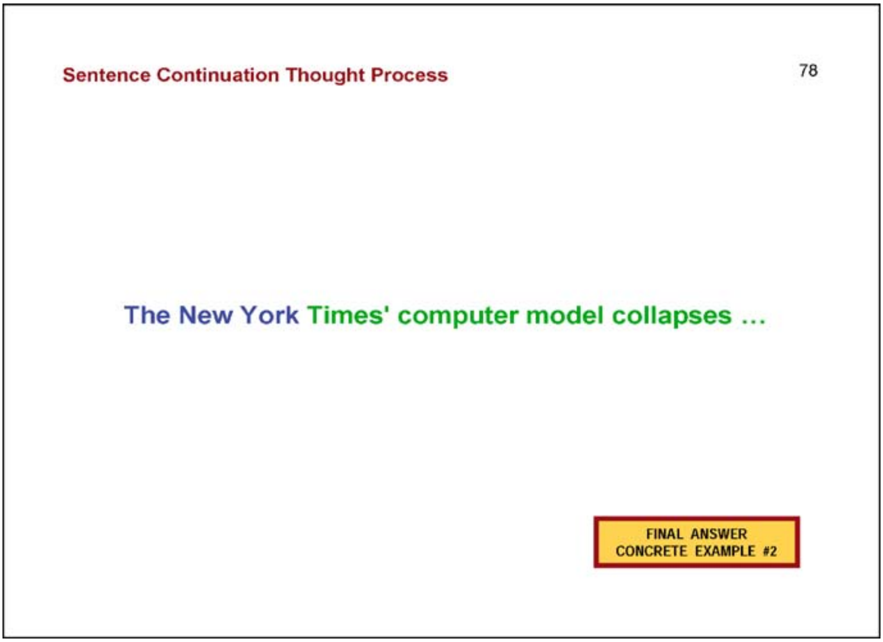
The New York Times' computer model collapses

P1 P2 P3 P4 P5 P6 P7 P8

6 swirls

W4 W5 W6 W7 W8

SUMMARY OF THOUGHT PROCESS STEPS

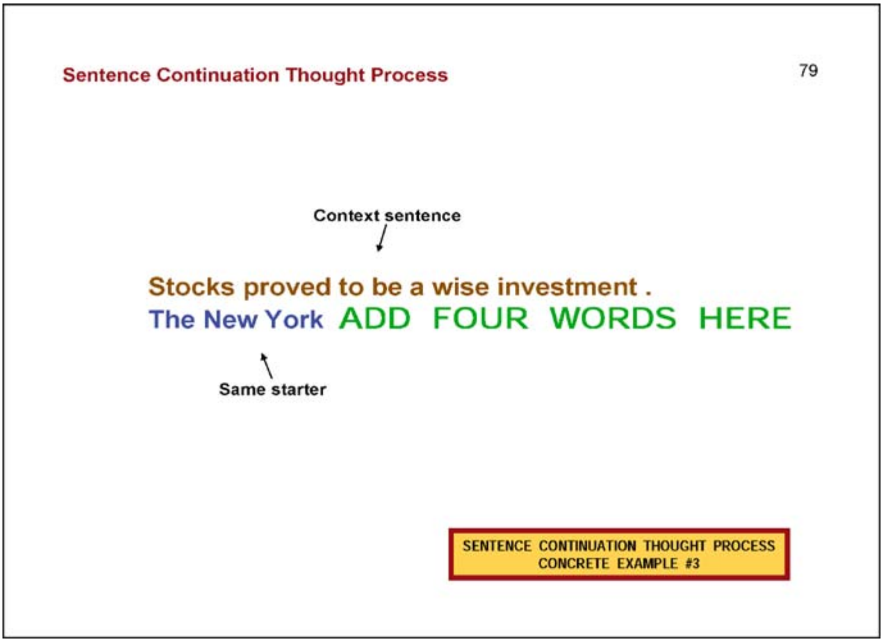


Sentence Continuation Thought Process

78

The New York Times' computer model collapses …

FINAL ANSWER
CONCRETE EXAMPLE #2



Sentence Continuation Thought Process

79

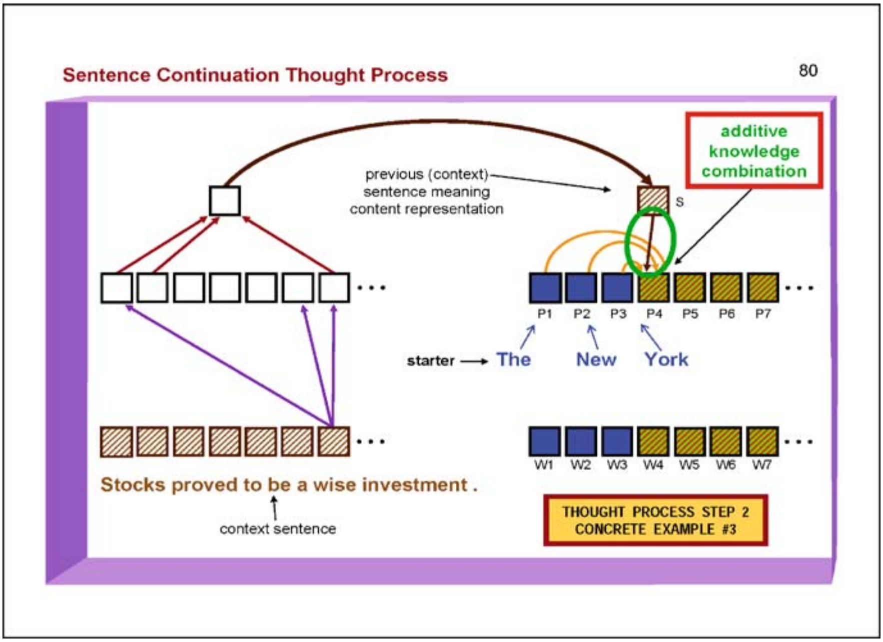
Context sentence

Stocks proved to be a wise investment .

The New York ADD FOUR WORDS HERE

Same starter

SENTENCE CONTINUATION THOUGHT PROCESS
CONCRETE EXAMPLE #3



Sentence Continuation Thought Process

80

additive knowledge combination

previous (context) sentence meaning content representation

S

P1 P2 P3 P4 P5 P6 P7

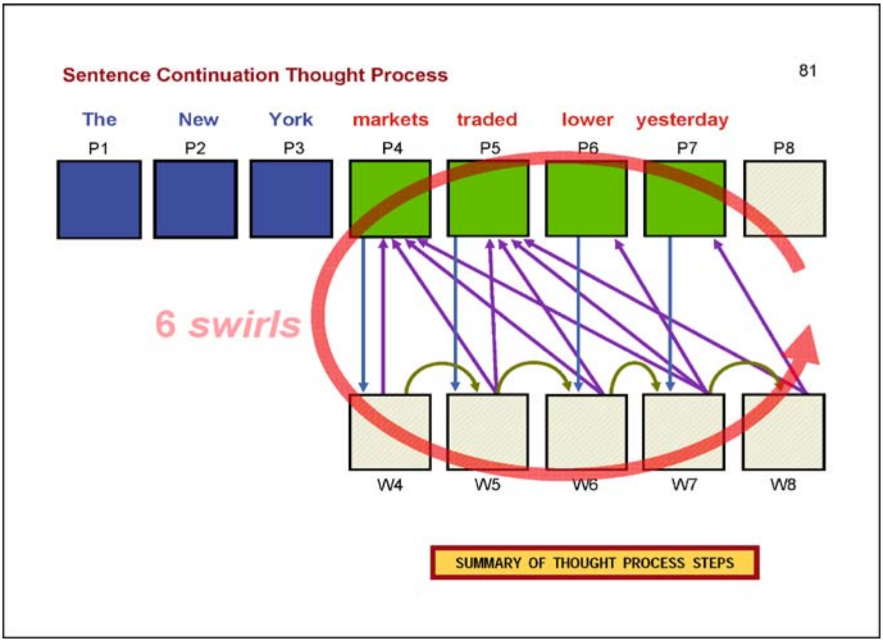
starter → The New York

W1 W2 W3 W4 W5 W6 W7

Stocks proved to be a wise investment .

context sentence

THOUGHT PROCESS STEP 2
CONCRETE EXAMPLE #3



Sentence Continuation Thought Process

81

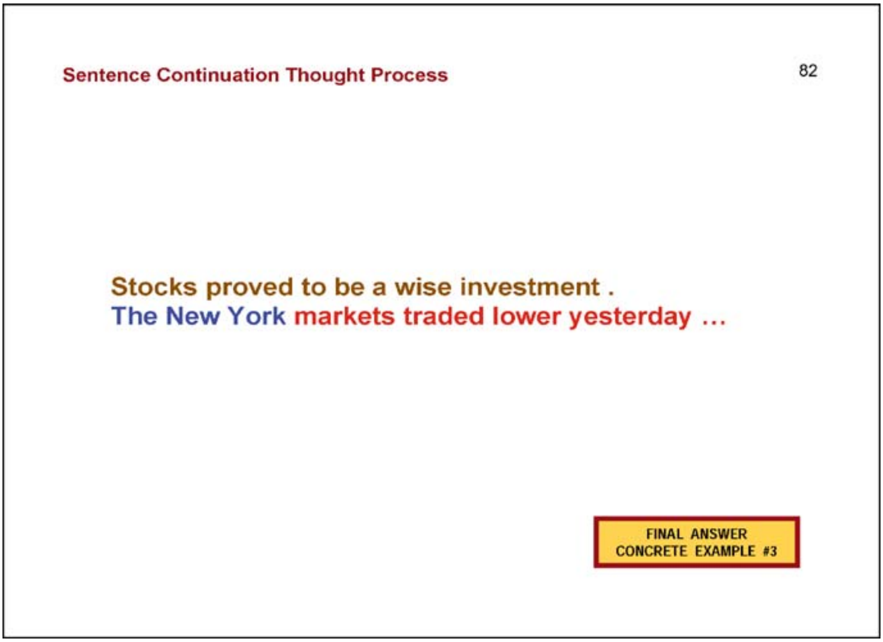
The New York markets traded lower yesterday

P1 P2 P3 P4 P5 P6 P7 P8

6 swirls

W4 W5 W6 W7 W8

SUMMARY OF THOUGHT PROCESS STEPS

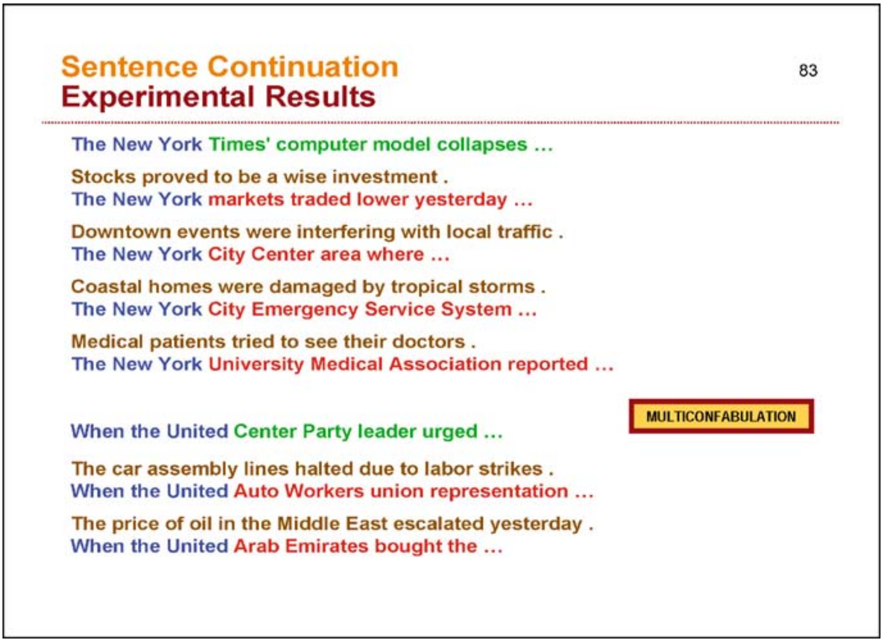


Sentence Continuation Thought Process

82

Stocks proved to be a wise investment .
The New York markets traded lower yesterday …

FINAL ANSWER
CONCRETE EXAMPLE #3



Sentence Continuation
Experimental Results

83

The New York Times' computer model collapses …

Stocks proved to be a wise investment .
The New York markets traded lower yesterday …

Downtown events were interfering with local traffic .
The New York City Center area where …

Coastal homes were damaged by tropical storms .
The New York City Emergency Service System …

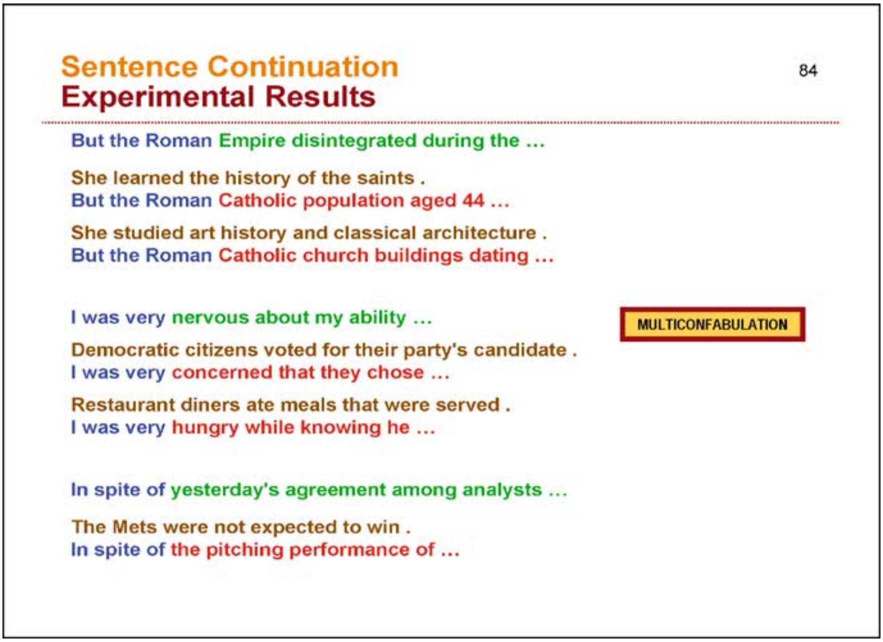
Medical patients tried to see their doctors .
The New York University Medical Association reported …

MULTICONFABULATION

When the United Center Party leader urged …

The car assembly lines halted due to labor strikes .
When the United Auto Workers union representation …

The price of oil in the Middle East escalated yesterday .
When the United Arab Emirates bought the …



Sentence Continuation
Experimental Results

84

But the Roman Empire disintegrated during the …

She learned the history of the saints .
But the Roman Catholic population aged 44 …

She studied art history and classical architecture .
But the Roman Catholic church buildings dating …

MULTICONFABULATION

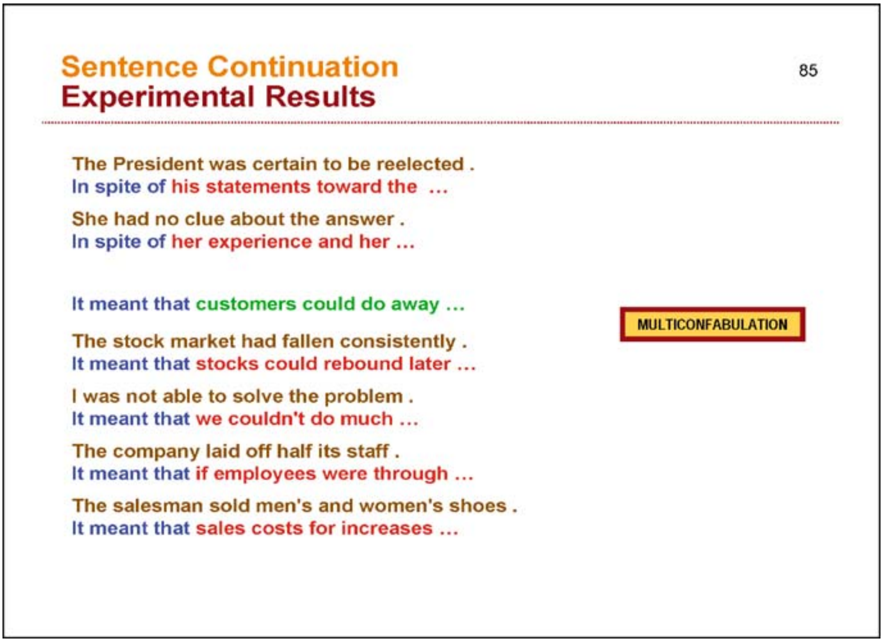
I was very nervous about my ability …

Democratic citizens voted for their party's candidate .
I was very concerned that they chose …

Restaurant diners ate meals that were served .
I was very hungry while knowing he …

In spite of yesterday's agreement among analysts …

The Mets were not expected to win .
In spite of the pitching performance of …



Sentence Continuation
Experimental Results

85

The President was certain to be reelected .
In spite of his statements toward the …

She had no clue about the answer .
In spite of her experience and her …

It meant that customers could do away …

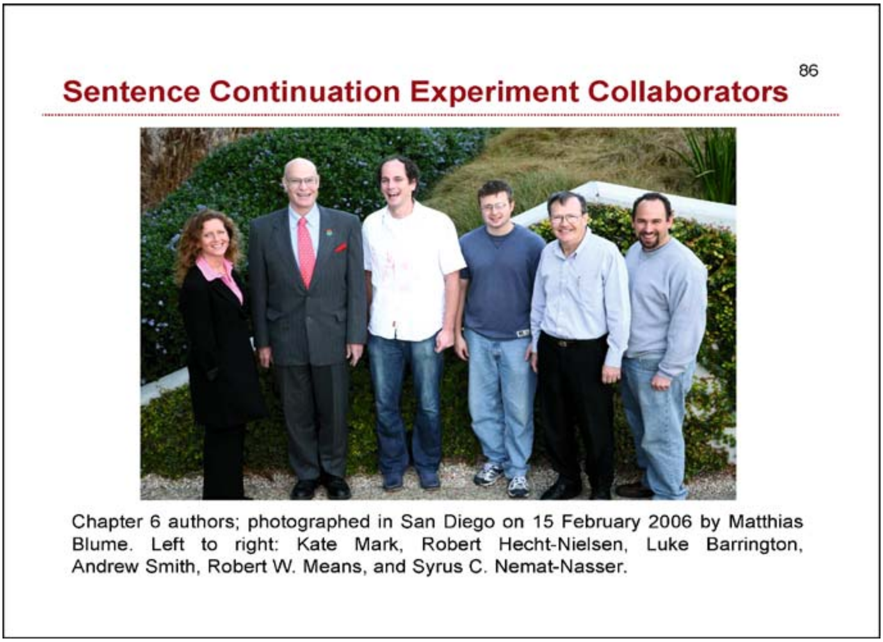
The stock market had fallen consistently .
It meant that stocks could rebound later …

I was not able to solve the problem .
It meant that we couldn't do much …

The company laid off half its staff .
It meant that if employees were through …

The salesman sold men's and women's shoes .
It meant that sales costs for increases …

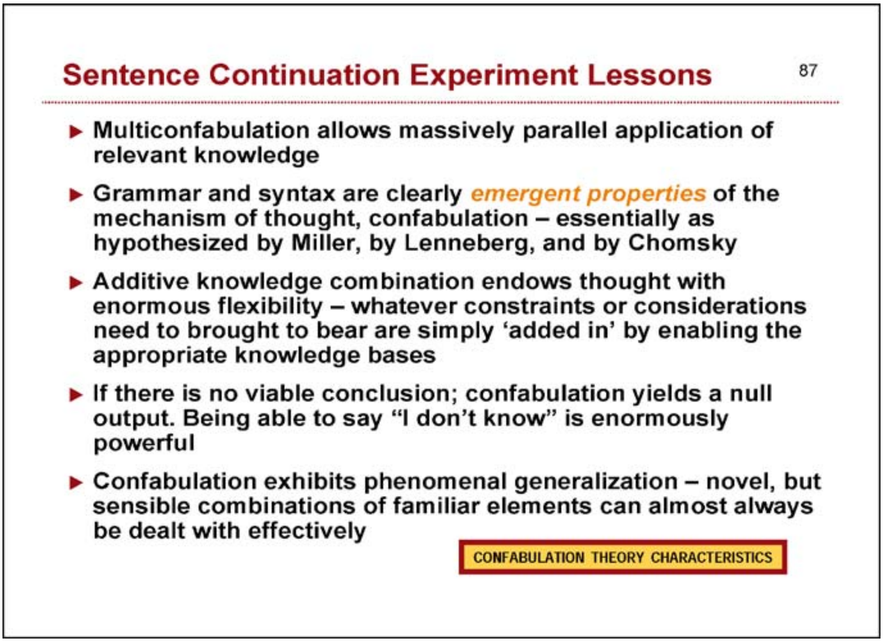
MULTICONFABULATION



Sentence Continuation Experiment Collaborators

86

Chapter 6 authors; photographed in San Diego on 15 February 2006 by Matthias Blume. Left to right: Kate Mark, Robert Hecht-Nielsen, Luke Barrington, Andrew Smith, Robert W. Means, and Syrus C. Nemat-Nasser.

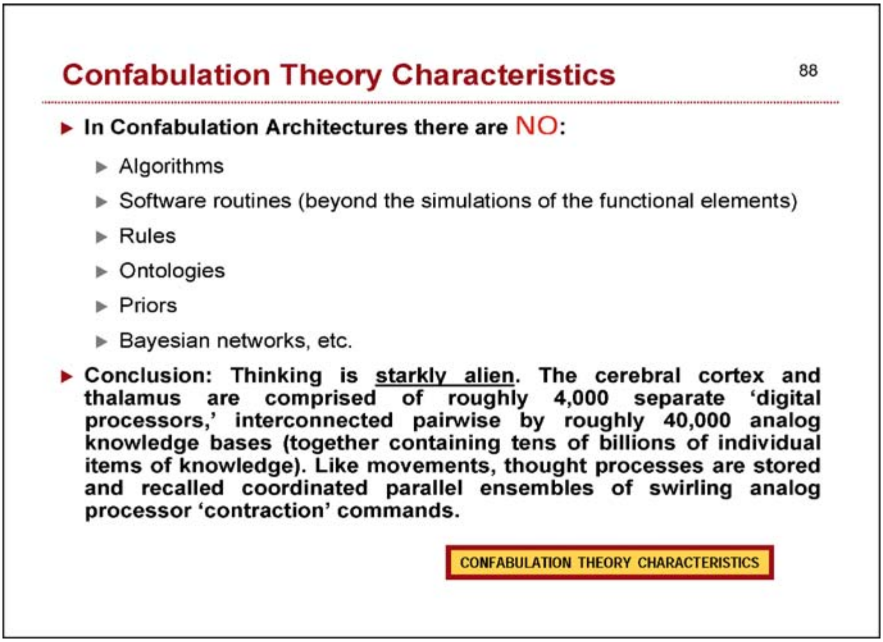


Sentence Continuation Experiment Lessons

87

- Multiconfabulation allows massively parallel application of relevant knowledge
- Grammar and syntax are clearly *emergent properties* of the mechanism of thought, confabulation – essentially as hypothesized by Miller, by Lenneberg, and by Chomsky
- Additive knowledge combination endows thought with enormous flexibility – whatever constraints or considerations need to brought to bear are simply ‘added in’ by enabling the appropriate knowledge bases
- If there is no viable conclusion; confabulation yields a null output. Being able to say “I don’t know” is enormously powerful
- Confabulation exhibits phenomenal generalization – novel, but sensible combinations of familiar elements can almost always be dealt with effectively

CONFABULATION THEORY CHARACTERISTICS

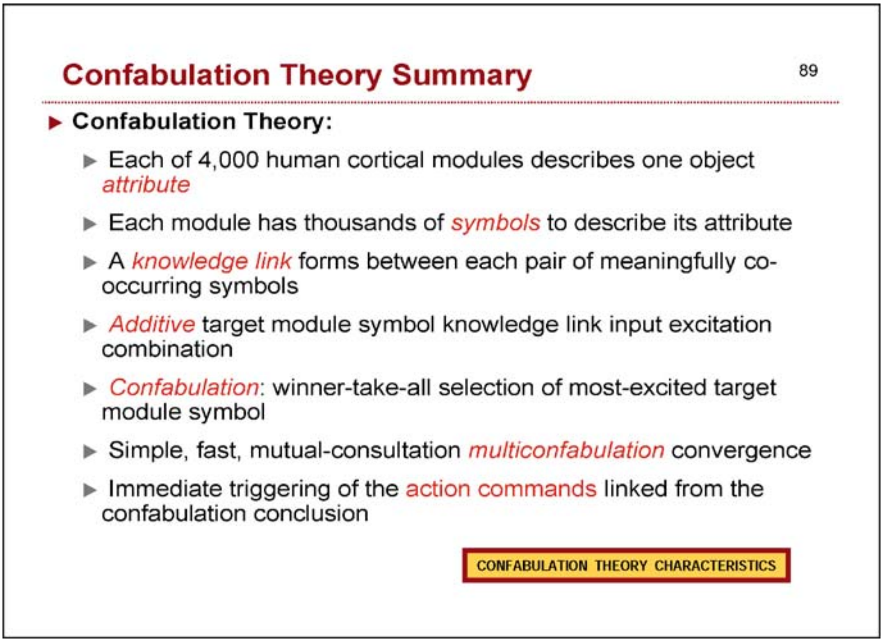


Confabulation Theory Characteristics

88

- **In Confabulation Architectures there are NO:**
  - Algorithms
  - Software routines (beyond the simulations of the functional elements)
  - Rules
  - Ontologies
  - Priors
  - Bayesian networks, etc.
- **Conclusion: Thinking is starkly alien. The cerebral cortex and thalamus are comprised of roughly 4,000 separate ‘digital processors,’ interconnected pairwise by roughly 40,000 analog knowledge bases (together containing tens of billions of individual items of knowledge). Like movements, thought processes are stored and recalled coordinated parallel ensembles of swirling analog processor ‘contraction’ commands.**

CONFABULATION THEORY CHARACTERISTICS

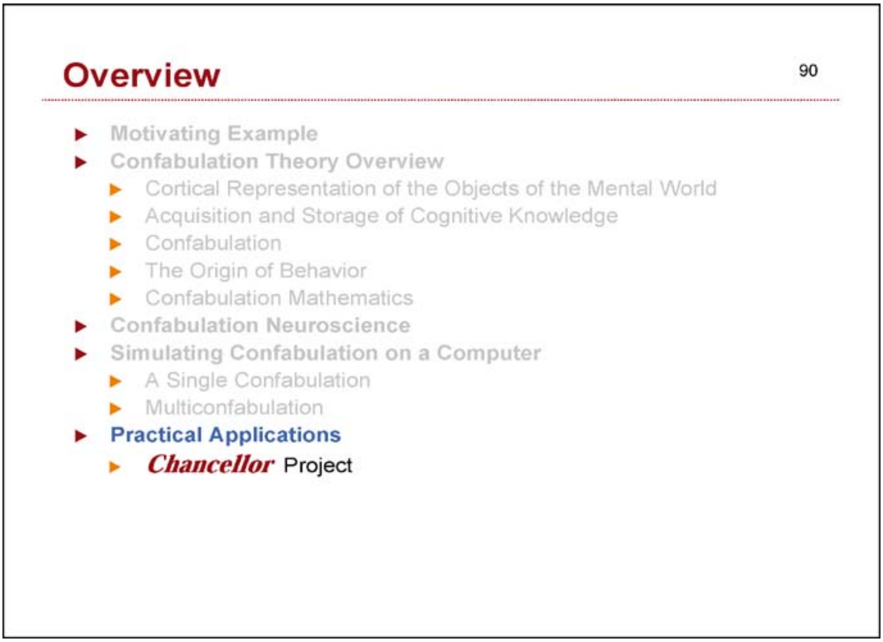


Confabulation Theory Summary

89

- **Confabulation Theory:**
  - Each of 4,000 human cortical modules describes one object *attribute*
  - Each module has thousands of *symbols* to describe its attribute
  - A *knowledge link* forms between each pair of meaningfully co-occurring symbols
  - *Additive* target module symbol knowledge link input excitation combination
  - *Confabulation*: winner-take-all selection of most-excited target module symbol
  - Simple, fast, mutual-consultation *multiconfabulation* convergence
  - Immediate triggering of the action commands linked from the confabulation conclusion

CONFABULATION THEORY CHARACTERISTICS



Overview

90

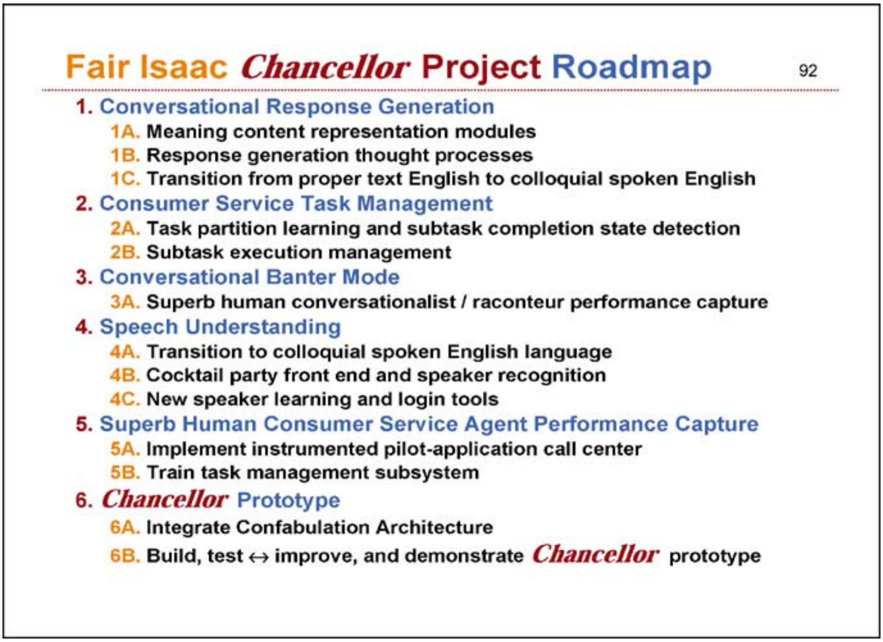
- **Motivating Example**
- **Confabulation Theory Overview**
  - Cortical Representation of the Objects of the Mental World
  - Acquisition and Storage of Cognitive Knowledge
  - Confabulation
  - The Origin of Behavior
  - Confabulation Mathematics
- **Confabulation Neuroscience**
- **Simulating Confabulation on a Computer**
  - A Single Confabulation
  - Multiconfabulation
- **Practical Applications**
  - ***Chancellor*** Project

Chancellor™ – Cat Food for Zeus

91

Chancellor

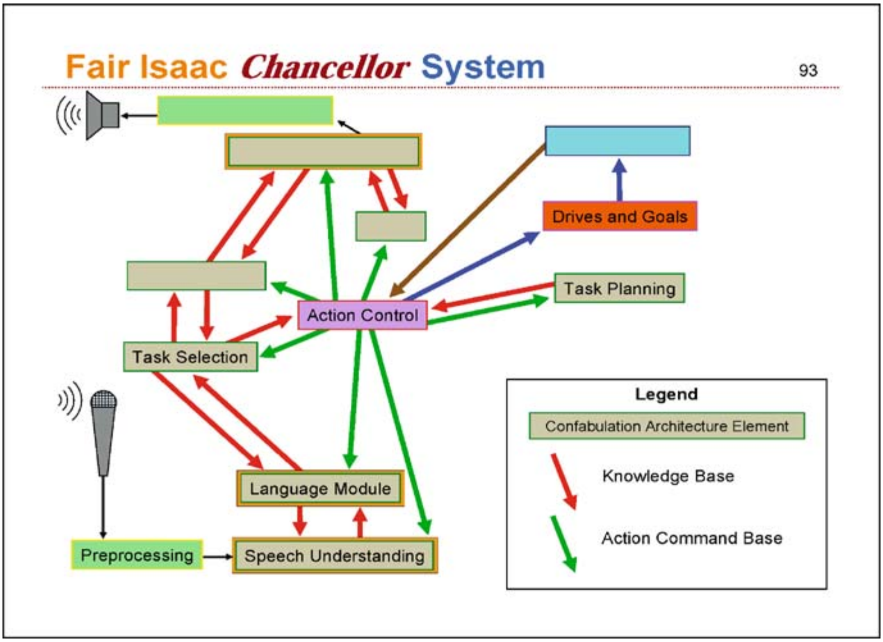
Chancellor is a trademark of Fair Isaac Corporation.



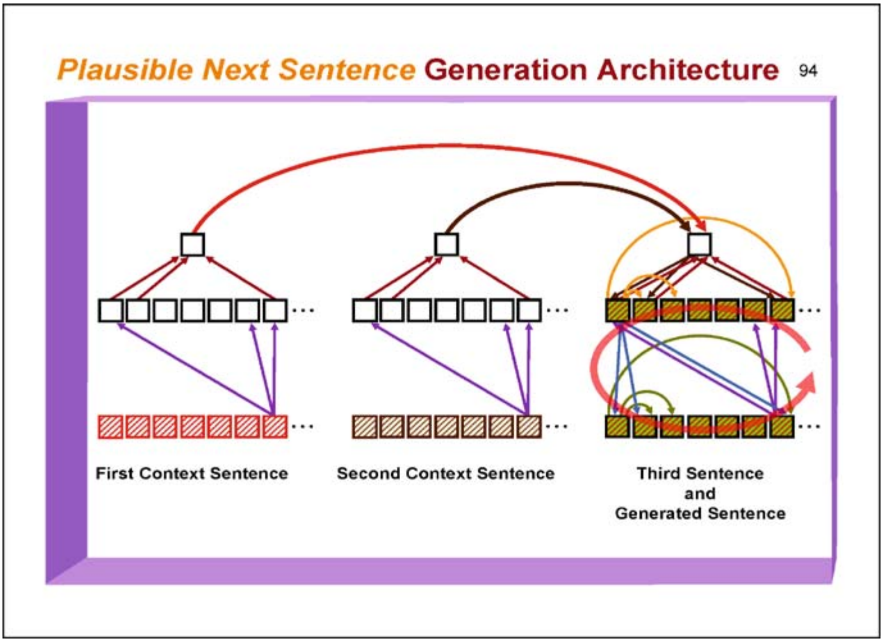
Fair Isaac *Chancellor* Project Roadmap

92

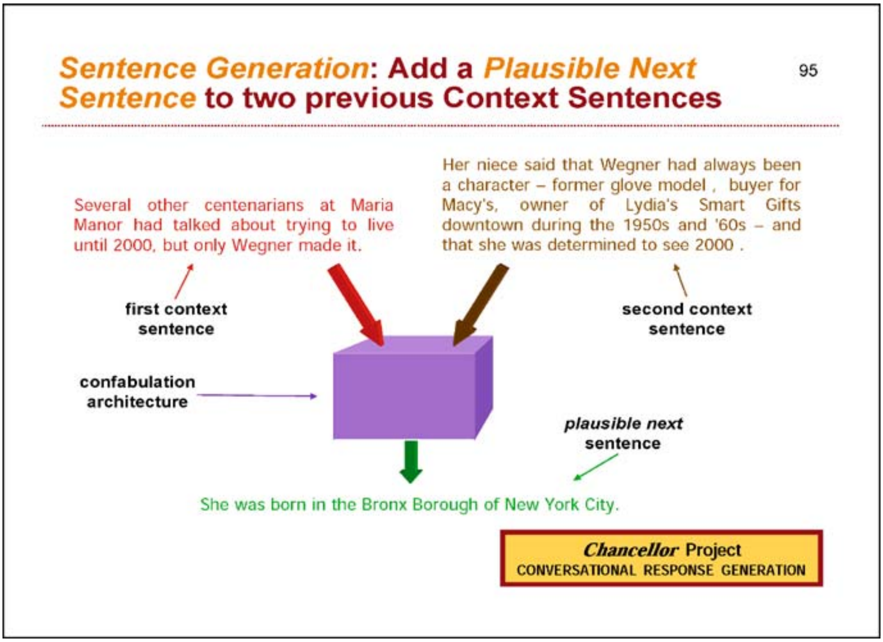
1. Conversational Response Generation
   - 1A. Meaning content representation modules
   - 1B. Response generation thought processes
   - 1C. Transition from proper text English to colloquial spoken English
2. Consumer Service Task Management
   - 2A. Task partition learning and subtask completion state detection
   - 2B. Subtask execution management
3. Conversational Banter Mode
   - 3A. Superb human conversationalist / raconteur performance capture
4. Speech Understanding
   - 4A. Transition to colloquial spoken English language
   - 4B. Cocktail party front end and speaker recognition
   - 4C. New speaker learning and login tools
5. Superb Human Consumer Service Agent Performance Capture
   - 5A. Implement instrumented pilot-application call center
   - 5B. Train task management subsystem
6. *Chancellor* Prototype
   - 6A. Integrate Confabulation Architecture
   - 6B. Build, test ↔ improve, and demonstrate *Chancellor* prototype



Fair Isaac *Chancellor* System 93



***Plausible Next Sentence* Generation Architecture** 94



Sentence Generation: Add a *Plausible Next Sentence* to two previous Context Sentences

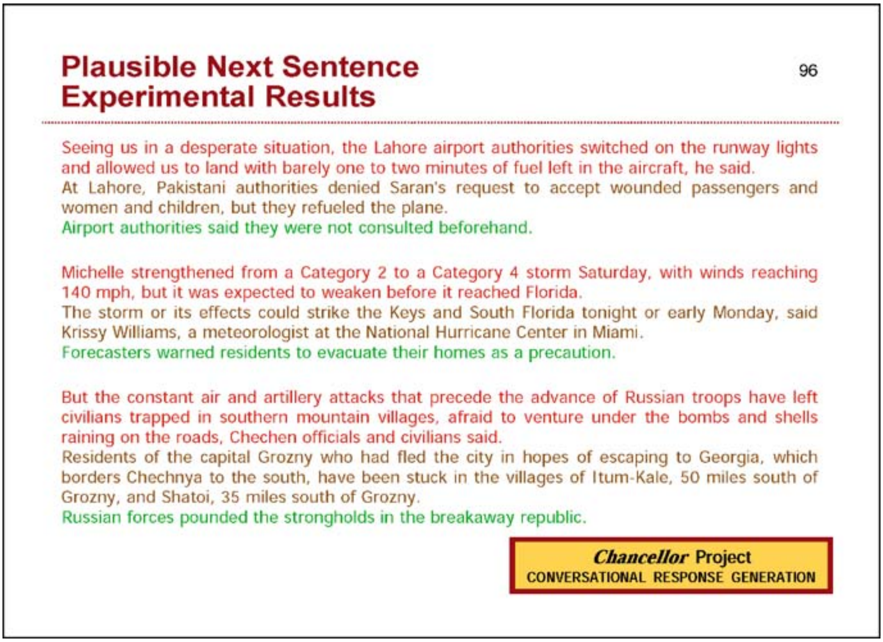
95

Several other centenarians at Maria Manor had talked about trying to live until 2000, but only Wegner made it.

Her niece said that Wegner had always been a character – former glove model , buyer for Macy's, owner of Lydia's Smart Gifts downtown during the 1950s and '60s – and that she was determined to see 2000 .

She was born in the Bronx Borough of New York City.

Chancellor **Project**
CONVERSATIONAL RESPONSE GENERATION



Plausible Next Sentence Experimental Results

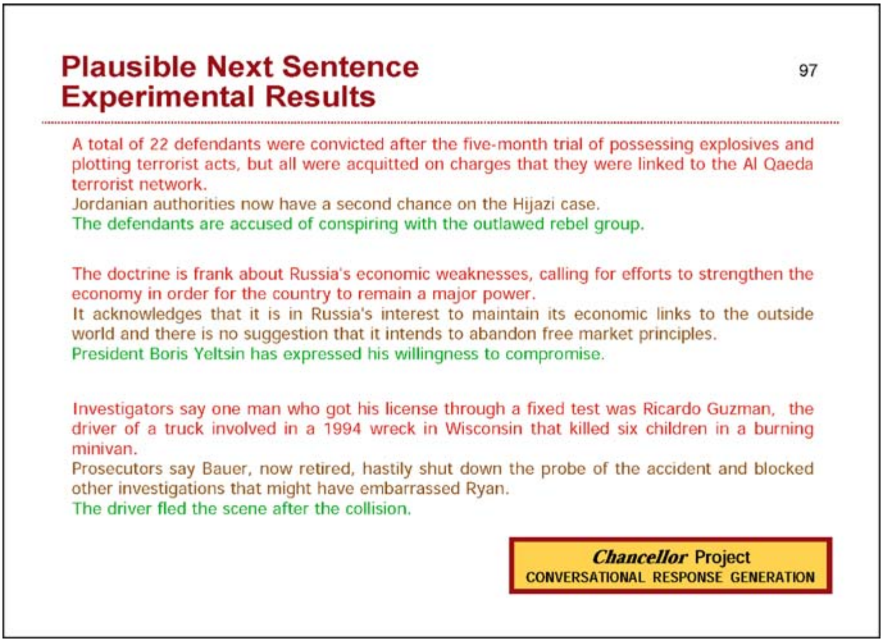
96

Seeing us in a desperate situation, the Lahore airport authorities switched on the runway lights and allowed us to land with barely one to two minutes of fuel left in the aircraft, he said.
At Lahore, Pakistani authorities denied Saran's request to accept wounded passengers and women and children, but they refueled the plane.
Airport authorities said they were not consulted beforehand.

Michelle strengthened from a Category 2 to a Category 4 storm Saturday, with winds reaching 140 mph, but it was expected to weaken before it reached Florida.
The storm or its effects could strike the Keys and South Florida tonight or early Monday, said Krissy Williams, a meteorologist at the National Hurricane Center in Miami.
Forecasters warned residents to evacuate their homes as a precaution.

But the constant air and artillery attacks that precede the advance of Russian troops have left civilians trapped in southern mountain villages, afraid to venture under the bombs and shells raining on the roads, Chechen officials and civilians said.
Residents of the capital Grozny who had fled the city in hopes of escaping to Georgia, which borders Chechnya to the south, have been stuck in the villages of Itum-Kale, 50 miles south of Grozny, and Shatoi, 35 miles south of Grozny.
Russian forces pounded the strongholds in the breakaway republic.

Chancellor **Project**
CONVERSATIONAL RESPONSE GENERATION



Plausible Next Sentence Experimental Results

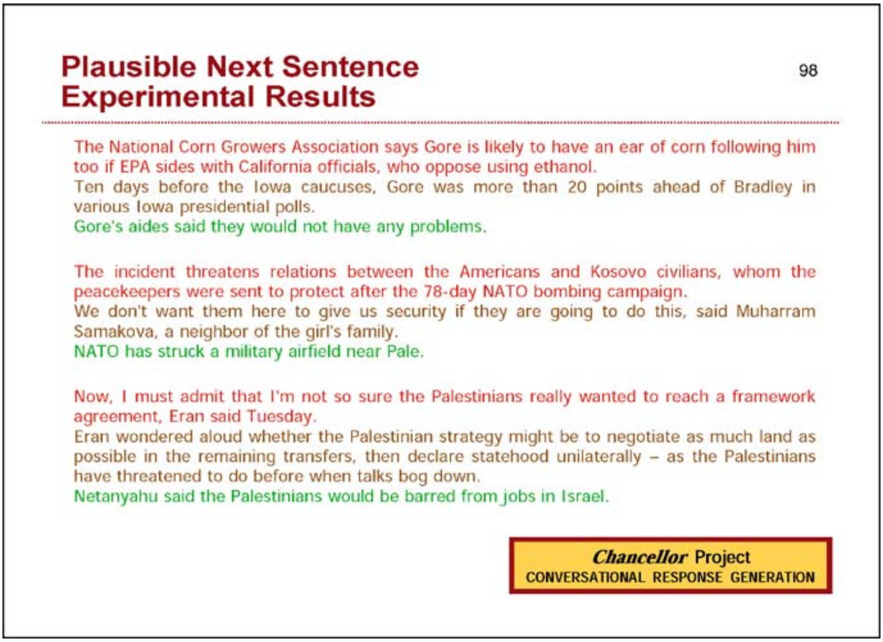
97

A total of 22 defendants were convicted after the five-month trial of possessing explosives and plotting terrorist acts, but all were acquitted on charges that they were linked to the Al Qaeda terrorist network.
Jordanian authorities now have a second chance on the Hijazi case.
The defendants are accused of conspiring with the outlawed rebel group.

The doctrine is frank about Russia's economic weaknesses, calling for efforts to strengthen the economy in order for the country to remain a major power.
It acknowledges that it is in Russia's interest to maintain its economic links to the outside world and there is no suggestion that it intends to abandon free market principles.
President Boris Yeltsin has expressed his willingness to compromise.

Investigators say one man who got his license through a fixed test was Ricardo Guzman, the driver of a truck involved in a 1994 wreck in Wisconsin that killed six children in a burning minivan.
Prosecutors say Bauer, now retired, hastily shut down the probe of the accident and blocked other investigations that might have embarrassed Ryan.
The driver fled the scene after the collision.

Chancellor **Project**
CONVERSATIONAL RESPONSE GENERATION



Plausible Next Sentence Experimental Results

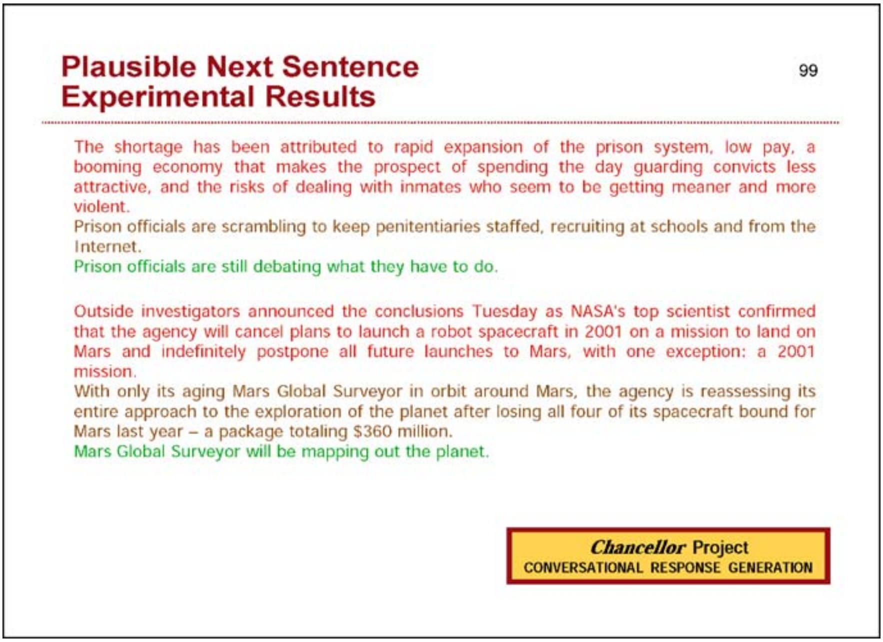
98

The National Corn Growers Association says Gore is likely to have an ear of corn following him too if EPA sides with California officials, who oppose using ethanol.
Ten days before the Iowa caucuses, Gore was more than 20 points ahead of Bradley in various Iowa presidential polls.
Gore's aides said they would not have any problems.

The incident threatens relations between the Americans and Kosovo civilians, whom the peacekeepers were sent to protect after the 78-day NATO bombing campaign.
We don't want them here to give us security if they are going to do this, said Muharram Samakova, a neighbor of the girl's family.
NATO has struck a military airfield near Pale.

Now, I must admit that I'm not so sure the Palestinians really wanted to reach a framework agreement, Eran said Tuesday.
Eran wondered aloud whether the Palestinian strategy might be to negotiate as much land as possible in the remaining transfers, then declare statehood unilaterally – as the Palestinians have threatened to do before when talks bog down.
Netanyahu said the Palestinians would be barred from jobs in Israel.

Chancellor **Project**
CONVERSATIONAL RESPONSE GENERATION



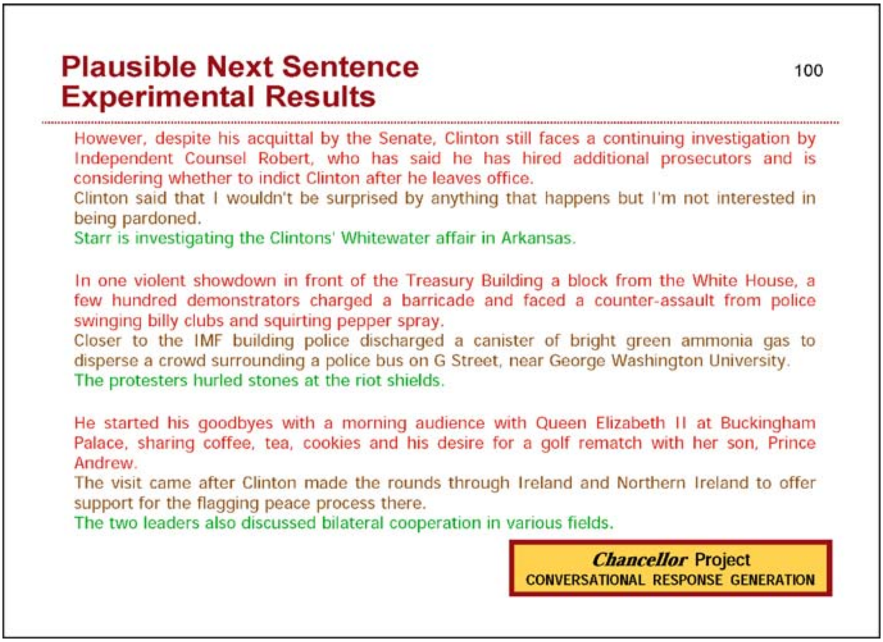
Plausible Next Sentence Experimental Results

99

The shortage has been attributed to rapid expansion of the prison system, low pay, a booming economy that makes the prospect of spending the day guarding convicts less attractive, and the risks of dealing with inmates who seem to be getting meaner and more violent.
Prison officials are scrambling to keep penitentiaries staffed, recruiting at schools and from the Internet.
Prison officials are still debating what they have to do.

Outside investigators announced the conclusions Tuesday as NASA's top scientist confirmed that the agency will cancel plans to launch a robot spacecraft in 2001 on a mission to land on Mars and indefinitely postpone all future launches to Mars, with one exception: a 2001 mission.
With only its aging Mars Global Surveyor in orbit around Mars, the agency is reassessing its entire approach to the exploration of the planet after losing all four of its spacecraft bound for Mars last year – a package totaling $360 million.
Mars Global Surveyor will be mapping out the planet.

Chancellor Project
CONVERSATIONAL RESPONSE GENERATION



Plausible Next Sentence Experimental Results

100

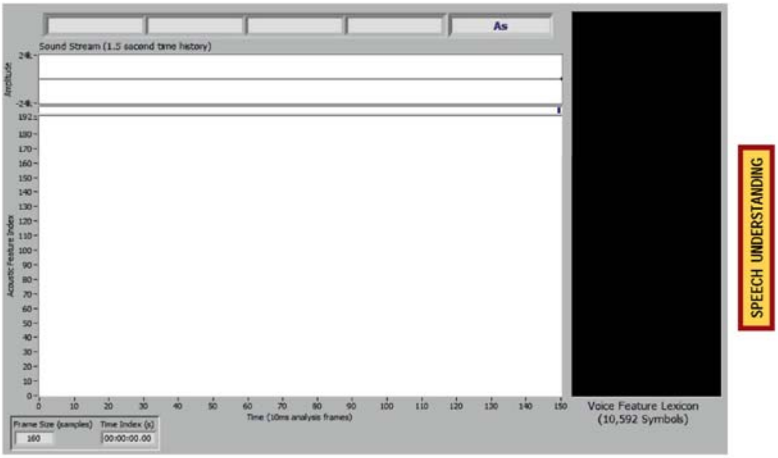
However, despite his acquittal by the Senate, Clinton still faces a continuing investigation by Independent Counsel Robert, who has said he has hired additional prosecutors and is considering whether to indict Clinton after he leaves office.
Clinton said that I wouldn't be surprised by anything that happens but I'm not interested in being pardoned.
Starr is investigating the Clintons' Whitewater affair in Arkansas.

In one violent showdown in front of the Treasury Building a block from the White House, a few hundred demonstrators charged a barricade and faced a counter-assault from police swinging billy clubs and squirting pepper spray.
Closer to the IMF building police discharged a canister of bright green ammonia gas to disperse a crowd surrounding a police bus on G Street, near George Washington University.
The protesters hurled stones at the riot shields.

He started his goodbyes with a morning audience with Queen Elizabeth II at Buckingham Palace, sharing coffee, tea, cookies and his desire for a golf rematch with her son, Prince Andrew.
The visit came after Clinton made the rounds through Ireland and Northern Ireland to offer support for the flagging peace process there.
The two leaders also discussed bilateral cooperation in various fields.

Chancellor Project
CONVERSATIONAL RESPONSE GENERATION

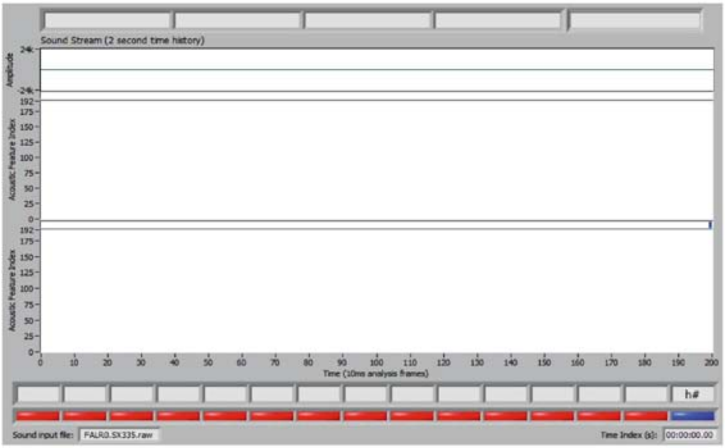
Speech Understanding Confabulation Architecture 101



Frame Size (samples): 160

Time Index (s): 00:00:00.00

Speechstream Segmentation 102



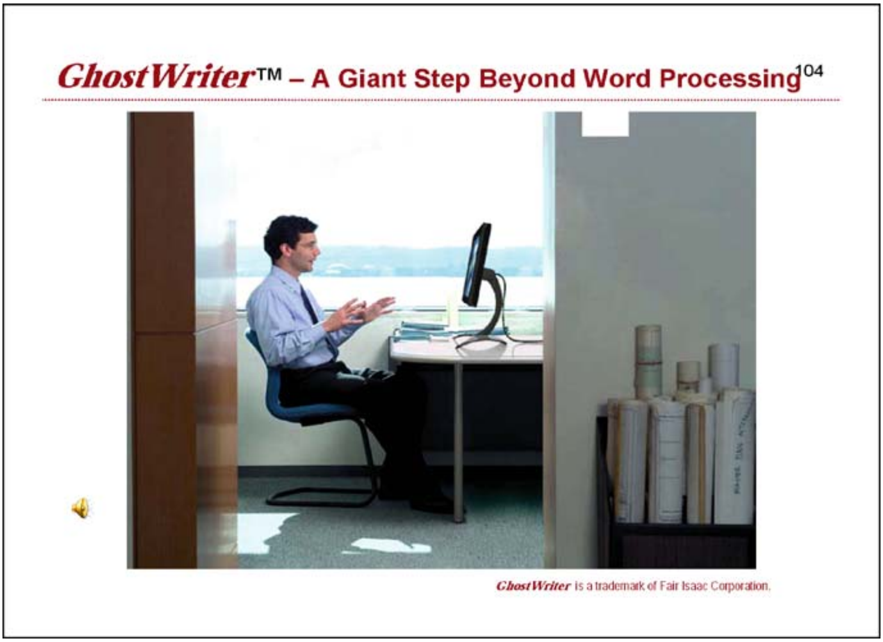
Sound input file: FALR0.SX335.raw

Time Index (s): 00:00:00.00

Chancellor **– Cat Food for Zeus**

103

Chancellor



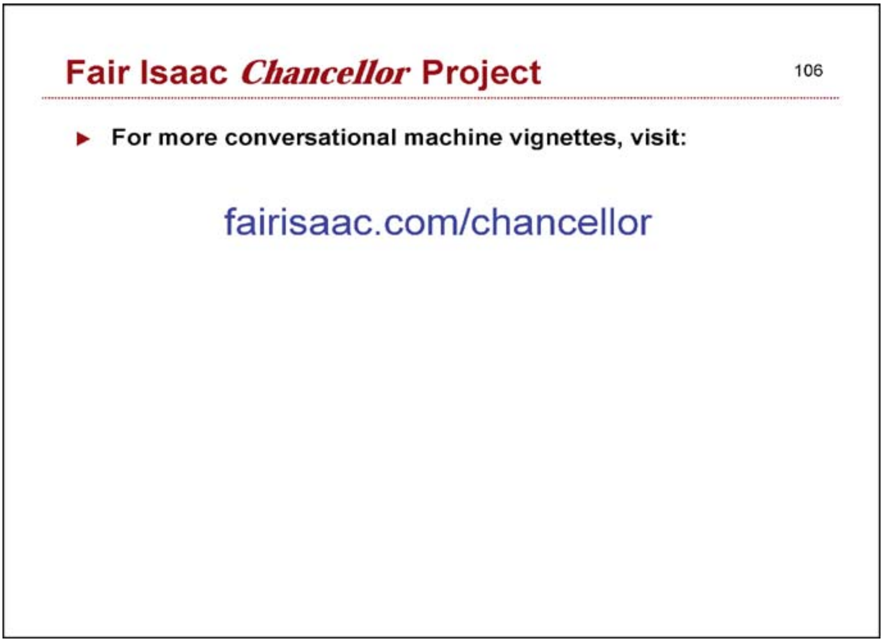
***GhostWriter*™ – A Giant Step Beyond Word Processing**

104

GhostWriter is a trademark of Fair Isaac Corporation.

Chancellor – Shopping for a Handbag at *EarthMall*™ 105

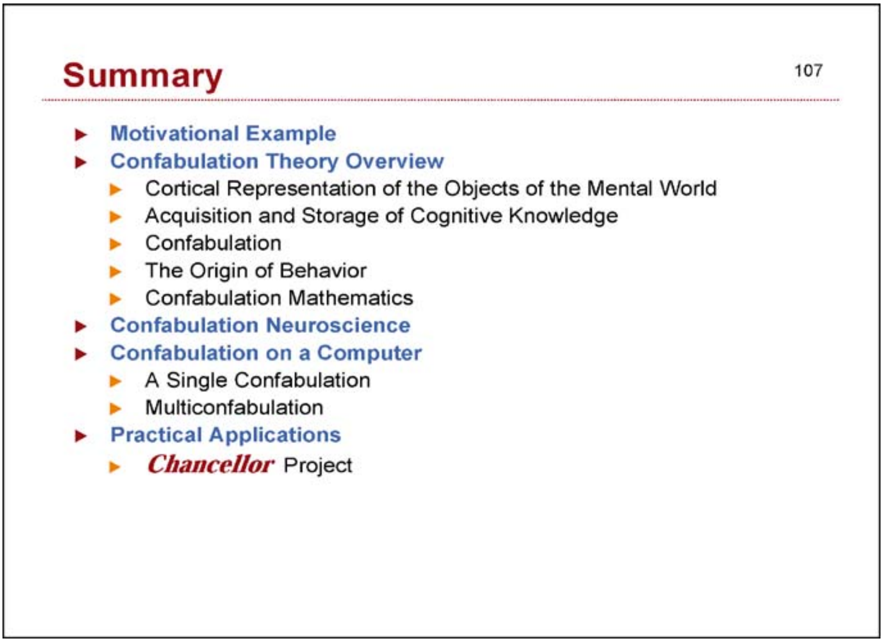
EarthMall is a trademark of Fair Isaac Corporation.



Fair Isaac *Chancellor* Project 106

- For more conversational machine vignettes, visit:

fairisaac.com/chancellor



Summary

107

- Motivational Example
- Confabulation Theory Overview
  - Cortical Representation of the Objects of the Mental World
  - Acquisition and Storage of Cognitive Knowledge
  - Confabulation
  - The Origin of Behavior
  - Confabulation Mathematics
- Confabulation Neuroscience
- Confabulation on a Computer
  - A Single Confabulation
  - Multiconfabulation
- Practical Applications
  - ***Chancellor*** Project

Research Sponsors

108

The support of this research by:

FairIsaac
60 Years
ONR

Fair Isaac Corporation (NYSE: FIC)

Office of Naval Research

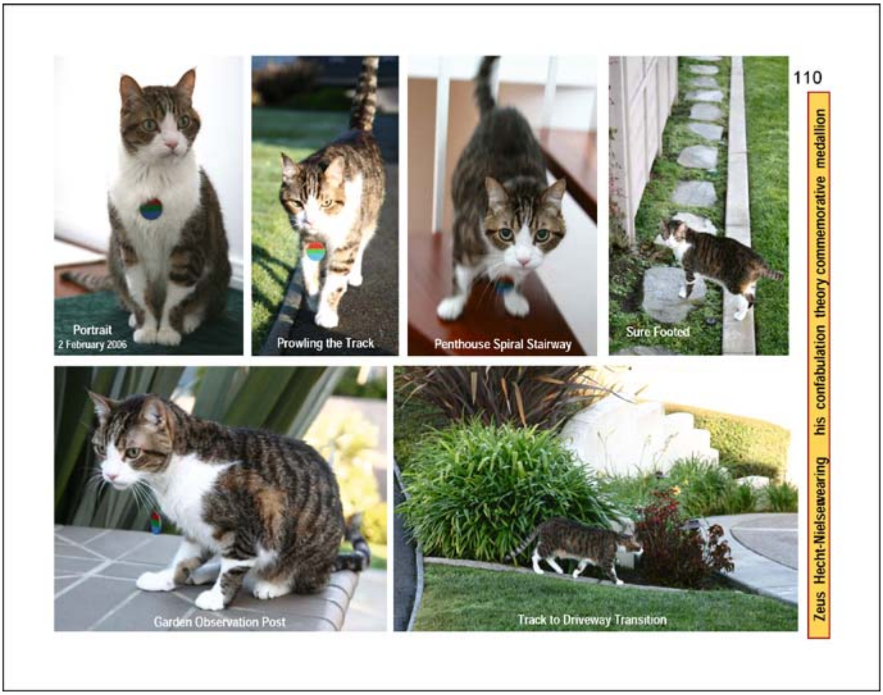
is gratefully acknowledged.

Domestic Cat Zeus Hecht-Nielsen 109

Co-Investigator, Behavioral Observation Subject, Muse

An Indispensable Contributor to the Confabulation Discovery

Official Portrait
(shown with his favorite toy – a purple string)
Tranquility Base Here
Prowling the Track
(Hecht-Nielsen Compound – Park Side)
Rolling Behavior on the Track
Awaiting Ingress
(Hecht-Nielsen Compound – House Side)
Breakfast Time



110

Zeus Hecht-Nielsewearing his confabulation theory commemorative medallion

3 The Mathematics of Thought[3]

3.1 The Constructs of Confabulation Mathematics

Confabulation theory (with which the reader is assumed to be somewhat familiar from Chaps. 1 and 2 and the video presentation) is based upon four mathematical constructs:

1. A collection of N finite sets of *symbols* (each such set is termed a *thalamocortical module*).
2. A directed R +– weighted graph having all of the symbols of all of the modules as its nodes. Each edge of the graph is termed a *knowledge link* or *item of knowledge*.
3. A “winners-take-all” intersymbol competition operation (termed *confabulation*) which is carried out within a module over a finite time span – in accordance with an externally supplied *thought-command signal*.
4. A mapping (termed *skill knowledge*) between each symbol of a module and a subset of the set of *action commands* associated with that module.

Since the fourth of these constructs is not strictly a part of animal cognition, i.e., thought (the focus of this exposition), skill knowledge and action commands will not be discussed further here.

The next section provides a look at the basic essence of confabulation mathematics by considering the case of a single, isolated, confabulation operation. However, as with almost any aspect of biology, there are many refinements, embellishments, and improvements that have been added onto this basic framework since its initial evolutionary emergence hundreds of millions of years ago. Sections 3.3 through 3.7 consider a few of these. For expositional simplicity, and because the author believes them to be roughly correct, the tenets, positions, and views of confabulation theory will be presented as if they were facts.

[3] This chapter is based on the original publication Hecht-Nielsen R (2006) The mathematics of thought. In: Yen GY, Fogel DB (eds) Computational intelligence: Principles and practice. IEEE Computational Intelligence Society, Piscataway, NJ, pp 1–16, and is adapted here in accordance with IEEE copyrights.

3.2 A Single Confabulation

The most basic cognitive information-processing operation (thought process) consists of a single "pure" confabulation. This section describes the mathematics of this simplest case.

During wakefulness, the cognitive brain is constantly functioning. When a single symbol is made active as the result of winning a completed confabulation competition (details to follow below), that symbol often then serves, briefly, as an input to subsequent confabulations, via knowledge links emanating from it (see Fig. 3.1). A winning symbol which has faded can be briefly restored by means of a confabulation with no inputs, and then used as an assumed fact input to subsequent confabulations (as long as the module has not been used since the symbol faded). This ability to briefly restore a module's last-active symbol with a confabulation is called *short-term* (or *working*) *memory*.

Figure 3.1 considers four modules (on the left) that have recently been confabulated, each now *expressing* one active symbol (the generalization to an arbitrary number of modules is obvious). Label these symbols α, β, γ, and δ. A fifth module (on the right) is about to undergo confabulation, based upon knowledge link inputs to its symbols from these four *assumed fact* symbols α, β, γ, and δ.

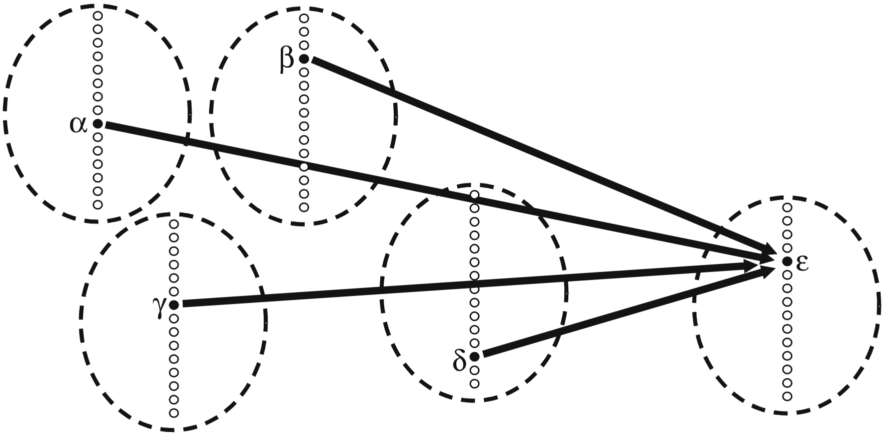


Fig. 3.1. Five *thalamocortical modules.* Four items of knowledge are shown, connecting, respectively, symbols α, β, γ, and δ (each belonging to one of the modules on the left) to symbol ε of the module on the right (which is about to undergo *confabulation*). Knowledge link inputs from *source symbols* α, β, γ, and δ (also referred to as *assumed facts*) to symbols of this fifth module (each assumed fact symbol often sources hundreds of knowledge links) cause these *target symbols* on the fifth module to become excited (knowledge link input excitations are additive). The target symbol with the highest input excitation – here symbol ε – will win the confabulation competition and be made *active.* ε can then serve as an assumed fact for subsequent confabulations on other modules

If an item of knowledge connects assumed fact symbol α to symbol λ of the fifth module, its *weight* is

$$\log_c(p(\alpha|\lambda)/p_0) + B, \tag{3.1}$$

where $p(\alpha|\lambda)$ is the *antecedent support probability* (see Chaps. 4 and 7 for details) that the link's *source symbol* α will be active, given the assumption that the link's *target symbol* λ is active, and c, p_0 and B are positive constants. As discussed further in Sect. 3.5, this link weight form was chosen by evolution because:

1. These weights can be implemented by neuronal synapses.
2. Summation of excitations delivered by knowledge links is a surrogate computation which approximates the general-purpose "conclusion quality measure" cogency (explained below).

As discussed further in Sect. 3.5, knowledge links form on the basis that the involved ordered pair of symbols are observed to meaningfully co-occur as the brain responds to the information environment in which it lives. Only certain externally specified ordered pairs of modules can develop knowledge links between their symbols (in the brain these ordered pairs are determined by the genetically controlled axonal "wiring pattern"). The set of all knowledge links proceeding from symbols of one module to symbols of another is termed a *knowledge base.*

The *input excitation* I(λ) received by symbol λ of the fifth module of Fig. 3.1 from knowledge links from the four assumed fact symbols α, β, γ, and δ is given by the sum of the weights of the involved knowledge links (the four terms in brackets):

$$\begin{aligned} I(\lambda) \equiv {} & [\log_c(p(\alpha|\lambda)/p_0) + B] \\ & + [\log_c(p(\beta|\lambda)/p_0) + B] \\ & + [\log_c(p(\gamma|\lambda)/p_0) + B] \\ & + [\log_c(p(\delta|\lambda)/p_0) + B] \end{aligned} \tag{3.2}$$

(with the convention that if a knowledge link does not exist from one or more of the assumed facts to λ, the nonexistent "knowledge link" has weight zero).

Confabulation of the fifth, or *answer*, module of Fig. 3.1 then consists of simply selecting that symbol ε having the highest input excitation. This symbol ε is termed the *conclusion* of the confabulation.

Since the logarithm is a strictly monotonically increasing function, maximizing I(λ) yields the same conclusion as maximizing the *confabulation product* $p(\alpha|\lambda) \cdot p(\beta|\lambda) \cdot p(\gamma|\lambda) \cdot p(\delta|\lambda)$. The significance of this fact is elucidated by:

Theorem 3.1: *The Fundamental Theorem of Cognition* (see Chap. 4): Given nonexceptional assumed facts α, β, γ, and δ, and *viable* answer module symbol λ, then the following exact relationship holds between *cogency* $p(\alpha\beta\gamma\delta|\lambda)$ (where juxtaposition of symbols means AND) and the *confabulation product* $p(\alpha|\lambda) \cdot p(\beta|\lambda) \cdot p(\gamma|\lambda) \cdot p(\delta|\lambda)$:

$$\begin{aligned}[p(\alpha\beta\gamma\delta|\lambda)]^4 = {} & [p(\alpha\beta\gamma\delta\varepsilon)/p(\alpha\lambda)] \\ & \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\beta\lambda)] \\ & \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\gamma\lambda)] \\ & \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\delta\lambda)] \\ & \cdot [p(\alpha|\lambda) \cdot p(\beta|\lambda) \cdot p(\gamma|\lambda) \cdot p(\delta|\lambda)]. \qquad \square\end{aligned}$$

Confabulation theory postulates that animal neurological evolution has found ways to ensure that the product of the first four bracketed quantities of the right-hand side of this equation is approximately constant for all *viable* conclusions λ (conclusions receiving non-zero-weighted knowledge links from all four of the assumed facts α, β, γ, and δ). Given this assumption, *confabulation* (i.e., maximizing the confabulation product $p(\alpha|\lambda) \cdot p(\beta|\lambda) \cdot p(\gamma|\lambda) \cdot p(\delta|\lambda)$) is clearly approximately equivalent to maximizing cogency $p(\alpha\beta\gamma\delta|\lambda)$.

Maximization of cogency (a quantity often referred to in other mathematical contexts as likelihood [Duda et al. 2000]) is postulated by confabulation theory to be the central underlying mathematical foundation of all aspects of cognition (seeing, hearing, language, reasoning, planning, control of movement and thought, etc.).

One example of the value of cogency maximization can be seen in:

Theorem 3.2 (see Chap. 4): If $\alpha\beta\gamma\delta \Rightarrow \varepsilon$ uniquely, then $\varepsilon = \lambda$ uniquely maximizes cogency $p(\alpha\beta\gamma\delta|\lambda)$. □

Thus, when considering a set of assumed facts which imply a unique conclusion (an event that would occur commonly when playing chess or doing mathematical proofs), cogency maximization yields Aristotelian logic. Thus, when logic pertains, animals behave logically. Of course, logical reasoning does not apply to many real-world situations (such as parking your car). In these instances, animals simply maximize cogency and go with the conclusions derived there from. As shown by the examples presented in Chaps. 4 and 6, the conclusions reached in this manner are, in general, quite serviceable.

Since the mathematics of confabulation is simple, an obvious question is: "Why wasn't confabulation theory discovered long ago?" A key reason is a decades-long intellectual constipation brought about by what might be called the "Bayesian religion."

The Bayesian religion is a dogmatic belief structure (often mistakenly viewed as a set of incontrovertible facts), currently held by perhaps 100,000 researchers and practitioners worldwide, underpinned by roughly the following seductive, compelling line of argument:

1. The "Bayes" pattern classifier (i.e., the classifier that maximizes *a posteriori* probability $p(\varepsilon|\alpha\beta\gamma\delta)$) is known to be mathematically optimal (lowest possible average error rate) (Duda et al. 2000). Note that the Bayes classifier chooses that class ε which has the highest probability of being correct, given the *pattern measurements* $\alpha\beta\gamma\delta$.
2. Animals are excellent pattern classifiers; so, in light of the proven mathematical fact 1, it seems obvious that animals must function as Bayesian pattern classifiers.
3. Given article of faith 2, animals are almost certainly also going to be "Bayesian" in other cognitive realms, meaning that the "best" conclusion ε to select in any situation will be the one which has the highest probability of being correct, given the available facts (i.e., cognition maximizes $p(\varepsilon|\alpha\beta\gamma\delta)$).

Since direct maximization of $p(\varepsilon|\alpha\beta\gamma\delta)$ is, in general, not practically feasible, Bayesians often use the *naïve Bayes formula:*

$$\begin{aligned} &\operatorname{argmax}(\varepsilon)\ p(\varepsilon|\alpha\beta\gamma\delta) \\ &\approx \operatorname{argmax}(\varepsilon)\ p(\varepsilon)\cdot p(\alpha|\varepsilon)\cdot p(\beta|\varepsilon)\cdot p(\gamma|\varepsilon)\cdot p(\delta|\varepsilon) \end{aligned} \tag{3.3}$$

as a surrogate. Note that the right-hand side of this formula is exactly the same as the confabulation product, but with an additional factor: the *prior* $p(\varepsilon)$ – the *a priori* probability of conclusion ε. An enormous number of practical applications of the naïve Bayes formula (see Wikipedia entries and literature cited therein) – a non-rigorous ad hoc mathematical relationship – have yielded excellent results, bolstering devoted belief in the Bayesian religion.

For some things, such as classical statistical pattern classification (Duda et al. 2000) and causal relationship analysis (Pearl 2000), the Bayesian approach may be the correct mathematical method to apply.

However, for explaining cognition, the Bayesian approach leads to architectural designs that seem irreconcilable with the facts of neuroscience. In effect, tenets 2 and 3 of the Bayesian religion (both of which are false – see Chap. 4) steered thousands of able researchers down the wrong trail (which yielded "$E = mc^3$"). Discovery of confabulation theory required a maverick mathematician following a decades-long circuitous research route (see Sect. 3.8) through the lands of cortical and thalamic neuroscience, neuronal symbol representation, synfire chain knowledge links, and neuronal attractor networks.

The fact that the naïve Bayes formula has been so successful is probably due to the fact that, in many practical "cognitive-type" problems, the prior values $p(\varepsilon)$ of the conclusions with the highest confabulation product values (many of which are often "reasonably high quality" answers) do not vary enough from one another to affect the quality of final choice much. Also, many times, heuristic kludges – such as excluding high-frequency and low-frequency words in linguistic processing – are used to eliminate unusually high or low value priors that would cause substantial damage. In effect, a great deal of "Bayesian analysis," as it is actually practiced, is cogency maximization.

3.3 Multiconfabulation

One of the most important discoveries within confabulation theory is of *multiconfabulation* (see Chaps. 6 and 7). A multiconfabulation is an ensemble of contemporaneous, mutually interacting, converging confabulations. The vast majority of human thought processes involve multiconfabulations. The mutual interactions which occur between the individual confabulations within a multiconfabulation often involve the simultaneous, parallel application of millions of items of *relevant* knowledge (only knowledge links emanating from viable conclusions are involved). In effect, the individual confabulations involved in the multiconfabulation constantly "compare notes" via knowledge link interactions as they continue to converge – ensuring mutual consistency between their final *consensus* of conclusions. This ensures that all known constraints and all available relevant knowledge are applied in crafting the final consensus. This massively parallel application of relevant knowledge during multiconfabulations is proposed as a primary explanation for the extreme information-processing effectiveness of thought.

The key to convergence of the individual confabulations during a multiconfabulation is twofold. First, the pattern in which knowledge links are applied in succession at first causes the excitation of large numbers of potential conclusion symbols. However, these knowledge link application patterns then "loop back," causing newly excited candidate conclusions to send knowledge link excitation to already-excited candidate conclusions (such a cyclic pattern of thought is termed *swirling*). The first crucial fact is that, on many occasions, many of the already-excited candidate conclusion symbols DO NOT receive significant "confirmatory" excitation from this subsequent knowledge link input (while some other candidate conclusion symbols **DO** receive "confirmatory" excitation). The list of candidate conclusion symbols is then shortened to eliminate those which did not receive confirmation (this list shortening is akin to "tightening" a muscle). [A brief aside: Fuzzy logic theorists might interpret some aspects of the initial brief "reordering" phase of swirling as *meaning fuzzification*. They might interpret the subsequent rapid convergence to the final "hard" consensus of conclusions as *defuzzification*. See the *Video Presentation Notes* for more details.]

As multiconfabulation proceeds, certain modules have their thought control input signals tightened more than others – causing their lists to shrink faster and, eventually, to converge to a single conclusion first. This order is part of the stored thought process. As convergences occur among the set of modules being confabulated; these converged modules have their final conclusion symbols *locked*. These locked symbols cause action commands to be launched, which, among other things, often cause other modules to be locked with no symbol expressed. For example, a language phrase module expressing the symbol for **New York** would cause the next phrase module to be locked in a null state (because it is not needed since **York** is already represented on the previous module – see Chap. 7 for details).

Multiconfabulation clearly illustrates the alien nature of thought. Thinking involves digital symbolic processors (modules) with different symbols having different graded excitation levels communicating via analog knowledge links being controlled by graded spatiotemporal thought maneuvers. This is about as far as one can imagine getting from Turing-style computation. The good news is that there will not be any need to agonize over Gödel paradoxes and exotic computability questions, because the "current state of the world-dependent," and therefore intrinsically unrepeatable, nature of thought processes means that thinking is inherently non-deterministic. Animal life is thereby endowed with exquisite unpredictability and spontaneity.

As technologists now proceed to develop artificial brains, the alien nature of cognition will introduce a variety of roadblocks. For one thing, much of what today's technologists know will be obsolete and largely inapplicable. Nonetheless, futile attempts to hybridize computer science-based algorithms with artificial cognition (akin to trying to build a hybrid of a Cessna and an eagle) will surely be irresistible to many. The large-scale transition from computers to artificial brains will, as with the transition from vacuum tubes to transistors, probably have to wait until a fresh new generation of technologists – people who grow up on confabulation theory and who possess extensive knowledge of neuroscience – can be produced. At UCSD the cadres who will train this new generation are being prepared today.

The path forward to ubiquitous artificial brains is not likely to be easy, fast, or inexpensive. But it will surely be exciting and fun, with many unexpected twists and turns.

3.4 Symbols: The Universal Language of Cognition

The roughly 4,000 thalamocortical modules which comprise human cerebral cortex and first-order thalamus are each equipped with a set of *symbols* (with each symbol represented by roughly 60 neurons of a special pyramidal symbol-representing neuronal population, probably located in layers II/III of the cortical patch of the module – see Chaps. 5 and 8). Each module typically possesses thousands of symbols, mostly formed in childhood. Symbols are highly stable: once formed, they typically remain unchanged for life (likely with the help of some, as yet unknown, active repair and maintenance processes).

Each module is responsible for describing one *attribute* that an object of the mental world may possess. Its symbols are the *descriptors* of that attribute. For example, a module in a language area might represent the name of an object. Its symbols would then be name words (Mary, John, airplane, etc.). A module in a visual area might represent the color of an object. Its symbols would include: red, green, blue, etc.

Symbols are the stable *terms of reference* that must exist if knowledge is going to be accumulated over decades. Any theory of thought clearly must have such an element. Further, our personal experience teaches us that in mental arenas as varied as language, dance and exercise, driving, visual objects, taste and smell, and texture, the average human possesses thousands to hundreds of thousands of categorical descriptors of each attribute of objects in these realms. Fixed sets of stable symbolic descriptors, one such set for each object attribute, is clearly the simplest explanation for these firm facts.

It might be assumed that if an object is "present in the mental universe" at a particular point in time, that each module describing an attribute pertaining to that object will have a single symbol expressed on it describing the object. This is wrong. Symbols become expressed on modules only as a result of a deliberately commanded, recently completed, confabulation.

If a confabulation has been completed, then at most one symbol will be *active* on that module (if all of the symbol excitation levels are too low a confabulation can yield no symbol – which means "I don't know"). If a confabulation is still in progress, then an *expectation* consisting of multiple *highly excited* (a signaling state lower than the active state) symbols may be expressed on that module – a common occurrence during multiconfabulation.

In summary, usually, only a subset of the attributes which pertain to a particular object are being described by symbols at any particular point in time. The set of involved modules is determined by the confabulations which have recently been commanded. It is likely that no two experiences of a rose will ever be the same.

A key observation is that all symbols are, roughly, equivalent from a neurophysiological viewpoint. They are each a collection of about 60 excited or active symbol-representing neurons within layers II/III of the cortical patch of their module. This suggests that they form the *universal language* of the cognitive brain, i.e., the interactions between symbols are the same no matter what object attribute they represent (visual, auditory, motor process, thought process, plan element, odor, tactile texture, etc.). As will be seen below, this is a huge advantage over past forms of algorithmic and rule-based information-processing because, in the brain, all object attribute representations (i.e., symbols) are freely *interoperable*.

Modules are organized into *hierarchies*, one symbol in a higher-level module often representing multiple sets of particular symbols in collections of lower-level modules.

In humans, the modules used to describe language object attributes form the core *hub* of cognition. These are connected via knowledge links to and from modules belonging to almost every other functional category.

Modules which interface directly with extracortical structures (primary visual cortex, primary motor cortex, primary auditory cortex, etc.) rarely express single symbols. Almost every moment they are in use, these modules are expressing *graded blends* of multiple symbols. This is because such blends are more accurate representations of the extracortical data. Thus, primary motor cortex might be expressing a graded blend of five symbols in order to accurately define a postural goal that, by its expression, initiates a movement (cerebellum – the autopilot of the brain – then smoothly and competently carries out this movement, adapting its commands to reach the cortically commanded postural goal in the specified transit time independent of limb load or drag resistance). [Note: Transit time is specified by the rate at which the graded symbol blend expression is carried out: a faster onset of the postural goal symbol blend yields a faster transit time.] A primary auditory module might be expressing eight symbols at different graded levels of excitation at one time – in order to represent sounds coming from the attended source (see Chap. 7 for more details).

3.5 Knowledge Links

Functionally and anatomically, brains are dominated by synapses (roughly 10^{14} in human cortex, in contrast with roughly 10^{10} neurons). So it is natural that cognition is based upon vast numbers of knowledge links, which are used massively in parallel (with confabulation, a simple, fast, winners-take-all competition among symbols, the only "information-processing operation" required).

The average human adult probably possesses billions of knowledge links. Isaac Newton, Albert Einstein, and Francis Crick might have each had over 100 billion. In our experiments (DVD *video presentation*, Chaps. 4–7) (which have employed up to billions of knowledge links) the average symbol that is expressed on a module as a conclusion has an average of about 200 knowledge links emanating from it in each knowledge base that its module sources. So, during a multiconfabulation it is not unusual for many millions of *relevant* (i.e., emanating from symbols that are viable conclusions) knowledge links to be used. This automatic, massively parallel application of relevant knowledge during thought surely accounts for some of its effectiveness. No previously studied approach to information-processing has ever had this characteristic.

Humans accumulate knowledge links (and symbols) at a prodigious rate during childhood (probably an average of many links per second) and often continue accumulating them throughout life. If true, this will have profound implications for our views of human nature, education, etc. For example, a child returning home after a day at school might report that she "learned nothing" that day. In reality, she probably began the process of establishing over a hundred thousand new knowledge links. Humans (and other animals) are extremely "smart."

As noted in Sect. 3.2, each knowledge link has a weight of the form

$$\log_c(p(\alpha|\lambda)/p_0) + B\,. \tag{3.4}$$

A key strength of confabulation theory is that this mathematical form fits well with the facts of neuroscience.

Knowledge link synapses are all strong. The weakest have weight B, which might be a value of 30. The strongest have weight $\log_c(1/p_0) + B$, which might be a value of 50 (e.g., if the logarithm base $c = 1.5849$). As always, the biological details are more complicated than this because only a fraction of the neurons of a symbol actually receive knowledge link synapses – see Sect. 3.6 below and

Chaps. 5 and 8 – but the introductory discussion in this section can safely ignore these further complications.

The constant B is the *bandgap* (a term coined by Dr. Robert W. Means of my Fair Isaac research team), in analogy with solid-state physics. What B implies is that there are no moderately strengthened synapses – only weak *vestigial* synapses (which are necessary – see below – but unused) and strong knowledge link synapses (which have already been massively strengthened) of minimum strength B.

The dependency of the synapse strength on a logarithmic function of $p(\alpha|\lambda) = p(\alpha,\lambda)/p(\lambda)$ is highly consistent with the neuroscience facts. *Hebbian* learning is a saturable pre-synaptic (axonal terminal process) strengthening varying directly with the joint firing probability $p(\alpha,\lambda)$ of the pre- and post-synaptic neurons (which code α and λ, respectively).

The recently discovered *Marder–Turrigiano variable post-synaptic receptivity* learning (Marder and Prinz 2003, 2002; Turrigiano and Nelson 2004, 2000; Turrigiano et al. 1998) is a saturable efficacy factor, varying directly with $1/p(\lambda)$, implemented in the neurotransmitter transduction zone of the post-synaptic neuron, which multiplies the Hebb pre-synaptic output. Thus, overall synaptic efficacy is directly related to the product $p(\alpha,\lambda) \cdot [1/p(\lambda)] = p(\alpha|\lambda)$.

When combined with their saturable character, the combination of Hebb and Marder–Turrigiano learning yields roughly the above knowledge link weight formula (Eq. 3.4).

The vast majority of excitatory cortico-cortical synapses situated in the proper places to implement knowledge links seem *vestigial* (see also Chaps. 5 and 8). When tested in brain slices using patch clamps (Cowan et al. 2001) these synapses rarely cause any depolarization of the target neuron. This has been interpreted as meaning that the vast majority of synapses are "unreliable." Even when they do function, their depolarization impact on their target cell is small.

This seeming paradox – that a vast majority of cortical excitatory synapses of the type most likely used to implement knowledge links are weak and unreliable – has puzzled neuroscientists for years. A variety of schemes have been devised to explain how such synapses might function. But the evidence suggests that they *do not* function.

Few neuroscientists are willing to accept the notion that brains deliberately create, and assiduously maintain, trillions of hardware elements (vestigial knowledge link final synapses) that are unused. Yet confabulation theory suggests exactly this design.

Vast numbers of vestigial synapses must be present in order to support *instantaneous learning*. Without this seeming "waste," most cognitive learning, and therefore survival, would not be possible.

When a symbol pair are first seen to co-occur "non-casually" (e.g., the target symbol is used shortly thereafter as an assumed fact), a new knowledge link is immediately created by temporary strengthening of the involved synapses. This might occur via the *long-term potentiation* (LTP) mechanism (Cowan et al. 2001), but more likely occurs via another as-yet-unknown mechanism, acting upon the (vestigial) transponder-neuron-to-target-symbol-neuron synapses (see Sect. 3.6). Permanent strengthening of these synapses, if warranted, then occurs over the following few sleep periods (Chaps. 1, 5, 7, 8; Sejnowski and Destexhe 2000). Hippocampus and entorhinal cortex somehow "index" the involved symbol pairs so that, if a particular knowledge link is deemed to be strongly associated with reductions of drive or goal states, it can be consolidated into permanence over the following 100 hours or so.

For any knowledge link to be used, its knowledge base must be deliberately *enabled*. Knowledge base enablement is hypothesized by the theory as a primary function of *higher-order thalamus* (Sherman and Guillery 2006; Casagrande et al. 2005; Paxinos and Mai 2004). Enablement ensures that active symbols do not broadcast excitation through all the knowledge links they source – but only the ones that belong to deliberately enabled knowledge bases involved in a currently ongoing confabulation or multiconfabulation. As with the control of confabulation, knowledge link enablement is thought to be controlled via thought-control inputs emanating from sub-cortical *thought-control nuclei* (roughly analogous to the motor nuclei involved in movement).

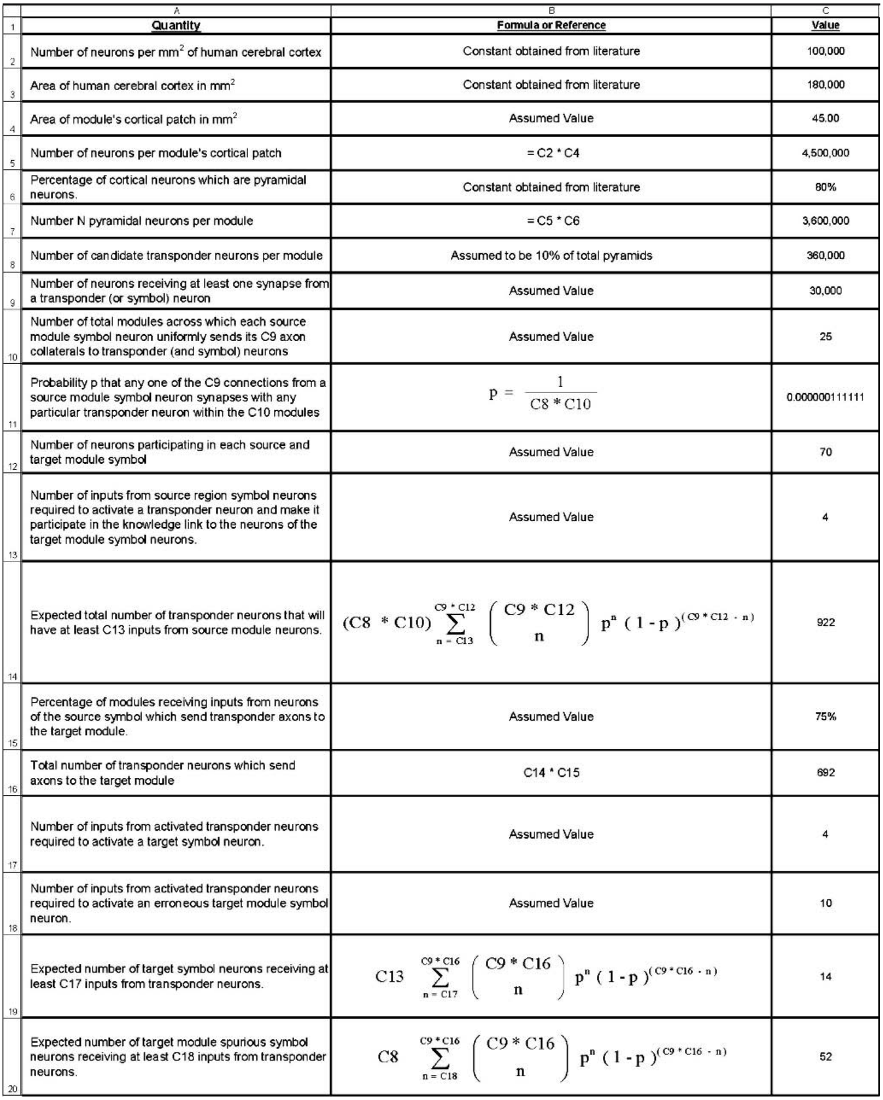
Since knowledge links form within genetically dictated knowledge bases strictly on the basis of meaningful symbol pair co-occurrence, it is common for pairs of symbols in modules with disparate attributes (e.g., visual and language) to be linked. Thought processes often involve expression of disparate assumed facts on multiple modules as inputs to an immediately subsequent confabulation.

For example, selection of a next plan execution step might involve excitation of a planning module's symbols by knowledge links arriving from visual, language, auditory, and olfactory assumed facts. The conclusion symbol selected (the descriptor of the next plan execution step) is that symbol receiving the highest total excitation from these knowledge links. That this works is astounding. Yet it does (as shown in Chap. 6)! This *universal interoperability* of cognitive knowledge (empowered by the *additive knowledge combination* cogency maximization mathematics of confabulation) is one of the most powerful characteristics of animal thought.

3.6 Neuronal Implementation of Knowledge Links

Over the past decade I have been investigating statistical models of how two-stage synfire chain knowledge links might be implemented in a pre-wired cortex. Model A, introduced in 2002 into the curriculum of my UCSD ECE-270 year-long graduate sequence, models one knowledge link. Model B considers 5,000 such links all emanating from the same symbol (this is a "worst case" analysis to show that this hypothesized neuronal implementation can support huge numbers of knowledge links without mutual interference). Model A is presented below.

Model A is implemented as the Microsoft Excel spreadsheet shown in Fig. 3.2. It models the single knowledge link shown in Fig. 3.3. Both the source and target



	A	B	C
1	Quantity	Formula or Reference	Value
2	Number of neurons per mm^2 of human cerebral cortex	Constant obtained from literature	100,000
3	Area of human cerebral cortex in mm^2	Constant obtained from literature	180,000
4	Area of module's cortical patch in mm^2	Assumed Value	45.00
5	Number of neurons per module's cortical patch	= C2 * C4	4,500,000
6	Percentage of cortical neurons which are pyramidal neurons.	Constant obtained from literature	80%
7	Number N pyramidal neurons per module	= C5 * C6	3,600,000
8	Number of candidate transponder neurons per module	Assumed to be 10% of total pyramids	360,000
9	Number of neurons receiving at least one synapse from a transponder (or symbol) neuron	Assumed Value	30,000
10	Number of total modules across which each source module symbol neuron uniformly sends its C9 axon collaterals to transponder (and symbol) neurons	Assumed Value	25
11	Probability p that any one of the C9 connections from a source module symbol neuron synapses with any particular transponder neuron within the C10 modules	$p = \frac{1}{C8 * C10}$	0.000000111111
12	Number of neurons participating in each source and target module symbol	Assumed Value	70
13	Number of inputs from source region symbol neurons required to activate a transponder neuron and make it participate in the knowledge link to the neurons of the target module symbol neurons.	Assumed Value	4
14	Expected total number of transponder neurons that will have at least C13 inputs from source module neurons.	$(C8 * C10) \sum_{n = C13}^{C9 * C12} \binom{C9 * C12}{n} p^n (1 - p)^{(C9 * C12 - n)}$	922
15	Percentage of modules receiving inputs from neurons of the source symbol which send transponder axons to the target module.	Assumed Value	75%
16	Total number of transponder neurons which send axons to the target module	C14 * C15	692
17	Number of inputs from activated transponder neurons required to activate a target symbol neuron.	Assumed Value	4
18	Number of inputs from activated transponder neurons required to activate an erroneous target module symbol neuron.	Assumed Value	10
19	Expected number of target symbol neurons receiving at least C17 inputs from transponder neurons.	$C13 \sum_{n = C17}^{C9 * C16} \binom{C9 * C16}{n} p^n (1 - p)^{(C9 * C16 - n)}$	14
20	Expected number of target module spurious symbol neurons receiving at least C18 inputs from transponder neurons.	$C8 \sum_{n = C18}^{C9 * C16} \binom{C9 * C16}{n} p^n (1 - p)^{(C9 * C16 - n)}$	52

Fig. 3.2. Knowledge link Model A. See text for details

symbol of Model A are assumed to be represented by a collection of 70 neurons (located in layers II/III of the cortical patches of their respective thalamocortical modules – row 12 in Fig. 3.2).

Each source symbol neuron is assumed to send 30,000 axon collaterals, each ending in a synapse, to transponder neurons (row 9) located within a total of 25 modules (row 10). Each transponder neuron is assumed to send 30,000 axon collaterals, each ending in a synapse (row 10), to symbol-representing neurons

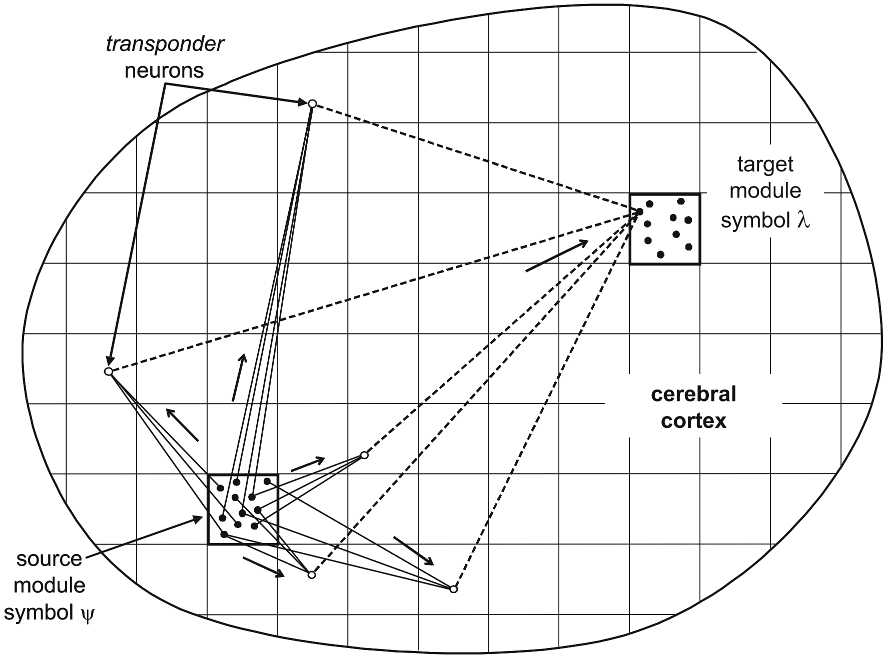


Fig. 3.3. Hypothesized neuronal implementation of a knowledge link. See text for details

located within a total of 25 modules (row 10). Each transponder neuron and symbol-representing neuron within the involved modules is assumed to have a uniformly equal chance of receiving each of the synapses directed at its population (rows 11, 14, 19, and 20).

Model A assumes that a transponder neuron must receive inputs from four source symbol neuron synapses (row 13) in order to become highly excited and participate in the transponder *re-representation* of the symbol.

The modules that the transponder neurons send axon collaterals to are not exactly the same ones that the source symbol neurons send to. However, it is assumed that 75% of them are the same (row 15). This is Model A's way of dealing realistically, but in a mathematically simple way, with the much more complicated (global and local), genetically specified patterns of excitatory axon distribution in human cortical white matter.

Model A examines a single knowledge link by first calculating the probability (row 11) that one of the 360,000 neurons which are candidate transponder neurons (rows 2 through 8) in each of the 25 modules that receive the outputs from the 70 source symbol neurons (row 12) will receive a synapse from a source symbol neuron. This probability is denoted by p (row 11).

Given p, Model A then calculates (using the binomial distribution – row 14) the expected number of transponder neurons which will receive at least the minimum required four inputs (row 13) from source symbol neuron synapses. This calculation is predicated on the assumption that these synapses are uniformly and independently randomly distributed. This calculation shows that there will be about 922 transponder neurons. By this mechanism, the initial set of 70 *active* source-symbol-representing neurons is "amplified" to 922 *highly excited transponder* neurons.

The synapses between the source symbol neurons and the transponder neurons are assumed to be moderately strong and reliable. These synapses are assumed to be moderately permanently strengthened, by a separate "non-Hebbian" process, on the basis that the transponder neurons must become a reliable momentary surrogate for the source symbol neurons for decades to come. This process must be "non-Hebbian" because the transponder neurons are not already excited before the inputs from the source neurons arrive. This is a neuroscience prediction of confabulation theory. Because of this modest strengthening, it is assumed that only four synapses from source neurons are required to excite a successful transponder neuron.

The excited state is less potent than the active state, and the outputs delivered by the transponder neurons are far less synchronized than those of the source symbol neurons – and for these reasons (and others not discussed here) the transponder neurons cannot go on to "ignite" an even larger set of "secondary" transponders. However – and this is the key to knowledge link implementation – the transponders are sufficiently numerous to make sufficiently strong connectivity to a sufficiently large subset of the target symbol neurons a virtual certainty. This is the next issue addressed by Model A.

The synapses which exist from transponder neurons to target symbol neurons are all highly strengthened. It is assumed that only four such synapses are required to highly excite a target symbol neuron (row 17), even if that neuron is initially inactive. The expected number of target symbol neurons which receive this number of transponder neuron synapses is again based upon uniform independent random distribution of these connections (row 19). Thus, Model A shows that 14 out of the 70 target symbol neurons will become highly excited by inputs from transponder neurons re-representing the source symbol.

The final issue addressed by Model A is to quantify the expected number of symbol-representing neurons of the target module that will be erroneously excited by this knowledge link (i.e., symbol-representing neurons that do not participate in representing the target symbol). Erroneous excitation is assumed to happen whenever a target module symbol neuron receives at least 10 unstrengthened transponder neuron synaptic inputs (row 18).

Erroneous symbol neurons are potentially a problem because, even though the probability of each symbol neuron being highly excited is low, there are a total of 360,000 symbol neurons in the target module. Row 20 calculates the expected number of erroneously highly excited target module symbol neurons using the binomial distribution. The result is 52 expected erroneous neurons. Since these are spread randomly across the symbols of the target module, no symbol's excitation level will likely be significantly affected. Thus, spurious activations are not a problem.

In summary, Model A shows that human cortical implementation of billions of knowledge links by two-stage synfire chains (Abeles 1991) may be possible. This crude model now needs to be followed up by more detailed models that take detailed neuroanatomy and neuronal behavior into account.

3.7 Neuronal Implementation of Confabulation

This section describes a current simple mathematical model (Model 5) of how a module implements confabulation. Model 5 is simply awful, as are all existing models. This is an area of research that needs a lot of work. However, now that we know that cortical information-processing is based entirely upon attractor networks, it should be possible to garner the talent and resources to attack this problem more vigorously.

The first mathematical model of an attractor network that explicitly converges to one of L fixed collections of active neurons (symbols) is that developed by Karen Haines and myself in 1988 (Haines and Hecht-Nielsen 1988). This network is constructed out of two Willshaw non-holographic associative memories (Willshaw et al. 1969) connected reciprocally.

Computer experiments with this network were first carried out during the mid-1990s by students in my UCSD ECE-270 graduate course **Neurocomputing** (a course which Haines helped develop during the period 1987–1992). These experiments yielded two conclusions: First, the network would converge to the symbol that was closest in Hamming distance to the starting state – as long as that Hamming distance was relatively small compared to the number of neurons in each symbol (all the symbols are represented by about the same number of, randomly selected, neurons). Second, under the above condition, the network would almost always converge in one cycle (Haines and I had proved convergence in 1988, under certain conditions, but did not know how quickly it would happen). Another key attribute of these networks (which Haines and I had also shown mathematically) was that they have no "spurious attractors" – at least under the assumptions we employed. The only attractors are the symbols.

During the 1990s, it became ever clearer to me that it would frequently be necessary for an initial mixture of tens, hundreds, or even thousands, of partially excited symbols to rapidly converge to the single most excited symbol. Our attractor network could clearly not meet this requirement.

By the early 2000s, Sommer and Palm (Sommer and Palm 1999), and Seung and his colleagues (Xie et al. 2001; Hahnloser et al. 2003) had also become interested in attractor networks with neuron collection attractors. For example, the Seung networks featured mutual excitation among the neurons of each symbol and mutual inhibition between neurons of different symbols. Both the Sommer and the Seung networks are superior to the Haines and Hecht-Nielsen network, in that they allow convergence with multiple competing partially excited symbols (for example, a target symbol receiving four of the knowledge links of Model A of Sect. 3.6 would have roughly 56 – row 19 times four – of its 70 neurons receiving high excitation). However, in all existing networks, the number of symbols which can successfully compete is still much too low, and convergence is much too slow (seconds or more, a time scale based upon a single neuron updating its state once per millisecond). Also, the amount of inhibitory interaction required in the Seung network seems neuroanatomically unrealistic.

In UCSD ECE-270 (as it is currently formulated) we study five attractor network models (Models 1–5). Model 5 is an improved version of the early Seung models. In Model 5, each symbol-representing neuron excites all of the other neurons that represent that symbol. [A complication, which the Haines and Hecht-Nielsen, Sommer, and Seung models already admirably accommodate, is that each symbol-representing neuron actually participates in representing many symbols.]

In Model 5, symbol-representing neurons (which are arranged in a regular 2-D array) are assumed to inhibit a fraction of the nearby symbol-representing neurons which they do not excite. This inhibition is assumed to be implemented by local inhibitory interneurons possessing many gap junctions between them (Fukuda et al. 2006). This has the effect of essentially producing a level of local inhibition based upon the maximum excitation level of nearby symbol-representing neurons.

Model 5 is successfully able to deal with starting mixtures of hundreds of partially excited symbols, and its inhibitory elements fit within the confines of known cortical neuroanatomy and neurophysiology (e.g., the gap junctions used in the model are known to exist). However, Model 5's convergence speed is still far too slow.

Development of Models 6 and 7 (having formal neurons and spiking neurons, respectively) is now underway. These models incorporate the full known thalamocortical loop from cortical layers II/III to layer V to layer VI to first-order NRT/thalamus, back to NRT/cortical layer IV, and thence back to layers II/III (Sherman and Guillery 2006; Casagrande et al. 2005). The goal of Model 6 is to understand the basic mechanism used to dramatically speed up the confabulation competition process in thalamocortical modules. The goal of Model 7 is to implement Model 6 using spiking neurons [e.g., via the Izhikevich phenomenological spiking neuron model (Izhikevich 2006, 2007)].

The ultimate goal of this research is to produce accurate predictions of in vivo confabulating thalamocortical module neuronal behavior which can be extracted and tested. This research is underway in the UCSD Confabulation Neuroscience Laboratory under sponsorship of ONR.

3.8 The Origins of Confabulation Theory

Any new scientific development has a key historical question attached to it: How did the discovery come about? In other words, upon whose shoulders did the final discoverer choose to stand? Since almost every scientific discovery (and certainly this one) is the product of centuries of largely unheralded toil by tens of thousands of scientists, it may seem unfair to single out a few. But the actual process of *doing* science almost always involves selection of a fixed set of clues developed by specific previous research and then years of trying to discover how these selected clues fit together in nature [Albert Einstein termed this capability to spot key clues “good taste” and attributed his scientific success largely to possession of it (Einstein 1961).] The owners of these “penultimate shoulders” deserve rich praise and thanks from the world. But only the final discoverer definitively knows who they are. Unequivocal identification of these key contributors to the confabulation theory discovery is the purpose of this section.

Of the many thousands of items of past research of which I am aware, the overwhelmingly most important and influential relative to the confabulation theory discovery were those of Anderson (Anderson et al. 1977; Hecht-Nielsen 1989), Willshaw (Willshaw 1969; Hecht-Nielsen 1989), Abeles (Abeles 1991), and Singer (Gray et al. 1989; Freiwald et al. 2001; Fries et al. 2002).

If we were to separate the knowledge links and symbol sets, which are mashed together in Anderson’s 1977 “Brain State in a Box” construct, we would essentially have a confabulation architecture. The BSB has, for almost 30 years, helped to strongly guide my research. The BSB’s fixed symbol sets for describing object attributes, Hebbian co-occurrence-based knowledge links between pairs of symbols, and attractor network processing have, for all this time, been, for me, essential elements of any acceptable theory of thought.

Willshaw’s 1969 “non-holographic associative memory” construct has, for over 35 years, provided me with a mathematical and conceptual understanding that the fundamental components of cognition must, in some way, be neuron collections. When, in 1988, Karen Haines and I developed, and theoretically investigated the capabilities of, attractor networks constructed from a pair of reciprocally connected non-holographic associative memory structures (Haines and Hecht-Nielsen 1988), it quickly became clear to me that this sort of attractor network, robustly converging to fixed, sparse collections of neurons, each collection representing one symbol, was another essential element of an eventual theory of thought.

Abeles’ 1991 “synfire chain” construct provided key insight into how billions of knowledge links, each connecting one symbol’s neuron collection to that of another, can be instantly formed, on demand, in a pre-wired brain.

Singer’s 1989 discovery of pairs of precisely synchronized cortical “feature detector” neurons *in vivo* at distances of hundreds of microns from one another [a discovery related to earlier synchronization hypotheses of Hebb (Hebb 1949), von der Malsburg (von der Malsburg 1981), Crick (Crick 1984), and Kryukov – now Monk Fiofan, (Kryukov et al. 1990), of which I was aware] provided me strong encouragement that symbols are briefly simultaneously active neuron collections in which detailed spike timing plays some important role.

The final realization – that a simple winners-take-all attractor network function in which symbols belonging to a single module compete on the basis of their additive excitation by incoming antecedent-support-probability-weighted knowledge links (i.e., *confabulation*) was all that was needed to explain thinking – took 15 more years.

An indispensable contributor to the successful completion of this final research phase was domestic cat Zeus Hecht-Nielsen (for photos see Chaps. 2 and 8), who functioned over this period as co-investigator, behavioral observation subject, and muse. On the majority of mornings from his birth in May 1990 to the present, Zeus and I have spent time together exploring the grounds of the family compound and then sharing breakfast. These thousands of hours of interaction and observation provided essential insights that led to the four main elements of confabulation theory. Without Zeus, there would be no confabulation theory.

My deepest thanks to James Anderson, David Willshaw, Moshe Abeles, Wolf Singer, and Zeus Hecht-Nielsen.

3.9 Discussion

The mathematics of thought, as sketched in this chapter, is probably close to its final form. The neuroscience of thought (e.g., as described here and in Chaps. 5 and 8) is, at the gross level (modules, symbols, confabulation, and behavioral triggering), probably also close to finalized. However, as illustrated by the primitive neuronal implementation models of knowledge links and confabulation sketched in Sects. 3.6 and 3.7, the neuroscience at more detailed levels has a long way to go. Hopefully, this chapter will help recruit more researchers to consider and investigate confabulation theory.

Acknowledgments

This work was supported by Fair Isaac Corporation and the Office of Naval Research.

4 Cogent Confabulation[4]

A new model of vertebrate cognition is introduced: maximization of *cogency* $p(\alpha\beta\gamma\delta|\varepsilon)$. This model is shown to be a direct generalization of Aristotelian logic, and to be rigorously related to a calculable quantity. A key aspect of this model is that in Aristotelian logic information environments it functions logically. However, in non-Aristotelian environments, instead of finding the conclusion with the highest probability of being true (a popular past model of cognition), this model instead functions in the manner of the "duck test," by finding that conclusion which is most supportive of the truth of the assumed facts.

4.1 Introduction

An appealing model of cognition (Bender 1996; Nilsson 1998; Pearl 2000 – see Chap. 3) is to generalize Aristotelian implication $\alpha\beta\gamma\delta \Rightarrow \varepsilon$ by finding that symbol ε which maximizes *a posteriori probability* $p(\varepsilon|\alpha\beta\gamma\delta)$ (for concreteness, four assumed fact symbols α, β, γ, and δ, and a conclusion symbol ε, each drawn from its own separate module, with juxtaposition indicating Boolean AND, will be used in the discussion of this chapter; the generalization to arbitrary situations is obvious). However, as discussed in Sect. 4.3, this model of cognition is not correct. This chapter introduces a new model of vertebrate cognition: maximization of *cogency* $p(\alpha\beta\gamma\delta|\varepsilon)$ – and considers some related mathematical quantities.

Some terminology: Assuming that the combined assumed facts $\alpha\beta\gamma\delta$ are true, the set of all symbols λ (in the *answer module* from which conclusions are being sought) with $p(\alpha\beta\gamma\delta|\lambda) > 0$ is called the *expectation*, the elements of which, in descending order of their cogencies, are termed *candidate conclusions* or *answers*.

[4] This chapter is a reformatted reprint of the original publication Hecht-Nielsen R (2005) Cogent confabulation. Neural Networks 18:111–115, Copyright (2005), used with permission from Elsevier.

4.2 Cogency and Confabulation

Assume that $\alpha\beta\gamma\delta \Rightarrow \varepsilon$ exclusively in the answer module. Then $p(\alpha\beta\gamma\delta\varepsilon) > 0$ and $p(\alpha\beta\gamma\delta\lambda) = 0$ for all such other answer module symbols λ. Thus, $p(\alpha\beta\gamma\delta|\varepsilon) = p(\alpha\beta\gamma\delta\varepsilon)/p(\varepsilon) > 0$ and $p(\alpha\beta\gamma\delta|\lambda) = p(\alpha\beta\gamma\delta\lambda)/p(\lambda) = 0$ for all other symbols λ.

This establishes:

Theorem 4.1: *If* $\alpha\beta\gamma\delta \Rightarrow \varepsilon$ *exclusively, then maximization of cogency produces one and only one answer*: ε. □

Thus, surprisingly, in an Aristotelian logic information environment, maximizing cogency will produce logical answers. But what about more general environments? Conceptually, cogency maximization works like the *duck test*: if a duck-sized creature quacks like a duck, walks like a duck, swims like a duck, and flies like a duck (assumed facts $\alpha\beta\gamma\delta$), then we accept it as a duck (because duck, ε, is the symbol that, when it is seen, most strongly supports the probability of these assumed facts being true; i.e., ε maximizes $p(\alpha\beta\gamma\delta|\varepsilon)$). There is no logical guarantee that this creature is a duck; but maximization of cogency makes the decision that it is and moves on.

Of course, cogency $p(\alpha\beta\gamma\delta|\varepsilon)$ is a conceptual, notional quantity and can only be calculated in trivial situations. Consider the possibility of using *confabulation* (maximization of the product $p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)$ or, equivalently, the sum of the logarithms of these probabilities) as a surrogate for maximizing cogency. [It is assumed that all required pairwise conditional probabilities $p(\psi|\lambda)$ between symbols ψ and λ are known. This assumption is termed *exhaustive knowledge*.] Each meaningful non-zero $p(\psi|\lambda)$ is termed an individual *item of knowledge*.

An exact mathematical relationship between the confabulation product and cogency is now derived. Applying the probabilistic chain rule identity $p(abcde) = p(a|bcde) \cdot p(b|cde) \cdot p(c|de) \cdot p(d|e) \cdot p(e)$ to cogency, and using the fact that the AND operation commutes, we can write the quantity $p(\alpha\beta\gamma\delta|\varepsilon)$ in all four of the following ways:

$$\begin{aligned}
p(\alpha\beta\gamma\delta|\varepsilon) &= p(\alpha\beta\gamma\delta\varepsilon)/p(\varepsilon)\\
&= p(\alpha|\beta\gamma\delta\varepsilon) \cdot p(\beta|\gamma\delta\varepsilon) \cdot p(\gamma|\delta\varepsilon) \cdot p(\delta|\varepsilon)\\
p(\alpha\beta\gamma\delta|\varepsilon) &= p(\beta\gamma\delta\alpha\varepsilon)/p(\varepsilon)\\
&= p(\beta|\gamma\delta\alpha\varepsilon) \cdot p(\gamma|\delta\alpha\varepsilon) \cdot p(\delta|\alpha\varepsilon) \cdot p(\alpha|\varepsilon)\\
p(\alpha\beta\gamma\delta|\varepsilon) &= p(\gamma\delta\alpha\beta\varepsilon)/p(\varepsilon)\\
&= p(\gamma|\delta\alpha\beta\varepsilon) \cdot p(\delta|\alpha\beta\varepsilon) \cdot p(\alpha|\beta\varepsilon) \cdot p(\beta|\varepsilon)\\
p(\alpha\beta\gamma\delta|\varepsilon) &= p(\delta\alpha\beta\gamma\varepsilon)/p(\varepsilon)\\
&= p(\delta|\alpha\beta\gamma\varepsilon) \cdot p(\alpha|\beta\gamma\varepsilon) \cdot p(\beta|\gamma\varepsilon) \cdot p(\gamma|\varepsilon). \qquad (4.1)
\end{aligned}$$

Multiplying these equations together gives:

$$\begin{aligned}
[p(\alpha\beta\gamma\delta|\varepsilon)]^4 = {} & [p(\alpha|\beta\gamma\delta\varepsilon) \cdot p(\beta|\gamma\delta\varepsilon) \cdot p(\gamma|\delta\varepsilon)] \cdot [p(\beta|\gamma\delta\alpha\varepsilon) \cdot p(\gamma|\delta\alpha\varepsilon)\\
& \cdot p(\delta|\alpha\varepsilon)] \cdot [p(\gamma|\delta\alpha\beta\varepsilon) \cdot p(\delta|\alpha\beta\varepsilon) \cdot p(\alpha|\beta\varepsilon)] \cdot [p(\delta|\alpha\beta\gamma\varepsilon)\\
& \cdot p(\alpha|\beta\gamma\varepsilon) \cdot p(\beta|\gamma\varepsilon)] \cdot [p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)]. \qquad (4.2)
\end{aligned}$$

Applying Bayes' law to the conditional probabilities in the first four parentheses yields:

$$\begin{aligned}[\mathrm{p}(\alpha\beta\gamma\delta|\varepsilon)]^4 = {} & [\mathrm{p}(\alpha\beta\gamma\delta\varepsilon)/\mathrm{p}(\beta\gamma\delta\varepsilon)\cdot\mathrm{p}(\beta\gamma\delta\varepsilon)/\mathrm{p}(\gamma\delta\varepsilon)\cdot\mathrm{p}(\gamma\delta\varepsilon)/\mathrm{p}(\delta\varepsilon)]\\ & [\mathrm{p}(\beta\gamma\delta\alpha\varepsilon)/\mathrm{p}(\gamma\delta\alpha\varepsilon)\cdot\mathrm{p}(\gamma\delta\alpha\varepsilon)/\mathrm{p}(\delta\alpha\varepsilon)\cdot\mathrm{p}(\delta\alpha\varepsilon)/\mathrm{p}(\alpha\varepsilon)]\\ & [\mathrm{p}(\gamma\delta\alpha\beta\varepsilon)/\mathrm{p}(\delta\alpha\beta\varepsilon)\cdot\mathrm{p}(\delta\alpha\beta\varepsilon)/\mathrm{p}(\alpha\beta\varepsilon)\cdot\mathrm{p}(\alpha\beta\varepsilon)/\mathrm{p}(\beta\varepsilon)]\\ & [\mathrm{p}(\delta\alpha\beta\gamma\varepsilon)/\mathrm{p}(\alpha\beta\gamma\varepsilon)\cdot\mathrm{p}(\alpha\beta\gamma\varepsilon)/\mathrm{p}(\beta\gamma\varepsilon)\cdot\mathrm{p}(\beta\gamma\varepsilon)/\mathrm{p}(\gamma\varepsilon)]\cdot\\ & [\mathrm{p}(\alpha|\varepsilon)\cdot\mathrm{p}(\beta|\varepsilon)\cdot\mathrm{p}(\gamma|\varepsilon)\cdot\mathrm{p}(\delta|\varepsilon)].\end{aligned}\tag{4.3}$$

If any of the probabilities within the first four bracketed quantities on the right side of this equation are zero, but the fifth bracketed quantity is not zero, then this is said to be an *exceptional* case. Noting that the first probability in each of the first four parentheses equals p($\alpha\beta\gamma\delta\varepsilon$) and rearranging and simplifying yields:

Theorem 4.2: *Given non-exceptional assumed facts* α, β, γ, *and* δ, *and expectation element* ε, *then the following exact relationship holds between cogency* $p(\alpha\beta\gamma\delta|\varepsilon)$ *and the confabulation product* $\mathrm{p}(\alpha|\varepsilon)\cdot\mathrm{p}(\beta|\varepsilon)\cdot\mathrm{p}(\gamma|\varepsilon)\cdot\mathrm{p}(\delta|\varepsilon)$:

$$\begin{aligned}[p(\alpha\beta\gamma\delta|\varepsilon)]^4 = {} & [p(\alpha\beta\gamma\delta\varepsilon)/p(\alpha\varepsilon)]\cdot\\ & [p(\alpha\beta\gamma\delta\varepsilon)/p(\beta\varepsilon)]\cdot\\ & [p(\alpha\beta\gamma\delta\varepsilon)/p(\gamma\varepsilon)]\cdot\\ & [p(\alpha\beta\gamma\delta\varepsilon)/p(\delta\varepsilon)]\cdot\\ & [p(\alpha|\varepsilon)\cdot p(\beta|\varepsilon)\cdot p(\gamma|\varepsilon)\cdot p(\delta|\varepsilon)].\end{aligned}\qquad\square$$

To see a key implication of Theorem 4.2, consider the following concrete case: five distinct, but identical, modules. Each module possesses 10,000 symbols, each representing exactly one of the 10,000 most common words in a huge reference corpus of uncapitalized proper English text (novels, encyclopedias, news stories, etc.). Let assumed fact symbols α, β, γ, and δ be drawn from modules 1, 2, 3, and 4, respectively; representing a contiguous sequence of four words of text. Then let us consider which symbols ε, if any, from module 5 would make a suitable completion to this phrase $\alpha\beta\gamma\delta$. These ε's will be the symbols with the largest cogencies p($\alpha\beta\gamma\delta|\varepsilon$). To make the example even more concrete, let the assumed fact phrase be: $\alpha\beta\gamma\delta$ = `the train was going` and consider one possible expectation symbol ε, representing the word `south`. Then,

$$\begin{aligned}\mathrm{p}(\alpha\beta\gamma\delta\varepsilon)/\mathrm{p}(\alpha\varepsilon) &= \mathrm{p}(\texttt{the train was going south})/\\ &\quad\ \mathrm{p}(\texttt{the \_\_\_ \_\_\_ \_\_\_ south})\\ \mathrm{p}(\alpha\beta\gamma\delta\varepsilon)/\mathrm{p}(\beta\varepsilon) &= \mathrm{p}(\texttt{the train was going south})/\\ &\quad\ \mathrm{p}(\texttt{\_\_\_ train \_\_\_ \_\_\_ south})\\ \mathrm{p}(\alpha\beta\gamma\delta\varepsilon)/\mathrm{p}(\gamma\varepsilon) &= \mathrm{p}(\texttt{the train was going south})/\\ &\quad\ \mathrm{p}(\texttt{\_\_\_ \_\_\_ was \_\_\_ south})\end{aligned}$$

and

$$\begin{aligned}\mathrm{p}(\alpha\beta\gamma\delta\varepsilon)/\mathrm{p}(\delta\varepsilon) &= \mathrm{p}(\texttt{the train was going south})/\\ &\quad\ \mathrm{p}(\texttt{\_\_\_ \_\_\_ \_\_\_ going south})\end{aligned}$$

where an underline indicates a word position that is not being considered in calculating the probability.

Note that if ε (`south`) were replaced by any other expectation element (`north`, `east`, `west`, `fast`, `slow`, etc.) these ratios would probably change very little. Thus, in non-exceptional cases, these first four terms might function approximately as a positive constant independent of ε, making the product $p(\alpha|\varepsilon) \cdot p(\beta|\varepsilon) \cdot p(\gamma|\varepsilon) \cdot p(\delta|\varepsilon)$ and the fourth power of cogency approximately proportional. Under these circumstances, confabulation and maximizing cogency will give the same answers. Theorem 4.2 is postulated to be the "fundamental theorem" of vertebrate cognition.

While the above argument may be correct in the specific case of English phrase completion considered, how can we be sure that the first four terms of the fundamental theorem will, in general, be approximately constant for all expectation elements ε? In general, we cannot. Besides the problem of exceptions (which, it turns out, can be handled by explicitly learning all of them), badly designed modules or ill-mannered information environments can almost certainly cause these first four terms to not be approximately constant for all expectation elements.

Biological systems find ways of exploiting a variety of scientific, technological, and mathematical principles, but to do so, they must often improvise (by evolution) specific designs that conform to the requirements and limitations of the principle. Cellular biochemistry developed in the ocean and so we must carry the ocean around with us in order to use these innovations. Cognition is presumably like this. With the proper modules and knowledge, developed in the proper sequence via exposure to the proper information environments (which are all things that, ultimately, genetics, and therefore evolution, can control), the fundamental theorem of cognition can be exploited and confabulation can be *cogent*. Without these restrictions, confabulation probably does not work.

4.3 Confabulation Examples

Here are some examples of confabulation applied to phrase completion. As in the discussion of Theorem 4.2, a sequence of four words, each from its own module, is given as assumed facts α, β, γ, and δ (in some examples only one, δ, two, $\gamma\delta$, or three, $\beta\gamma\delta$, assumed facts are used). Confabulation is used to select the symbol ε of the next word after the assumed fact phrase. Each module has 10,000 symbols, representing the 10,000 most commonly encountered words in a 1.4×10^9-word proper English training corpus composed of books, news stories, encyclopedias, etc. For simplicity, capital letters are not used. The pairwise conditional probabilities were obtained by marching a five-contiguous-word window one word at a time from the beginning word of the training corpus to the end. Each time a contiguous sub-string of two to five words without internal punctuation appeared at the right-hand end of the window, counts were gathered and stored for that substring in the corresponding selections of four

10,000 × 10,000 matrices (one for each of the first four symbol modules paired with the fifth).

After the march through the training corpus, the probability p(ψ|λ) between symbols ψ and λ was approximated by c(ψ,λ)/c(λ), where c(ψ,λ) is the count of the number of times the word represented by symbol ψ (belonging to one of the first four modules) and the word represented by symbol λ (belonging to the fifth module) appeared together, and c(λ) is the total number of times symbol λ appeared in module five with any symbol of the ψ module during the march (c(λ) is equal to the sum of the c(ϕ,λ) across all symbols ϕ belonging to the ψ module). Only the pairwise conditional probabilities between symbols in the first four modules, conditioned on symbols of the fifth module, were computed because the phrase completion confabulations were always carried out with the fifth module as the answer module. Symbol pair counts below 3 were thrown out (set to zero) as accidental or meaningless, as were calculated p(ψ|λ) values below 0.001. This knowledge acquisition process yielded 5,251,335 meaningful items of knowledge. This knowledge acquisition process was carried out on a desktop computer in a few hours.

To test the system, a highly literate but non-technical person was asked to create a number of test word sequences to be completed. Confabulation was then applied to these assumed facts with the results shown below. The symbols determined by confabulation to be expectation elements are shown in decreasing order of the confabulation product value. When an expectation had seven or more answers, the words corresponding to the top six are shown in parentheses, followed by the total number of symbols in the expectation; if six or fewer answers were found, square brackets are used, and all of the corresponding words are shown:

- she could determine (whether, exactly, if, why, how, precisely) 8
- if it was not (immediately, clear, enough, true, properly, stupid) >999
- earthquake activity was [centered]
- for lack of a (unified, blockbuster, comprehensive, definitive, coordinated, protein) 111
- a lack of (urgency, oxygen, understanding, confidence, communication, enthusiasm) 407
- regardless of expected [outcome, length]
- cars drove down a (lane, freeway, highway, dirt, taxi, tying) 9
- driving west on interstate [highway, freeway]
- snow fell in (freezing, montana, portions, northwestern, northeastern) 11
- the facts point to [ ]
- threats of terrorist [attacks, retaliation, strikes, violence]

- the machine (tools, tool, guns, gun, operator, shop) 33
- children can learn [lessons, math, english]
- students can learn [lessons, math, english]
- college students can learn [math]
- knowledge of historical [facts, subjects, styles]
- questions that cannot be (answered, solved, resolved, avoided, addressed, yes) 9
- benefits from additional (cost-cutting, taxable, protections, taxes, acquisitions, payroll) 11
- limitations [expired, expires, imposed]
- her responsibility for taking [sole, matters]
- her responsibility for making [errors, matters, sure, references, choices, lethal]
- his responsibility for making (mistakes, matters, bombs, references, decisions, sure) 11
- his responsibility for taking [actions, sole, matters, decisions]
- mechanical failure [caused]
- crowded (commuter, marketplace, subway, courtroom, skies, sidewalk) 62
- they crowded (onto, lobby, shopping, shelters, around, into) 22
- beaches are covered with [pools]
- there were many (indications, surprises, instances, casualties, signs, exceptions) >999
- are easy to (install, dismiss, detect, locate, accumulate, criticize) 159
- microsoft makes software for [apple's, desktop, hardware]
- the green car turned [yellow] .

Notice the automatic and instantaneous emergence of “grammar” and “semantics.”

4.4 Discussion

In the experiments above, many of the completed phrases never appeared anywhere in the training corpus. This strong ability to correctly generalize to cases which are novel in detail, but which involve familiar elements and include no unlearned exceptions, is a favorable characteristic that confabulation seems to possess. In addition, tens of nonsense phrases (e.g., $\alpha\beta\gamma\delta =$ tune card fly bold) were tested as assumed facts for phrase completion and, in every case, confabulation returned no answers.

Some of the phrase completion examples presented above are perfectly valid sets of assumed facts (e.g., `the facts point to`), which should have at least some reasonable completions. Yet confabulation returns no answers. This is an indication that the knowledge used is not completely exhaustive. Thus, a collection of knowledge that falls somewhat short of being exhaustive seems to translate into a tendency to make errors of omission, not commission; a generally favorable attribute.

Synaptic learning of the $p(\psi|\lambda)$ probabilities (Chaps. 3, 5, and 7; Hecht-Nielsen 2004) and the high-speed parallel competitive "winners-take-all" implementation of confabulation by neuronal attractor networks (Anderson et al. 1977; Sommer and Palm 1999; Amit 1989) seem biologically plausible.

Now, as quickly as you can, select a next word for each of the following phrases:

- `company rules forbid taking`
- `mickey and minnie were`
- `capitol hill observers are`
- `paper is made from`
- `riding the carousel was`

The idea of finding that conclusion ε which has the highest probability of being true, $p(\varepsilon|\alpha\beta\gamma\delta)$, given the assumed facts, $\alpha\beta\gamma\delta$, has, for decades, been an attractive model of cognition. This attractiveness is seemingly bolstered by the (average error rate) optimality of *a posteriori* probability in pattern classification. A great deal of study has gone into this model (Bender 1996; Nilsson 1998; Pearl 2000). However, this is actually an awful model of cognition. Consider the following numerical calculations.

Given a set of assumed facts $\alpha\beta\gamma\delta$, let λ be a conclusion with *a priori* probability $p(\lambda) = 0.01$ and ε an alternative conclusion with $p(\varepsilon) = 0.0001$. Also assume that $p(\alpha\beta\gamma\delta|\lambda) = 0.01$ and $p(\alpha\beta\gamma\delta|\varepsilon) = 0.2$. Applying Bayes' law twice to *a posteriori* probability yields:

$$p(\psi|\alpha\beta\gamma\delta) = p(\alpha\beta\gamma\delta|\psi) \cdot [p(\psi) \,/\, p(\alpha\beta\gamma\delta)]. \tag{4.4}$$

Thus, $p(\lambda|\alpha\beta\gamma\delta) = 5 \cdot p(\varepsilon|\alpha\beta\gamma\delta)$ and so the policy of maximizing *a posteriori* probability will overwhelmingly choose λ over ε even though $p(\alpha\beta\gamma\delta|\varepsilon) = 20\ p(\alpha\beta\gamma\delta|\lambda)$. Thus, it should be possible to discern whether maximum *a posteriori* probability is a good model of cognition by seeing how strongly *a priori* probability enters into cognitive decision making. In performing the above phrase completions, you have produced relevant (although perhaps not statistically significant!) experimental data that bears on this question.

In an informal poll, typical answers for the completions were: `naps`, `happy`, `wondering`, `wood`, and `fun`. However, in each example, the word `the` is both a viable answer and, by far, the most frequent word in English; so, if maximization of *a posteriori* probability were a correct theory of cognition, you would surely have selected it, overwhelmingly (as the above calculation illustrates), in every case. Instead, as with the computer confabulation experiments presented in the previous section, you probably selected words with much higher cogencies than `the`.

Cogency, $p(\alpha\beta\gamma\delta|\varepsilon)$, the cognitive analog of class probability in pattern classification theory, and likelihood in probability and statistics, has always been there. Waiting. Perhaps its time has finally arrived.

Acknowledgments

Thanks to Adrian T. Fan, Kate Mark, Robert W. Means, and Syrus C. Nemat-Nasser for implementing the confabulation experiments and to Fair Isaac Corporation for long-term support of this research.

5 Confabulation Neuroscience I[5]

A fast winners-take-all competition process, termed *confabulation* (Chaps. 1–4; Chaps. 5–8), is proposed as the fundamental mechanism of all aspects of cognition (vision, hearing, planning, language, control of thought and movement, etc.). Multiple, contemporaneous, mutually interacting confabulations – in which millions of items of relevant knowledge are applied in parallel – are typically employed in thinking. At the beginning of such a *multiconfabulation*, billions of distinct, potentially viable, conclusion sets are considered. At the end, only one remains. This fast, massively parallel application of relevant knowledge (an alien kind of information-processing with no analog in today's computational intelligence, computational neurobiology, or computer science) is hypothesized to be the core explanation for the information-processing effectiveness of thought. This paper presents a synopsis of this *confabulation theory* of human cortical and thalamic function.

5.1 Introduction

Confabulation theory offers a comprehensive, concrete explanation for animal cognition. The theory hypothesizes the specific underlying mathematical mechanism of cognition, as well as the human neuronal implementation of that mechanism (specified at a "meta-level" of neurophysiological detail: summary descriptions of the dynamical behavior of hypothesized sub-groups of neurons).

Confabulation theory proposes that all aspects of cognition (seeing, hearing, command of movement and thought, planning, language, abstract thinking, etc.) are implemented using four fundamental elements: (1) a universal modular system for representing the objects of the mental world, (2) knowledge links, (3) confabulation, and (4) action command origination. These four elements are briefly sketched in the following four sections, emphasizing their human implementations. Section 5.6 summarizes the underlying mathematics of confabulation, and Sect. 5.7 sketches multiconfabulation. The concrete numerical parameter values stated here (cortical patch area, number of neurons representing one symbol, etc.) are guesses presented solely to help fix ideas.

[5] This chapter is based on the original publication Hecht-Nielsen R (2006) The mechanism of thought. In: Proceedings of the World Congress on Computational Intelligence. Vancouver, BC, Canada. July 16–21, IEEE Press, Piscataway, NJ, and is adapted here in accordance with IEEE copyrights.

5.2 Confabulation Theory Element 1

5.2.1 Cognitive World Object Representation

As illustrated in Fig. 5.1, human cerebral cortex is hypothesized to be exhaustively divided into roughly 4,000 discrete, localized, largely disjoint patches, each the cortical component of a *module*. The exact physical form, and functional details, of modules are not specified by the theory, and are not known. For example, applying Sutton and Strangman's "network of networks" hypothesis (Sutton and Strangman 2003), each module could be made up of collections of smaller "sub-modules." Another complicating issue is that excitatory cortical neurons having "pyramidal" morphology (which make up the majority of cortical neurons) almost surely are further divided into many distinct sub-categories that have distinctive connectivity (Markram 2003).

Each module is used to represent one *attribute* that an *object* (visual, auditory, conceptual, abstract, motor process, thought process, plan, etc.) of the

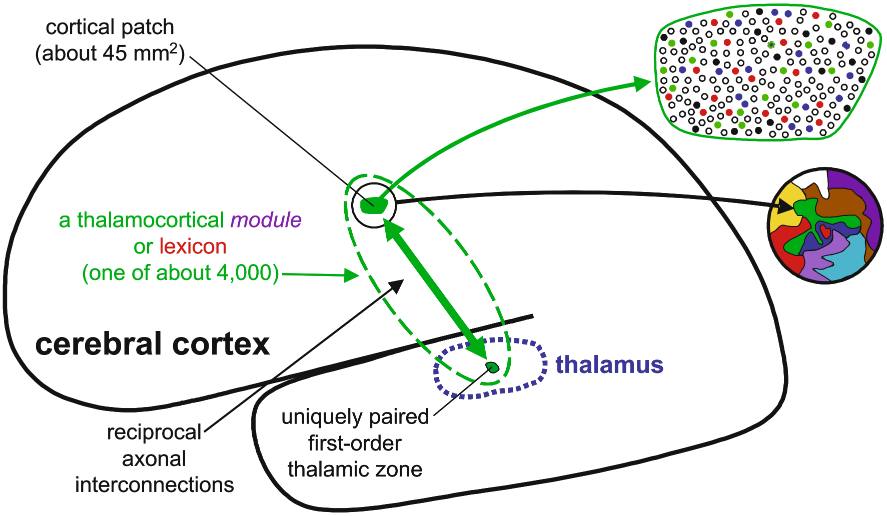


Fig. 5.1. A human thalamocortical *module*, as hypothesized by confabulation theory. Each module consists of a localized cortical patch extending through the full depth of cortex having a cortical surface area of roughly 45 mm^2 [out of a total of roughly 180,000 mm^2 for both hemispheres (Paxinos and Mai 2003)], and a small localized zone of first-order thalamus (Sherman and Guillery 2006; Casagrande et al. 2005) which is reciprocally axonally connected with the module's cortical patch. The upper-right enlarged notional depiction shows the neurons of layer III of the cortical patch of the module. These are the neurons which are hypothesized to represent the distinct symbols of the module. The "magnified" depiction beneath illustrates that actual modules are probably actually irregular in shape (each colored "blob" illustrating what an individual module may look like). When their function is being emphasized, rather than their physiology, modules are referred to as *modules*

cognitive mental universe may possess (see Fig. 5.2). This representation takes the form of the selection of a single *symbol* from among a set of (typically) thousands of symbols implemented by the module for describing its object attribute. For example, if a particular module is used to represent the English name of an object, it might be equipped with hundreds of thousands of symbols encoding everything from **aardvark** to **cloud** to **Milan** to **Zigmund**. When invoked (as part of a learned and stored thought process being deliberately executed), this module will typically *activate* one of its symbols to represent the name of the object being considered. All of its other symbols will be *inactive*.

Symbols are the durable, persistent *terms of reference* for describing the objects of the mental universe. Clearly, such fixed terms of reference must exist if knowledge is to be accumulated over long periods of time.

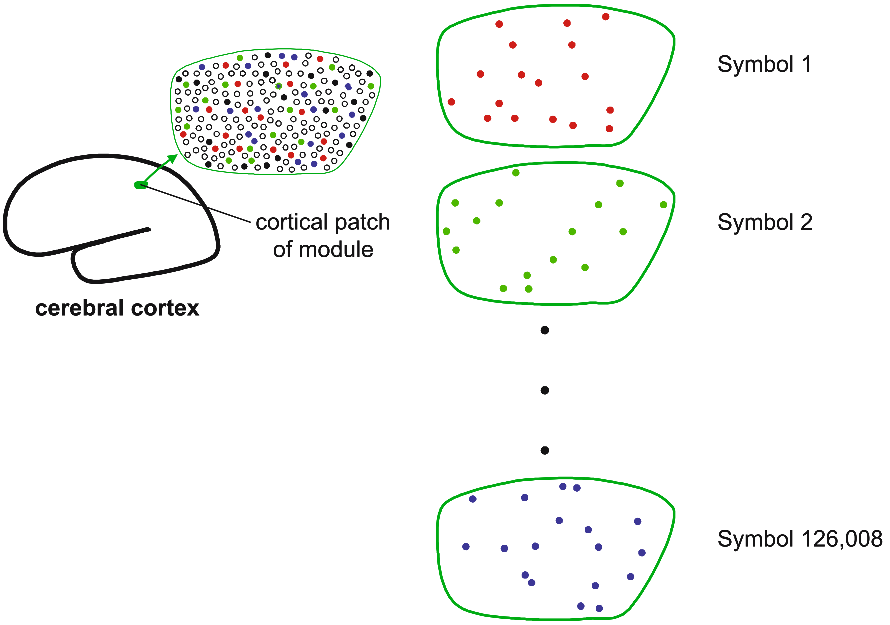


Fig. 5.2. Each module describes a single *attribute* that *objects* of the mental universe may possess. This description, when used, is in terms of selection of a single *symbol* (object attribute descriptor) from among a collection of (typically) thousands of distinct symbols implemented by the module (the particular module shown here is implementing 126,008 symbols). Each symbol is represented by a collection of roughly 60 *active* neurons (a sampling of which are shown for each symbol), each belonging to a special population of about 450,000 neurons (shown notionally in the enlarged depiction of the module to the upper right of the cortex illustration). The first key hypothesis of confabulation theory is that this is how the objects of the mental world are represented in cerebral cortex – by having selected attributes of the object each represented by a single active symbol in that attribute's module

Confabulation theory postulates that within each module's cortical patch there exists a population of neurons that function to represent symbols (Fig. 5.2). These neurons may reside within layer III; but their exact location is not important for the purposes of this synopsis. Since each square millimeter of full-depth cortex has roughly 100,000 neurons (Paxinos and Mai 2003), and this hypothesized symbol-representing neuron population contains roughly 10% of these, each module's 45 mm^2 cortical patch will possess about 450,000 of these symbol representation neurons. If a particular module implements, for example, 100,000 symbols, and each symbol of that module is represented by a collection of about 60 of these neurons (the reason why 60 are needed is discussed below), then it is easy to see that, on average, each symbol representation neuron will participate in representing about 13 different symbols.

The symbolic processing that is the heart of confabulation theory critically depends upon having the symbols of a module be functionally discrete and distinct – not fuzzy and overlapping. Worse yet, as will be discussed in Sect. 5.3 below, during confabulation, the symbols must compete with one another on the basis of their representing neuron's combined total input excitation. Thus, having each symbol representation neuron participate in representing many symbols would seem to be an invitation to "crosstalk" and "interference" between symbols. However, counterintuitively, such problems probably don't actually arise (Haines and Hecht-Nielsen 1988; Xie et al. 2001; Hahnloser et al. 2003).

5.3 Confabulation Theory Element 2: Knowledge Links

Confabulation theory hypothesizes that all aspects of cognition utilize a simple, uniform type of knowledge: antecedent support axonal links (Hecht-Nielsen 2004). Each such individual *knowledge link* (see Fig. 5.3) connects the neuron collection representing one symbol (termed the *source symbol* of the link) to neurons representing a second symbol (termed the *target symbol* of the link – usually a symbol in a different module from that of the source symbol).

Again, there are details and complications involved. First of all, these axonal links are not direct. They are probably implemented as two-stage Abeles synfire chains (Abeles 1991). The 60 neurons of the source symbol send axons to a million or more neurons scattered all over (many outside its module). Of these neurons, thousands receive sufficient synchronized input from multiple source symbol neurons to become highly *excited*. Thus, this first stage of the synfire chain "amplifies" the high activity of the 60 neurons representing the source symbol to high excitation of many thousands of *transponder* neurons (as these intermediate neurons of the chain are termed). Note that the synapses involved in this transponder neuron excitation process will, in general, not be strengthened, since the transponder neurons are typically not already active when the source symbol excitation arrives (thus failing the Hebb meaningful co-activity criterion for synapse strengthening). Of the thousands of transponder neurons that are excited by the momentary source symbol neuron activity, their statistical

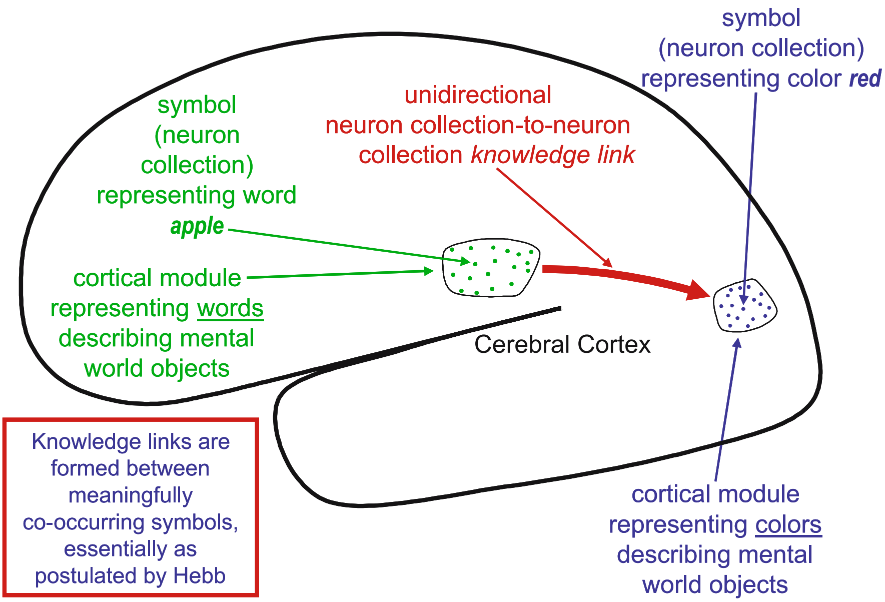

symbol (neuron collection) representing color *red*
unidirectional neuron collection-to-neuron collection *knowledge link*
symbol (neuron collection) representing word *apple*
cortical module representing words describing mental world objects
Cerebral Cortex
Knowledge links are formed between meaningfully co-occurring symbols, essentially as postulated by Hebb
cortical module representing colors describing mental world objects

Fig. 5.3. A *knowledge link*. Confabulation theory hypothesizes that all cognitive knowledge is stored in the form of these axonal communication links. Each individual knowledge link is between the collection of neurons representing a particular symbol (termed the *source symbol* of the knowledge link) and members of the collection of neurons representing a second symbol (termed the *target symbol* of the link). As presciently postulated by Hebb (Hebb 1949) in 1949, these pairwise neuron-collection-to-neuron-collection links are established on the basis that the involved source and target symbols are meaningfully active at the same time (this is termed *meaningful symbol co-occurrence*). The average human is hypothesized to possess many billions of knowledge links. That knowledge of such a simple kind can explain all of cognition is astounding, but that is precisely the second key hypothesis of confabulation theory

axonal distributions are assumed to be genetically programmed so that some reasonable fraction (say, 10%) of the neurons representing the target symbol will receive synapses from multiple (perhaps three to six) transponder neurons. During learning, these final link synapses will be strengthened, since the involved transponder neurons and the target symbol neurons will be meaningfully co-active. In other words, these synfire chain links from symbol to symbol will be formed only if Hebb's co-occurrence condition on the source and target symbols' neurons is met. [New knowledge links are immediately, but temporarily, established when the involved symbols first co-occur. Permanent strengthening, if warranted, then occurs over the following few sleep periods (Chap. 8; Sejnowski and Destexhe 2000).]

The set of all knowledge links connecting the symbols of one module to symbols of a second module is termed a *knowledge base*. For any knowledge link to be used, its knowledge base must be deliberately *enabled*. Knowledge base enablement is hypothesized by the theory as the primary function of higher-order thalamus (Sherman and Guillery 2006; Casagrande et al. 2005).

The average human is postulated to have many billions of individual knowledge links (each of which is termed an *item of knowledge*). This implies a learning rate well in excess of one link per second throughout life. If true, this will have profound implications for our views of human nature, education, etc. For example, a child returning home after a day at school might report that she "learned nothing" that day. In reality, she probably began the process of establishing tens of thousands of new knowledge links. Humans (and other animals) are extremely "smart."

Another implication of this knowledge link hypothesis is that in order for learning to take place "on demand" (i.e., without waiting many days for new, correctly connected axons to somehow form) there must probably be a vast "overwiring" of cortex to support immediate inauguration of the required synfire chain terminal synapses. Thus, confabulation theory also proposes that only a small fraction (roughly 1%) of the cortical synapses available for use in storing cognitive knowledge are actually used. Perhaps this is why so many excitatory cortical synapses seem "vestigial" and do not function reliably when tested with patch clamps. Synapses which are used to store cognitive knowledge are rare. Thus, the old saw that "we only use 10% of our brain" may need to be updated to: "we only use 1% of our knowledge synapses."

Modules are organized into *hierarchies*, one symbol in a higher-level module often representing multiple sets of particular symbols at lower levels. [Hierarchies are an old idea, the power of which is amply demonstrated by Fukushima's *Neocognitron* family of visual neural networks (Fukushima 2005, 1975; Fukushima et al. 1983).] In humans, the modules used to describe language object attributes form the core "hub" of cognition. These are connected (via knowledge links) to and from modules belonging to almost every other functional category of modules.

5.4 Confabulation Theory Element 3: Confabulation

Besides implementing a module of symbols for describing that module's mental object attribute; each module is also responsible for carrying out the *confabulation* operation (a "winners-take-all" competition process among the symbols of the module – see Fig. 5.4). Confabulation is postulated to be the only information-processing operation used in cognition. The hypothesized neuronal *attractor network* (Haines and Hecht-Nielsen 1988; Amari 1974; Anderson et al. 1977; Hopfield 1982; Cohen and Grossberg 1983; Kosko 1988; Amit 1989; Sommer and Palm 1999) mechanism by which modules (when deliberately commanded to do so by a single, graded, *thought-command signal* input to the module from an external source) carry out confabulation is illustrated in Fig. 5.9.

Figure 5.4 shows many symbols of the module (with symbol neurons notionally represented by colored circles having thicknesses directly related to their

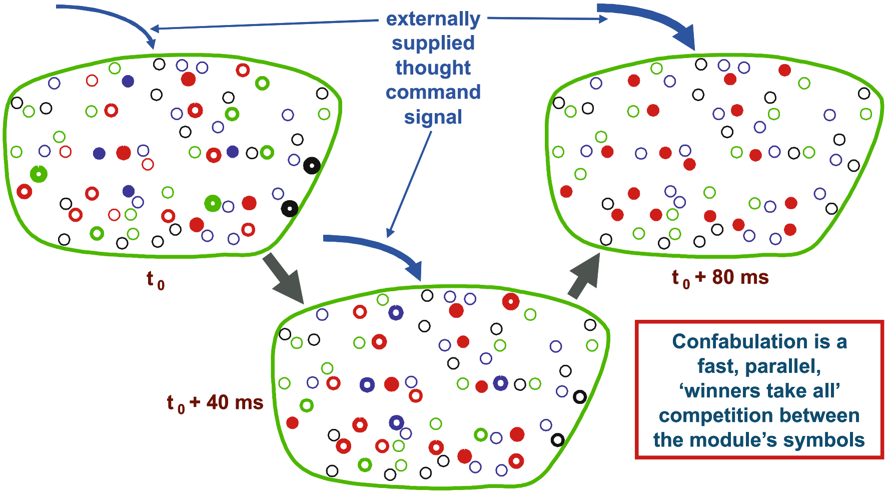


Fig. 5.4. When deliberately commanded, a module is hypothesized to function as a *neuronal attractor network*. Confabulation theory hypothesizes that all aspects of cognition are carried out using this single "information-processing operation," which is termed *confabulation* (this is the third key hypothesis of confabulation theory). Confabulation is a simple winners-take-all competition process among the symbols of a module (notionally illustrated here showing layer III of the module's cortical patch). Confabulation takes place only when the module receives a deliberate *thought-command signal* (which originates outside the module). The thought-command signal is analog (graded), not binary. At the starting time of the confabulation (here denoted by t_0), the thought-command signal level is low or zero. By rapidly increasing the strength of this command input (which arrives at all points of the module via a small number of parallel axons from an external source which ramify upon entering the module), the symbols compete with one another on the basis of their total excitation at t_0. The symbol with the highest initial total excitation wins the competition (in this case, the symbol represented by the red neurons). This symbol is termed the *conclusion* of the confabulation. Confabulation is often completed in about 80 ms. In light of the fact that each module is controlled by a single graded input (just as with the contraction of a muscle), modules can be viewed as the *muscles of thought*

excitation levels) receiving various levels of input excitation at a starting time t_0 from incoming knowledge links (as shown, each such knowledge link typically only effects a subset of each involved target symbol's neurons). Delivery of a thought-control signal input (illustrated in blue) of rapidly increasing amplitude (indicated by the line thickness of the blue arrow above the module in the figure) to the module then causes that symbol with the highest average level of total input excitation (among its 60 neurons) to end up having all 60 of its neurons highly active (for a brief moment). This symbol is termed the *conclusion* of the confabulation [confabulations can also end with multiple excited symbols or no symbol; but this complication is not discussed here (see Chap. 6)]. Confabulation involves a multitude of parallel, local interactions between the involved neurons during the roughly 80 ms required to complete the process.

In the confabulation theory view of cortical cognitive function (cortex does other things besides cognition, such as triggering behaviors – see next section), each module has its cognitive activity controlled by a small number of thought control input afferents from outside the module. Thus, the theory views a thalamocortical module as an exact analog of a muscle: a discrete "action unit" controlled by a single graded input. The exact origin of the thought control axonal inputs to modules is not known, but they probably arise in one or more sub-cortical nuclei (Herculano-Houzel et al. 1999; Fries et al. 2001; Freeman and Holmes 2005; Makeig et al. 2002). In summary, confabulation theory views modules as the *muscles of thought*.

5.5 Confabulation Theory Element 4: The Origin of Behavior

As illustrated in Fig. 5.5, confabulation theory hypothesizes that every time a confabulation operation carried out by any module yields a *definitive conclusion* (namely, one symbol – not multiple symbols or no symbol), a set of *action* (movement process and/or thought process) *commands* associated from that particular conclusion symbol are immediately launched. This explains the continual flow of behaviors that emerge, moment by moment, during wakefulness. In effect, each new behavior is a response to the latest update to the representation of the mental world state. Action commands originate in layer V of cortex and typically target sub-cortical nuclei such as motor or thought nuclei or the basal ganglia.

Most behaviors are small "housekeeping" functions (e.g., the next small segment of a movement or thought process); which are termed *microbehaviors*. Tens of microbehaviors are often implemented in one second. Higher-level behaviors (launched by conclusions reached on higher-level, more abstract, modules devoted to planning or large-scale behavior representation – often residing in frontal cortex), such as a decision to take a trip to Copenhagen, are launched much less frequently. Microbehaviors are typically executed instantly, whereas higher-level behaviors are often treated as "suggestions" and subjected to further scrutiny (e.g., by the basal ganglia and by cognitive modules executing plan evaluation thought processes) before being executed (or discarded).

In effect, behavior results (during wakefulness) from the action commands which are launched each time the state description of the mental world is updated (by activation of a new confabulation conclusion symbol in a module). Each successful confabulation launches the next set of action commands, and so on, endlessly until the next sleep period. The wizard homunculus standing behind the curtain pulling the levers of behavior is thereby exorcised. All non-reflexive and non-autonomic behavior is postulated to originate in this way (although many sub-cortical nuclei are involved in continuing, refining, and sustaining behaviors once they are started).

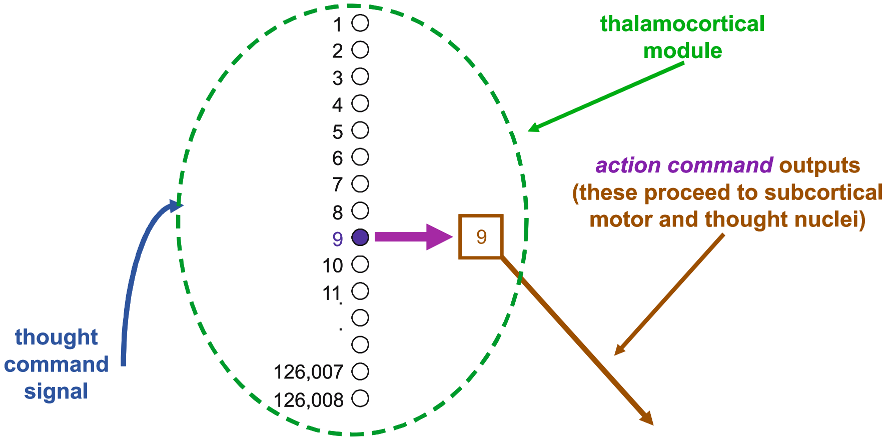

thalamocortical module
action command outputs (these proceed to subcortical motor and thought nuclei)
thought command signal
1
2
3
4
5
6
7
8
9
10
11
126,007
126,008
9

Fig. 5.5. The *conclusion → action principle* (the fourth and last key hypothesis of confabulation theory). Here, a module (illustrated, in consonance with Fig. 5.1, as an abstract "oval" structure containing a list of symbols to emphasize its module function) has successfully completed a confabulation operation (under control of its externally supplied thought-command signal) and reached a conclusion (symbol number 9). Whenever a module reaches a single conclusion it immediately causes a set of *action command* outputs to be launched (these outputs proceeding to sub-cortical brain nuclei from neurons in layer V of the module's cortical patch). The specific action command outputs that are launched are those which have been previously *associated from* this specific conclusion symbol via a completely separate, sub-cortically managed, *skill learning* process (Brown et al. 2004; Shibata et al. 2005). These action command outputs can cause behaviors to occur. The conclusion → action principle is hypothesized to be the origin of all non-autonomic and non-reflexive behavior. During wakefulness, many behaviors (most of them small "microbehaviors") are launched every second

5.6 Confabulation Theory Mathematics

Confabulation theory hypothesizes that the four key elements described above, are capable of explaining every aspect of cognition. Here, the underlying mathematics of confabulation is briefly discussed to see why this assertion may be tenable.

Confabulation theory proposes that the underlying mathematical process of cognition is *maximization of cogency* (see Chaps. 3 and 7). For example, consider four *assumed fact* symbols: α, β, γ, and δ (these are symbols being expressed on four different modules, as shown in Fig. 5.6; with each symbol transmitting excitation, via all the available knowledge links, to target symbols of a fifth *answer* module that is about to undergo confabulation). Confabulation theory hypothesizes that the fundamental underlying mathematical operation of cognition is to find that symbol ε of the answer module which maximizes cogency $p(\alpha\beta\gamma\delta|\varepsilon)$. Cogency is the probability of the assumed facts being true,

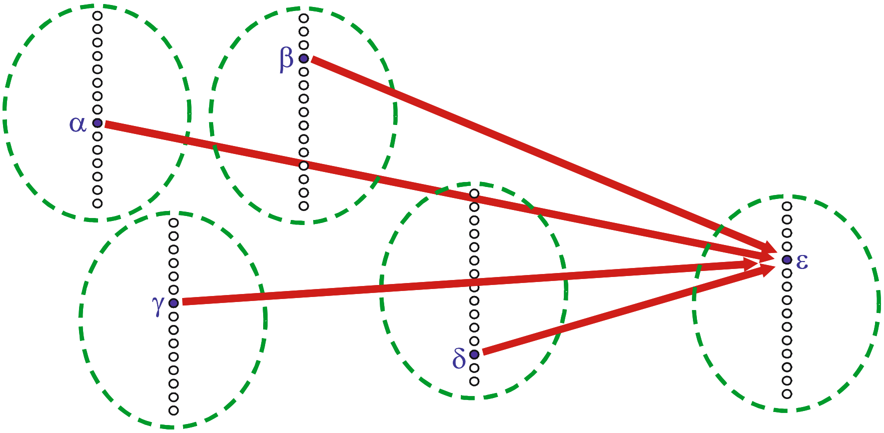


Fig. 5.6. Confabulation. Four modules, which have already reached conclusions α, β, γ, and δ in recently completed confabulations, are sending excitation via all available knowledge links from each of these symbols to the symbols of a fifth *answer* module that is about to be commanded to carry out confabulation. When completed, this next confabulation yields that symbol ε of the answer module with the highest total knowledge link excitation. The mathematics of confabulation shows that this conclusion ε will be that symbol which maximizes *cogency* p(αβγδ|ε). Confabulation is a general-purpose decision-making tool that animals use to carry out all aspects of cognition

given an assumption that the symbol ε is true. In other words, confabulation theory claims that each decision-making process involved in cognition is selection of that conclusion which is most supportive of the assumed facts being employed actually being true.

Cogency maximization is a radical departure from the "Bayesian" viewpoint that has dominated thinking in computational intelligence, computational neurobiology, and computer science for decades. This viewpoint initially arose in the work of R.A. Fisher 80 years ago. It was eventually established that, for any fixed system of object property measurement, the optimal pattern classifier is that which selects the class which has the highest probability of being the correct one, given the available measurements (Duda et al. 2000). This optimum classifier became known as the "Bayes classifier," because it involves a conditional probability which (Fisher argued) can be evaluated using Bayes' law from elementary probability theory. Since animals are excellent pattern classifiers, it became an article of faith that cognition must therefore be "Bayesian." This view was later expanded to the principle that the "best" conclusion to select in any situation will be the one which has the highest probability of being correct, given the available facts (Pearl 2000; Korb and Nicholson 2003). Although this "Bayesian" viewpoint has a strong intuitive appeal, and has yielded a wide range of valuable technological applications, it is an incorrect model of cognition (see Chap. 3).

An important property of cogency maximization (rigorously proven in Chap. 4) is that in a "logical information environment" (playing chess, doing mathematics, etc.) it yields the same result as classical Aristotelian logic:

Theorem 5.1: If $\alpha\beta\gamma\delta \Rightarrow \varepsilon$ uniquely, then ε uniquely maximizes cogency $p(\alpha\beta\gamma\delta|\varepsilon)$. □

Thus, the cogency maximization hypothesis implies that cognition is "logical" when that is possible (most of the information environments we encounter are not logical). In a "non-logical" environment (e.g., when parking your car), cogency maximization just picks the conclusion that best supports the probability of the available facts being true, and then moves on (there is no "logical guarantee" that the conclusions reached are correct, but that is not important).

As shown in Chap. 4:

Theorem 5.2 [*The Fundamental Theorem of Cognition*]: Given non-exceptional assumed facts α, β, γ, and δ, and expectation element ε, then the following exact relationship holds between cogency $p(\alpha\beta\gamma\delta|\varepsilon)$ and the confabulation product $p(\alpha|\varepsilon)\cdot p(\beta|\varepsilon)\cdot p(\gamma|\varepsilon)\cdot p(\delta|\varepsilon)$:

$$\begin{aligned}[p(\alpha\beta\gamma\delta|\varepsilon)]^4 = {} & [p(\alpha\beta\gamma\delta\varepsilon)/p(\alpha\varepsilon)] \\ & \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\beta\varepsilon)] \\ & \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\gamma\varepsilon)] \\ & \cdot [p(\alpha\beta\gamma\delta\varepsilon)/p(\delta\varepsilon)] \\ & \cdot [p(\alpha|\varepsilon)\cdot p(\beta|\varepsilon)\cdot p(\gamma|\varepsilon)\cdot p(\delta|\varepsilon)],\end{aligned}$$ □

confabulation (which maximizes the quantity $[p(\alpha|\varepsilon)\cdot p(\beta|\varepsilon)\cdot p(\gamma|\varepsilon)\cdot p(\delta|\varepsilon)]$) can, under the mild mathematical condition that the product of the first four terms of the right-hand side of this equation is approximately constant for all viable conclusions ε (a condition which confabulation theory postulates animal neurological evolution has been able to satisfy), approximately maximize cogency. This is important, because cogency maximization itself cannot be carried out in practice.

In summary, the central mathematical discovery of confabulation theory is that, under the mathematical condition of Theorem 5.2, which brain evolution is hypothesized to have found ways to satisfy hundreds of million years ago, there is no need for *a priori* probabilities in cognition. (except in the *antecedent support* probability weightings of knowledge links, which cortical synapses seem to be able to implement; Chaps. 3 and 8). By applying an axonally implementable, strictly monotonic logarithmic transformation to the confabulation product of Theorem 5.2 [see (Hecht-Nielsen 2006) for details], the neurons of symbol λ of a confabulating module will be receiving total input excitation $I(\lambda)$ from the four knowledge links arriving from (in the specific example case considered in Theorem 5.2) assumed fact symbols α, β, γ, and δ, where:

$$\begin{aligned}I(\lambda) \equiv {} & [\ln(p(\alpha|\lambda)/p_0) + B] + [\ln(p(\beta|\lambda)/p_0) + B] \\ & + [\ln(p(\gamma|\lambda)/p_0) + B] + [\ln(p(\delta|\lambda)/p_0) + B].\end{aligned} \tag{5.1}$$

Thus, by eliminating *a priori* probabilities, and using axonal knowledge links with logarithmically-transformed pairwise symbol conditional (antecedent support) probability weighting (e.g., $[\ln(p(\alpha|\lambda)/p_0) + B]$ for the knowledge link from symbol α to symbol λ in the above equation), an astoundingly simple cogency-maximizing winners-take-all competition, in which millions of relevant items of knowledge are automatically and effectively applied in parallel, via the above simple additive knowledge-combination law, can be used to implement all cognitive functions at blazing speed.

5.7 Multiconfabulation

My colleagues at Fair Isaac Corporation and I have conducted two types of computer simulation experiments to explore multiconfabulation (see Chap. 6 for complete details): *sentence continuation without context*, and *sentence continuation with context* (Fig. 5.7). In each experiment, an ordered sequence of three words (termed the starter – shown in blue in Fig. 5.7) was supplied. Execution of the thought process then caused four phrases to be appended to this starter, the first four words of which were retained as the continuation. Some experimental results are shown in Fig. 5.8. Below, I briefly describe the confabulation architecture used in the experiments, the learning process employed to prepare it for use, and the thought process which was applied to create the continuations.

Each square in Fig. 5.9 represents a single computer-simulated module (hereinafter, module). The 82-module confabulation architecture employed in the experiments consists of two vertical groupings of modules. The context sentence

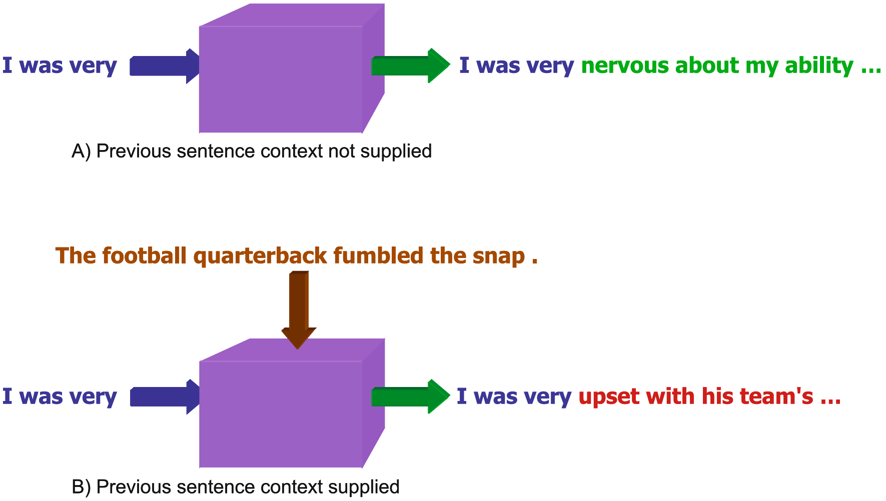

I was very
I was very nervous about my ability ...
A) Previous sentence context not supplied
The football quarterback fumbled the snap .
I was very
I was very upset with his team's ...
B) Previous sentence context supplied

Fig. 5.7. The two types of experiments carried out. **A / B** Sentence continuation without / with a previous context sentence being supplied. The "purple box" is the confabulation architecture used to create the continuations

The New York Times' computer model collapses ...
Stocks proved to be a wise investment .
The New York markets traded lower yesterday ...
Downtown events were interfering with local traffic .
The New York City Center area where ...
Coastal homes were damaged by tropical storms .
The New York City Emergency Service System ...
Medical patients tried to see their doctors .
The New York University Medical Association reported ...

When the United Center Party leader urged ...
The car assembly lines halted due to labor strikes .
When the United Auto Workers union representation ...
The price of oil in the Middle East escalated yesterday .
When the United Arab Emirates bought the ...

But the Roman Empire disintegrated during the ...
She learned the history of the saints .
But the Roman Catholic population aged 44 ...
She studied art history and classical architecture .
But the Roman Catholic church buildings dating ...

I was very nervous about my ability ...
Democratic citizens voted for their party's candidate .
I was very concerned that they chose ...
Restaurant diners ate meals that were served .
I was very hungry while knowing he ...

In spite of yesterday's agreement among analysts ...
The Mets were not expected to win .
In spite of the pitching performance of ...
The President was certain to be reelected .
In spite of his statements toward the ...
She had no clue about the answer .
In spite of her experience and her ...

It meant that customers could do away ...
The stock market had fallen consistently .
It meant that stocks could rebound later ...
I was not able to solve the problem .
It meant that we couldn't do much ...
The company laid off half its staff .
It meant that if employees were through ...
The salesman sold men's and women's shoes .
It meant that sales costs for increases ...

Fig. 5.8. Sample of experimental results. Using the same color scheme as Fig. 5.7, the first line of each text block is the result of a continuation trial without context. The subsequent lines are a supplied context sentence followed by the same starter and the continuation with context

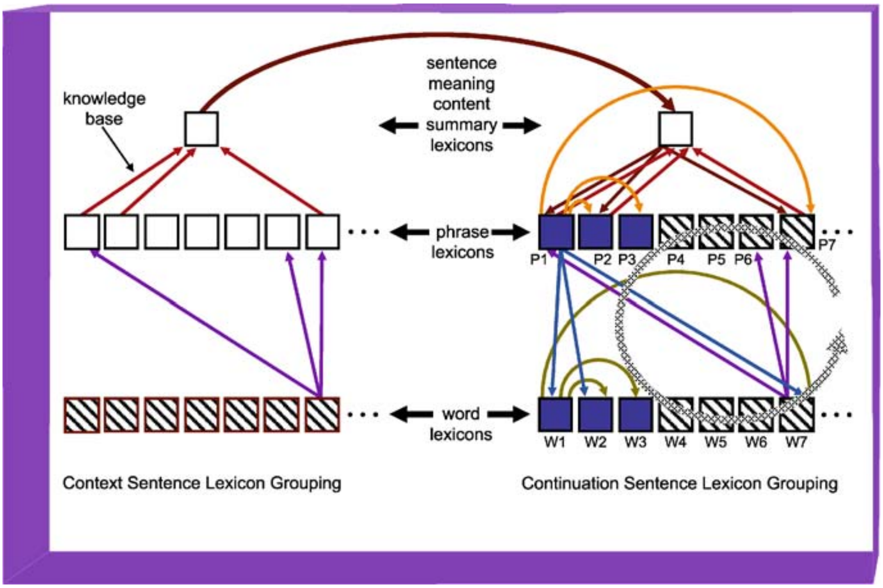

knowledge base
sentence meaning content summary lexicons
phrase lexicons
word lexicons
P1
P2
P3
P4
P5
P6
P7
W1
W2
W3
W4
W5
W6
W7
Context Sentence Lexicon Grouping
Continuation Sentence Lexicon Grouping

Fig. 5.9. Inside the "purple box" of Fig. 5.7: the confabulation architecture employed in the experiments (see text). The swirling red arrow indicates the sentence continuation thought process, during which the thought-control signals to the "unlocked" phrase modules (i.e., P4, P5, P6, and P7 – those not used to re-represent the starter at the phrase level) and word modules (W4, W5, W6, and W7) of the continuation grouping are progressively "tightened" to yield the confabulation consensus

grouping on the left is used to represent the context sentence, when one is employed. The continuation sentence grouping on the right is used to represent the starter and to carry out the thought process to obtain the confabulation consensus; which yields the continuation.

The 20 word modules comprising the bottom row of both groupings (of which the first seven are illustrated in Fig. 5.9) are used to represent individual English words or punctuation marks of a sentence (in order, from left to right). Each word module has 63,008 symbols representing common words and punctuations. Each punctuation mark is treated as a separate word. Capitalized words are represented by their own separate symbols (e.g., **exit** and **Exit** have separate symbols). In both groupings, a middle row of 20 phrase modules (of which the first seven are shown in Fig. 5.3) and the one top-level sentence meaning content summary module each have 126,008 symbols representing common words, multi-word phrases, and punctuations in English.

An individual symbol of a module can be unidirectionally connected to an individual symbol of certain other modules via a knowledge link. The ordered pairs of modules for which knowledge links are allowed (indicated by the knowledge base arrow patterns in Fig. 5.9), are now described.

In the context sentence grouping: each phrase module receives knowledge bases from all word modules except those that lie to its left, and the summary module receives knowledge bases from each of the phrase modules. In the continuation sentence grouping, every word (phrase) module receives knowledge bases from every word (phrase) module on its left; every phrase (word) module receives a knowledge base from every word (phrase) module, except those on its left (right); and the summary module sends knowledge bases to, and receives knowledge bases from, each of the phrase modules. Finally, a knowledge base links the summary module of the context grouping to that of the continuation sentence grouping. Thus, there are a total of 1,071 knowledge bases in this architecture.

Learning consisted of two phases: exposure of the continuation grouping of modules alone to single sentences (124 million examples), followed by exposure of the whole architecture to meaning-coherent pairs of successive sentences (70 million examples). Following completion of the first phase of learning and before the second began, the context grouping knowledge base collections shown with the same color in Fig. 5.9 were copied, in order, from their corresponding mates of the continuation grouping. The 70 million sentence pair examples were then used to build the knowledge base linking the context grouping summary module to the continuation grouping summary module. The examples used in learning were drawn from a varied proper English corpus containing novels, encyclopedias, news stories, etc.

In the initial learning phase, each of the single sentences was represented by the modules of the continuation grouping. The counts of pairwise symbol co-occurrences in pairs of modules linked by a knowledge base were accumulated. After all 124 million example sentences had been entered, those symbol pairs with meaningful co-occurrence counts had knowledge links established for them. Finally, during the second phase of learning, the meaning-coherent sentence pairs were entered and represented in temporal order on the two module groupings and symbol co-occurrence counts were accumulated for the summary module to summary module knowledge base. The knowledge links of this knowledge base were then constructed from these counts in the same manner as in the first learning phase. Learning required about a month using a desktop computer and yielded an average of about 2.5 million links per knowledge base.

As illustrated by the swirling red arrow in Fig. 5.9, the sentence continuation thought process involves a cyclic "tightening" of ongoing, parallel, unlocked phrase (P4, P5, P6, and P7) and word (W4, W5, W6, and W7) module confabulations until, progressively in temporal order, each phrase module confabulation converges to a single conclusion (together, the conclusions constitute the *confabulation consensus*). At the beginning of this process, billions of feasible permutations of four phrases in sequence are being considered in parallel. At the end, only one of these remains. After all phrase module confabulations have converged, the four consensus conclusions (the continuation) are read out from word modules W4, W5, W6, and W7. Sample continuations are shown in Fig. 5.8.

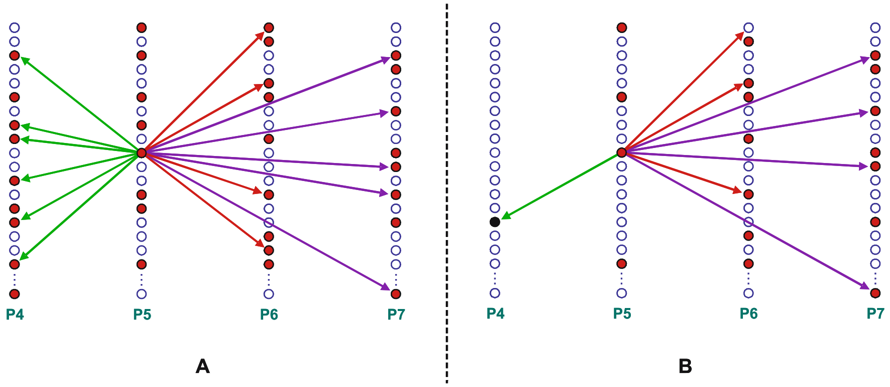


Fig. 5.10. Multiconfabulation; explanation for the power of thought. Before beginning a multiconfabulation, available external context can be used to restrict subsequent processing to not-unacceptable sets of "viable" symbols (red-filled symbols). Multiconfabulation then invokes knowledge links from each viable symbol to symbols of the other modules involved in the multiconfabulation. Here, the four confabulating phrase modules are illustrated (with details simplified). A small subset of the links from only one viable symbol on P5 to viable symbols on other modules is illustrated in **A** (each viable symbol, of which each module frequently has thousands, often has links to thousands of viable symbols on each other module). These knowledge links are shown in colored groups to illustrate that the set of all links from one module to a second module forms a knowledge base. Each knowledge base is independently enabled during multiconfabulation. At the beginning of multiconfabulation, these millions of relevant knowledge links (each connecting a viable symbol to another viable symbol) are being employed in parallel to create the *confabulation consensus*. As this sentence continuation multiconfabulation progresses (**B**), module P4 converges to a conclusion (black symbol) first, which then, by deactivating all knowledge links from (and to) all other P4 symbols, significantly restricts the other modules' remaining sets of viable conclusions. Eventually, only one symbol on each of W4, W5, W6, and W7 (see Fig. 5.9 and text) remains

The sentence continuation thought process implemented in these experiments is a discrete-time approximation to the continuous-time process that is hypothesized to take place in human cortex. First, context is applied via knowledge links brought in from the words and phrases of the starter and (if present) from the context sentence. These inputs, and the module tightenings that briefly follow it, serve to establish expectations (Hecht-Nielsen 2006) on each of the four involved phrase modules. This establishes the "initial configuration" shown in Fig. 5.10A.

The function of the thought process is to continue the sentence by selecting an ordered sequence of four word symbols from the tens of billions of candidate sequences typically included within the initial configuration. As multiconfabulation proceeds, knowledge links from all eight other modules are used to identify those symbols which are the most highly excited. Low-excitation symbols are dropped as the number of symbols on the expectation is gradually reduced (i.e., the thought control input level to the module is *tightened*).

The goal of the thought process is to complete each phrase module's confabulation in its proper order (P4, P5, P6, and P7). As each module's expectation shrinks, the combinatorial collection of possible continuations plummets rapidly, as illustrated in Fig. 5.10B. This ongoing "mutual consultation" process ensures that the final continuation chosen is that one which has one of the "highest levels of consistency" with both the external context and between its individual conclusions.

5.8 Discussion

The excellent "grammar" and "syntax" seen in the experiments reported here suggests that these exist only as emergent properties of confabulation. If so, then the central conjectures of Miller, of Lenneberg, and of Chomsky (essentially that language reflects the universal human cognitive mechanism) are correct (Chomsky 1980).

As with the properly coordinated and phased contractions of muscles during a successful movement, executing a successful thought process requires a properly coordinated and phased set of "contractions" of the involved modules to yield a confabulation consensus. Thinking is like moving; with modules functioning as the "muscles of thought." Confabulation theory postulates that thought is a phylogenetic outgrowth of movement.

In the experiments presented here, the thought process causes the answer modules to converge in temporal order. However, thought process convergence can be made to occur in any desired order. This dynamic "convergence control" aspect of thought processes makes it meaningless to imagine a rigorous universal mathematical criterion for selecting confabulation consensuses. Each individual module converges to that symbol with the highest confabulation product, but the thought process, by controlling the order (and rate) of module convergence, and thereby the target symbol confabulation products considered, determines the final consensus. Thus, animal thought processes in which there are many viable confabulation consensuses are inherently non-deterministic, endowing life with delightful unpredictability. Confabulation theory should keep neurophilosophers busy for decades.

Finally, note that the acquisition of a new thought process (here, sentence continuation with or without a context sentence) is primarily dependent upon having a supply of products (examples available for learning of pairwise symbol co-occurrences) of that thought process being carried out by an existing competent practitioner (in this experiment, pairs of consecutive sentences produced by skilled human writers). This is the "monkey-see/monkey-do" principle of confabulation theory (Chaps. 1 and 8). Developing the associated thought processes (confabulation consensus convergence control maneuvers) seems relatively easy, as the crude thought process used in these experiments illustrates. Thus, besides illustrating the universal mechanism of thought, these experiments may be providing us deep insight into the fundamental nature of animal thought process acquisition. This points up the enormous treasure that is our accumulated cultural and intellectual legacy – much of which exists solely in the fragile form of expertise possessed by a relatively small subset of individual humans.

The role of synchrony in confabulation theory is seen in the function of knowledge links and in the details of confabulation (see Chaps. 3, 6 and 7). An interesting alternative is *polysynchrony* (Izhikevich 2006, 2007).

All cognizing species are presumed to have functional analogs of the human structures and mechanisms described here (Karten 1991).

6 The Mechanism of Thought

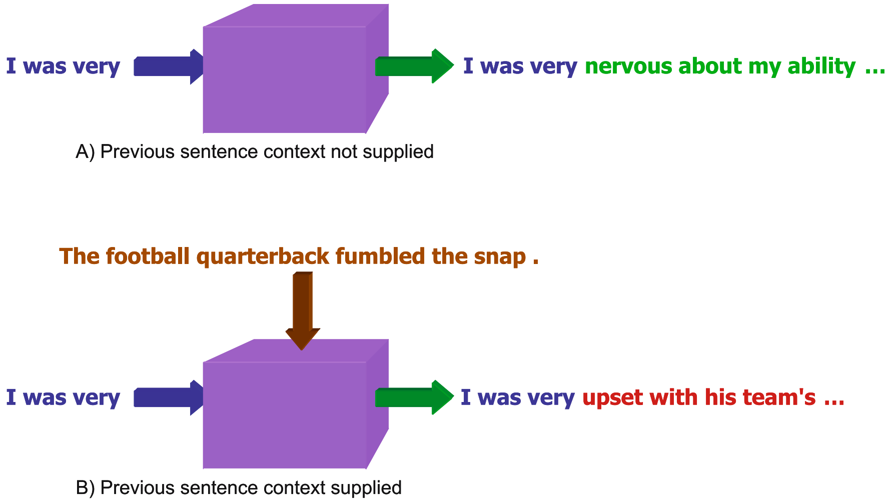
Robert Hecht-Nielsen[6]
Robert W. Means[7]
Kate Mark[7]
Syrus C. Nemat-Nasser[7]
Luke Barrington[7]
Andrew Smith[7]

A winners-take-all competition process, termed confabulation (Chaps. 1–5), is proposed as the fundamental mechanism of all aspects of cognition (vision, hearing, planning, language, control of thought and movement, etc.). Here, multiple, contemporaneous, mutually interacting computer-simulated confabulations – in which millions of items of relevant knowledge are applied in parallel – are considered. At the beginning of such a *multiconfabulation*, billions of distinct, potentially viable conclusion sets are considered. At the end, only one remains. This massively parallel, additive application of relevant knowledge is hypothesized to be the core explanation for the effectiveness, flexibility, and speed of thought.

6.1 Introduction

We conducted two types of computer simulation experiments to explore multiconfabulation: *sentence continuation without context*, and *sentence continuation with context* (Fig. 6.1). In each experiment, an ordered sequence of three words (termed the continuation sentence *starter* – shown in blue) was supplied. Execution of the thought process then caused four phrases to be appended to this starter, the first four words of which were retained as the continuation. Some experimental results are shown in Fig. 6.2. Below, we briefly describe the confabulation architecture used in the experiments, the learning process employed to prepare it for use, and the thought process which was applied to create the continuations (see Appendixes 6.A and 6.B for complete details).

[6] Computational Neurobiology, University of California, San Diego, La Jolla 92093-0407, USA
[7] Fair Isaac Corporation, San Diego, California 92130, USA


I was very
I was very nervous about my ability ...
The football quarterback fumbled the snap .
I was very
I was very upset with his team's ...

A) Previous sentence context not supplied

B) Previous sentence context supplied

Fig. 6.1. The two types of experiments carried out. **A**) / **B**) Sentence continuation without / with a previous context sentence being supplied. The "purple box" is the confabulation architecture used to create the continuations

Confabulation theory (Chaps. 1, 3, 4, 5, 7 and 8) hypothesizes that human cerebral cortex, and its locally paired first-order thalamus (Sherman and Guillery 2006; Casagrande et al. 2005), are divided into about 4,000 distinct, geographically localized, functional *thalamocortical modules* [which presumably have analogs in all other cognizing species (Karten 1991)]. Each individual module is devoted to describing one *attribute* that an object of the mental world may possess. To carry out this description, each module implements thousands of permanent *symbols* (most established in childhood), each symbol being represented by roughly 60 neurons belonging to a specialized population within the cortical portion of the module. When a particular module is being used to describe an object (most modules are hypothesized to typically use only one symbol at a time for this purpose), the neurons representing the symbol being employed are firing actively and synchronously, and the other symbol-representation neurons of the module are not.

When deliberately and individually commanded, a module implements an information-processing operation termed *confabulation* (Chaps. 1, 3 and 4): a fast, parallel, winners-take-all competition process [presumed to be carried out by a neuronal attractor network (Anderson et al. 1977; Kosko 1988; Haines and Hecht-Nielsen 1988; Hahnloser et al. 2003; Xie et al. 2001; Sommer and Palm 1999; Amit 1989; Cohen and Grossberg 1983; Hopfield 1982; Amari 1974) implemented within the module] between the symbols, with the winner being that symbol which is currently receiving the highest sum of knowledge link input excitation. Confabulation theory contends that this simple information-processing operation can explain all of cognition.

The New York Times' computer model collapses ...
Stocks proved to be a wise investment .
The New York markets traded lower yesterday ...
Downtown events were interfering with local traffic .
The New York City Center area where ...
Coastal homes were damaged by tropical storms .
The New York City Emergency Service System ...
Medical patients tried to see their doctors .
The New York University Medical Association reported ...

When the United Center Party leader urged ...
The car assembly lines halted due to labor strikes .
When the United Auto Workers union representation ...
The price of oil in the Middle East escalated yesterday .
When the United Arab Emirates bought the ...

But the Roman Empire disintegrated during the ...
She learned the history of the saints .
But the Roman Catholic population aged 44 ...
She studied art history and classical architecture .
But the Roman Catholic church buildings dating ...

I was very nervous about my ability ...
Democratic citizens voted for their party's candidate .
I was very concerned that they chose ...
Restaurant diners ate meals that were served .
I was very hungry while knowing he ...

Fig. 6.2. Sample of experimental results (see Appendix 6.B for the complete listing). Using the same color scheme as Fig. 6.1, the first line of each text block is the result of a continuation trial without context. The subsequent lines are a supplied context sentence followed by the same starter and the continuation with context

6.2 Confabulation Architecture

Figure 6.3 shows the confabulation architecture used in the experiments. Each square in Fig. 6.3 represents a single computer-simulated module (for complete details see Appendix 6.A at the end of this Chap.). This 82-module confabulation architecture consists of two vertical groupings of modules. The context sentence grouping on the left is used to represent the context sentence, when one is employed. The continuation sentence grouping on the right is used to represent the starter and to carry out the thought process to obtain the confabulation consensus, which yields the continuation.

The 20 word modules comprising the bottom row of both groupings (of which the first seven are illustrated in Fig. 6.3) are used to represent individual English words or punctuation marks of a sentence (in order, from left to right). Each word module has 63,008 symbols representing common words and punctuations. Each punctuation mark is treated as a separate word.

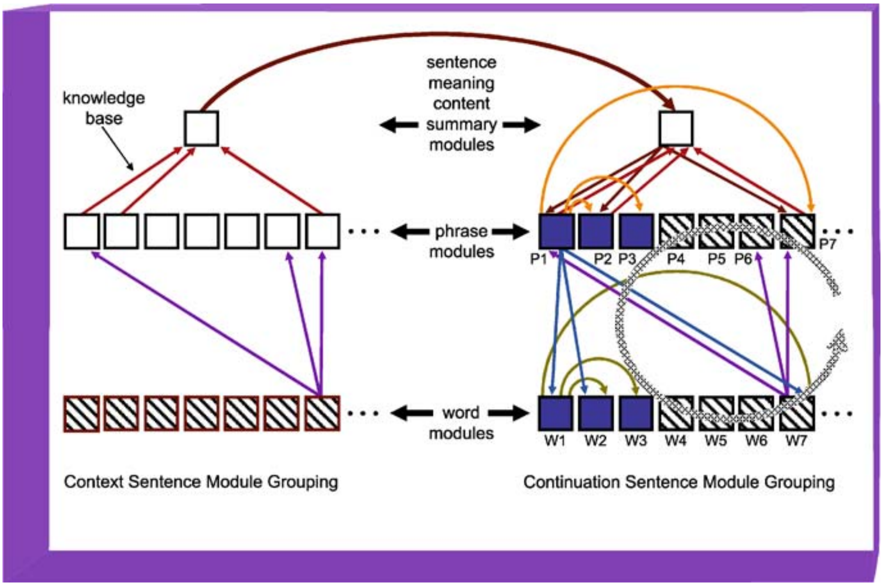

knowledge base
sentence meaning content summary modules
phrase modules
word modules
P1
P2
P3
P4
P5
P6
P7
W1
W2
W3
W4
W5
W6
W7
Context Sentence Module Grouping
Continuation Sentence Module Grouping

Fig. 6.3. Inside the "purple box" of Fig. 6.1: the confabulation architecture employed in the experiments (see text). The swirling red arrow indicates the sentence continuation thought process. The thought-control signals to both the "unlocked" phrase modules (i.e., P4, P5, P6, and P7 – those not used to re-represent the starter at the phrase level) and word modules (W4, W5, W6, and W7) of the continuation grouping are progressively "tightened" to eventually yield the confabulation consensus

Capitalized words are represented by their own separate symbols (e.g., **exit** and **Exit** have separate symbols). In both groupings, a middle row of 20 phrase modules (of which the first seven are shown in Fig. 6.3), and the one top-level sentence meaning content summary module, each have 126,008 symbols representing common words, multi-word phrases, and punctuations in English.

An individual symbol of a module can be unidirectionally connected to an individual symbol of certain other modules via a *knowledge link* [the basic unit of knowledge in a confabulation architecture – each envisioned as a two-stage synfire chain (Abeles 1991; Chaps. 3, 5, 7 and 8) in human cortex]. The bundle of all knowledge links between the symbols of each such ordered module pair is termed a *knowledge base*. As discussed further below, knowledge links are formed in response to meaningfully repeated symbol co-occurrences in bodies of English text presented to, and represented within, the architecture during learning. The ordered pairs of modules for which knowledge links are allowed (indicated by the knowledge base arrow patterns in Fig. 6.3), are now described.

In the context sentence grouping: each phrase module receives knowledge bases from all word modules except those that lie to its left, and the summary module receives knowledge bases from each of the phrase modules. In the continuation sentence grouping: every word (phrase) module receives knowledge bases from every word (phrase) module on its left; every phrase (word) module receives a knowledge base from every word (phrase) module, except those on its left (right); and the summary module sends knowledge bases to, and receives knowledge bases from, each of the phrase modules. Finally, a knowledge base links the summary module of the context grouping to that of the continuation sentence grouping. Thus, there are a total of 1,071 knowledge bases in this architecture.

6.3 Learning

Learning consisted of two phases: exposure of the continuation grouping of modules alone to single sentences (124 million examples), followed by exposure of the whole architecture to pairs of immediately successive sentences (70 million sentence pair examples). Following completion of the first phase of learning and before the second began, the context grouping knowledge base collections shown with the same color in Fig. 6.3 were copied, in order, from their corresponding mates of the continuation grouping. The 70 million sentence pair examples were then used to build the knowledge base linking the context grouping summary module to the continuation grouping summary module. The examples used in learning were drawn from a varied proper English corpus containing novels, encyclopedias, news stories, etc.

In the initial learning phase, each of the single sentences was represented by the modules of the continuation grouping. The counts of pairwise symbol co-occurrences in pairs of modules linked by a knowledge base were accumulated. After all 124 million example sentences had been entered, those symbol pairs with meaningful co-occurrence counts had [essentially as suggested by Hebb (Hebb 1949) – see Appendix 6.A for details] knowledge links established for them. Finally, during the second phase of learning, the meaning-coherent sentence pairs were entered and represented in temporal order on the two module groupings and symbol co-occurrence counts were accumulated for the summary module to summary module knowledge base. The knowledge links of this knowledge base were then constructed from these counts in the same manner as in the first learning phase. Learning required about a month using a desktop computer and yielded an average of about 2.5 million links per knowledge base.

6.4 Thought Process

As illustrated by the swirling red arrow in Fig. 6.3, and now summarized using Fig. 6.4, the sentence continuation thought process involves a cyclic "tightening" of ongoing, parallel, unlocked phrase (P4, P5, P6, and P7) and word (W4, W5, W6, and W7) module confabulations until, progressively in temporal order, each phrase module confabulation converges to a single conclusion (which

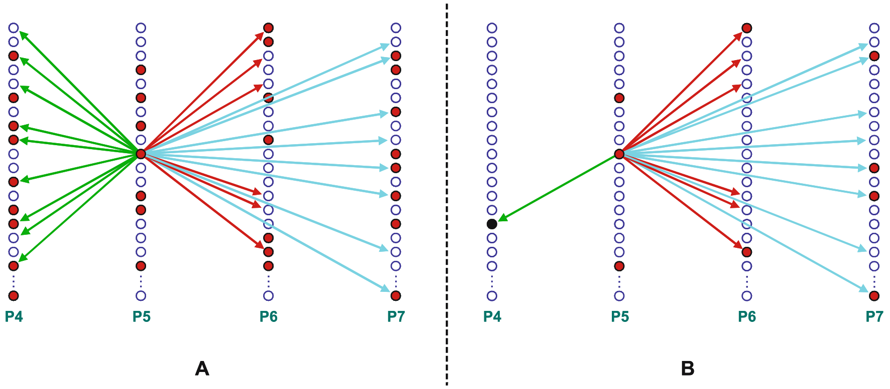


Fig. 6.4. Multiconfabulation; explanation for the power and speed of thought. Here, the four confabulating phrase modules are illustrated (with details simplified). Before beginning a multiconfabulation, available external context can be used to restrict subsequent processing to not-unacceptable sets of *viable* symbols (red-filled circles in **A**). Multiconfabulation then invokes knowledge links from each viable symbol to symbols of the other modules involved in the multiconfabulation. A small subset of the links from only one viable symbol on P5 to symbols on other modules is illustrated in **A** (each viable symbol, of which each module frequently has thousands, often has links to hundreds of symbols on each other module). These knowledge links are shown in colored groups to illustrate that the set of all links from one module to a second module form a *knowledge base*. Each knowledge base is independently deliberately *enabled*, and often employed multiple times, during a multiconfabulation. In the beginning stages of a typical multiconfabulation (**A**), millions of these *relevant* knowledge links (meaning that each such knowledge link emanates from a viable symbol on one of the modules) are often being employed. As the multiconfabulation progresses (**B**), module P4 converges to a conclusion (black symbol) first, which then automatically deactivates all knowledge links from all other P4 symbols. This significantly restricts, and reorders, the other modules' remaining sets of viable conclusions (*expectations*). Eventually, only one symbol on each of W4, W5, W6, and W7 (see Fig. 6.3 and text) remains: the *confabulation consensus*, which, in these experiments, is the sentence continuation

jointly constitute the confabulation consensus). At the beginning of this process, billions of feasible permutations of four phrases in sequence are being considered in parallel. At the end, only one of these remains. After all phrase module confabulations have converged, the four consensus conclusions (the continuation) are read out from word modules W4, W5, W6, and W7. Sample continuations are shown in Fig. 6.2.

The sentence continuation thought process implemented in these experiments is a discrete-time approximation to the continuous-time process that confabulation theory hypothesizes takes place in human cortex and thalamus. First, context is applied via knowledge links brought in from the words and phrases of the starter and (if present) from the context sentence. These inputs, and the module tightenings that briefly follow them, serve to establish *expectations* (Chaps. 3, 5 and 9) (subsets of symbols that future steps in the thought process will be restricted to) on each of the four involved phrase modules. This establishes the "initial configuration" shown in Fig. 6.4A.

The function of the thought process is to continue the sentence by selecting an ordered sequence of four word symbols from the tens of billions of candidate sequences typically included within the initial configuration. As multiconfabulation proceeds, knowledge links from subsequent phrase and word modules are used by each confabulating module to identify those symbols, from among the ever-shrinking expectation sets, which are the most highly excited. Low-excitation symbols are dropped as the number of symbols on each module's expectation is gradually reduced (i.e., the thought control input level to the module is *tightened*).

The key to this convergence is cyclic application of millions of knowledge links, which causes elimination of symbols which were (earlier in the thought process) considered potentially viable conclusions, but which this newly applied knowledge does not support. These knowledge link applications also serve to reorder the remaining viable symbols in terms of their updated approximate cogencies.

The goal of the particular thought process employed in these experiments is to complete each phrase module's confabulation in its proper order (P4, P5, P6, and P7). As each module's expectation shrinks, the combinatorial collection of remaining possible continuations plummets rapidly, as illustrated in Fig. 6.4B. This ongoing "mutual consultation" process ensures that the final confabulation consensus (i.e., sentence continuation) chosen is that one which has one of the "highest levels of consistency" with both the external context and between its individual conclusions.

6.5 Discussion

Confabulation architectures differ from past information-processing systems in that they do not employ algorithms, software, rules, priors, ontologies, etc. Their capabilities derive entirely from co-occurrence-based knowledge links obtained from exposure to raw data (proper English text in the experiments described here) and a simple mutually interacting multiple confabulation convergence (multiconfabulation) procedure.

In the first two examples of Fig. 6.2, note that the only difference was the presence of a context sentence in the second. The knowledge link inputs providing the context sentence meaning content to the multiconfabulation that produced the second example's continuation were essentially just "added in" (see Appendix 6.A for details). This illustrates that applying additional context information or constraints to a thought process is a simple manner of providing additional knowledge link input. Past use of that knowledge in connection with confabulation on a particular module need never have occurred. Thus, the use of an *additive knowledge combination law* (see Appendix 6.A) is surely another key secret to the extreme generalization ability of thought. Confabulation architectures allow arbitrary combinations of established knowledge of different types to be brought to bear on the outcome of a thought process, even if that combination has never been used before. For example, auditory and linguistic knowledge might be freely applied in the determination of the name of a particular aircraft heard flying, but unseen, nearby. A few seconds later, when the aircraft becomes visible, the same thought process can be applied again, now also using visual knowledge input.

The above characteristics possessed by confabulation architectures, the value of which is clearly illustrated by the ability of the specific architecture explored here to create the continuations shown in Fig. 6.2, are unprecedented.

The role of neuronal synchrony in confabulation theory is seen in knowledge links and confabulation (Chaps. 5 and 8). An alternative (which may play a role in the detailed neuronal implementation of some aspects of these functional components) is *polysynchrony* (Izhikevich 2006, 2007).

The excellent "grammar" and "syntax" seen in the experiments reported here suggests that these exist only as emergent properties of confabulation. If so, then the central conjectures of Miller, of Lenneberg, and of Chomsky [essentially that language reflects the universal human cognitive mechanism (Chomsky 1980)] are correct.

As with the properly coordinated and phased contractions of muscles during a successful movement, executing a successful thought process requires a properly coordinated and phased set of "contractions" of the involved modules to yield a confabulation consensus. Thinking is like moving, with modules functioning as the "muscles of thought."

In the experiments presented here, the thought process used causes the answer modules to converge in temporal order. However, thought process convergence can be made to occur in any desired order. This dynamic "convergence control" aspect of thought processes makes it meaningless to imagine a rigorous universal mathematical criterion for selecting confabulation consensuses. Each individual module converges to that symbol with the highest confabulation product (Chaps. 3 and 7), but the thought process, by controlling the order (and rate) of module convergence, and thereby the specific target symbol confabulation products considered, determines the final consensus. Thus, animal thought processes in which there are many viable confabulation consensuses are inherently non-deterministic (the slightest symbol input excitation perturbation, often at many points during the multiconfabulation, can switch the final consensus to one of a multitude of nearly-equally-good alternatives), endowing animal life with delightful inherent unpredictability.

Finally, note that the acquisition of a new thought process (here, sentence continuation with or without a context sentence) is primarily dependent upon having a supply of products (examples available for learning of pairwise symbol co-occurrences) of that thought process being carried out by existing competent practitioners (in this experiment, pairs of consecutive sentences produced by skilled human writers). This is the "*monkey-see/monkey-do*" skill acquisition principle (Chap. 8). Developing the associated confabulation consensus convergence maneuvers seems relatively easy, as the crude thought process used in these experiments illustrates. Thus, besides illustrating the universal mechanism of thought, these experiments may be providing us deep insight into the fundamental nature of animal thought process acquisition.

Author Contributions

Robert Hecht-Nielsen provided the concepts, principles, mathematical theory, experiment definition, learning procedure, the initial confabulation architecture and thought process designs; and wrote the chapter. Robert W. Means implemented and significantly improved these designs, and wrote the initial draft of Appendix 6.A. With the assistance of Syrus C. Nemat-Nasser, Luke Barrington, and Andrew Smith, Robert W. Means also carried out the learning procedure and further improved the confabulation architecture and thought process designs. Kate Mark produced the experimental examples and, once the architecture and thought process were finalized and frozen, ran the experiments.

Acknowledgments

Thanks to David Busby for help with development of some of the software used in implementing the confabulation architecture. This research was supported by Fair Isaac Corporation.

Appendix 6.A: Methods

Using Fig. 6.A.1 below, this appendix describes the sentence continuation thought process and remaining architecture details. The exposition starts with a more detailed description of the thought process; followed by a "third pass" discussion providing all remaining details.

For simplicity, Fig. 6.A.1 presumes that the first three phrase modules are used to represent the starter (i.e., re-representation of the starter at the phrase level did not involve any multi-word phrase symbols) – which was the most common situation in the experiments (the extension to other cases is straightforward). For experiments involving continuation with context, it is assumed that the context sentence has been represented with the context grouping and that the relevant symbols of its summary module have delivered excitation (via the knowledge links of the summary-module-to-summary-module knowledge base) to the symbols of the continuation grouping summary module. For continuations without context, the continuation grouping summary module is blank.

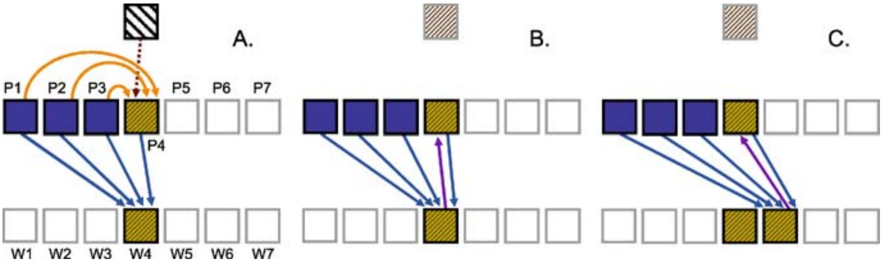


Fig. 6.A.1. Sentence continuation thought process sequence. Only the first seven word and phrase modules of the continuation grouping are shown. These are the only modules involved in steps of the P4 convergence thought process described here. Subsequent phrase and word modules are involved in convergence of P5, P6, and P7

Step A of the thought process (Fig. 6.A.1) involves first forming an expectation (a list of symbols receiving sufficient excitation) on P4, the fourth phrase module, by means of an initial light level of confabulation contraction (i.e., a small shortening of the module's expectation symbol list), during which knowledge links deliver excitation from the single symbols active on each of P1, P2, and P3 (which represent the three words of the starter). This P4 expectation is illustrated in Fig. 6.A.1 as a striped coloration fill. Next in Step A is knowledge link input to this P4 expectation from the (usually many) excited symbols of the summary module (if a context sentence is being used). This context input eliminates those expectation symbols which do not receive input from one or more of these links (this part of Step A is skipped if this is a continuation without context). Finally, an initial light contraction of module W4 (while knowledge links from the single symbols on P1, P2, and P3, and the excited symbols of the P4 expectation, are delivering excitation to W4's symbols), yields an initial expectation on W4 (again illustrated as a light color fill).

In Step B, the W4 expectation first sends excitation to P4 and the P4 confabulation is tightened further – typically reducing the number of symbols within the P4 expectation (i.e., narrowing the possible set of symbols that can be selected in this position). Then, the knowledge links from P1, P2, P3, and P4 are used to feed excitation to W4 as its confabulation is tightened a bit further. This "up and down" cycle is repeated until either a single symbol (a conclusion) emerges on P4, or the expectation on P4 stops changing. This ends Step B.

If Step B has yielded a conclusion, then P4 is locked and the steps, beginning with Step A, are repeated for the next unlocked phrase module (with excitation from P4 joining that from the summary module and the three initial phrase modules). If a conclusion is not reached (i.e., there are still multiple viable symbols on the P4 expectation after cycling), then Step C is invoked. Step C is like Step B, except that now W5 is fed by knowledge links from expectation symbols on P1, P2, P3 and P4 and lightly confabulated, yielding an expectation, which is then fed to P4, closing the loop. As in Step B, this "down and up" cycle is repeated until either a conclusion is reached on P4, or the P4 expectation stabilizes. If a conclusion is reached, P4 is locked and the steps, beginning with Step A, are repeated for the next unlocked phrase module. If a conclusion is not reached, then "Step D" (not shown in Fig. 6.A.1) is invoked. Step D is the same as Step C except that it features W6. If a conclusion is not reached, then "Step E" (not shown in Fig. 6.A.1) is invoked. Step E is the same as Step D except that it features W7 (the final word module of the continuation). If Step E still does not lead to convergence of P4 (a not uncommon occurrence), then all of the expectations of W4, W5, W6, and W7 are simultaneously used to excite P4, which is then strongly confabulated to yield a single conclusion. In principle, this final confabulation could lead to a null conclusion; in practice, however, this never happened during our experiments.

Once P4 is locked, the steps are repeated, in sequence, for each remaining unlocked phrase module. The final result is conclusions expressed on W4, W5, W6, and W7 – the confabulation consensus. Notice that in this thought process there is never any explicit consideration given to which words end up in the consensus – only that convergence to a consensus be achieved. There are no algorithms, rules, ontologies, productions, etc. involved in thought; only the thought-control signal manipulations required to achieve convergence to a confabulation consensus.

In cerebral cortex and thalamus, this "progressive tightening" thought process (the red circular arrow in Fig. 6.3) would presumably proceed smoothly and continuously in time and more in parallel, with each successive phrase module conclusion triggering the next step of the thought process.

Note that in the sentence continuation thought process, unless the choice of each successive phrase module conclusion is definitively decided by the candidate words considered so far, then additional candidate words (W module expectation symbols), from later in the sentence, are considered. This is a key thought process design principle, which ensures that the earlier words of the confabulation consensus agree with the later ones – even though the words are, in the end, selected in causal sequential order.

Words entered into the word modules of a grouping are automatically re-represented (via a simple thought process) (Chap. 7) in terms of phrases starting at the sentence beginning and progressing to the right. In effect, each phrase symbol selected is the longest applicable phrase available at that position. The phrase module used to represent a contiguous string of words being re-represented is always the one located (see Fig. 6.3) immediately above the module representing the first word of that string. The phrase modules intermediate between those having a symbol locked into them are locked with no symbol expressed. For example, if the P4 confabulation were to converge to the symbol representing **New York**, then P4 would be locked with this symbol expressed, and P5 would be locked with no symbol expressed [by a microbehavior launched by the conclusion symbol of **New York** – for details see Chap. 7]. Then P6 would be the next unlocked phrase module.

Once re-representation of a context sentence has been completed at the phrase level of a grouping (during entry of a sentence), the sentence summary module (responding to inputs from all of the active phrase modules) is lightly confabulated to yield an expectation including each symbol expressed on the phrase modules of that grouping. The relative excitation levels of these symbols reflect their multiplicity of appearance at the phrase level.

During learning (Hecht-Nielsen 2005, 2006), after the symbol co-occurrence counts were accumulated, $p(\psi|\lambda)$ was approximated by $c(\psi,\lambda)/c(\lambda)$, where $c(\psi,\lambda)$ is the count of the number of times the word represented by symbol ψ (belonging to the source module of the link within a designated knowledge base) and the word represented by symbol λ (belonging to the target module of a candidate link within a designated knowledge base) appeared together, and $c(\lambda)$ is the total number of times symbol λ appeared in the target module with any symbol of the source module ($c(\lambda)$ is equal to the sum of the $c(\phi,\lambda)$ across all symbols ϕ belonging to the source module). Candidate knowledge links with $c(\psi,\lambda) < 3$ or with $c(\psi,\lambda)/c(\lambda) < 0.0001$ were discarded as not meaningful. The only exception to this was the content summary module to content summary module knowledge base, produced during the second phase of leaning, where we discarded links having $c(\psi,\lambda) < 25$ or $c(\psi,\lambda)/c(\lambda) < 0.0001$.

Remaining details of the thought process are now described.

During each step of confabulation, the *input excitation* (Chaps. 3 and 7) $I(\lambda)$ of symbol λ (assumed, for concreteness, to be receiving knowledge links from four assumed fact symbols: α, β, γ, and δ) was always initially computed using the formula:

$$\begin{aligned} I(\lambda) \equiv {} & [\log_2 (p(\alpha|\lambda)/p_0) + B] \\ & + [\log_2 (p(\beta|\lambda)/p_0) + B] \\ & + [\log_2 (p(\gamma|\lambda)/p_0) + B] \\ & + [\log_2 (p(\delta|\lambda)/p_0) + B] \\ & = \log_2 [p(\alpha|\lambda) \cdot p(\beta|\lambda) \cdot p(\gamma|\lambda) \cdot p(\delta|\lambda)] \\ & - 4 \log_2 (p_0) + 4B, \end{aligned} \tag{6.A.1}$$

where $\log_2$ is the logarithm to the base 2, B is the *bandgap* (Chap. 7), and p_0 is the smallest meaningful antecedent support probability $p(\psi|\lambda)$ value (in these experiments we used $B = 30$ and $p_0 = 0.0001$). The formula for other numbers of link inputs to a symbol is the same, but with one term in the initial expression per knowledge link. In the case where a link's source symbol is not an assumed fact (previous confabulation conclusion or external input), but is part of an expectation, then the input excitation is the above term (e.g., $[\log_2 (p(\alpha|\lambda)/p_0) + B]$) multiplied by the excitation level of that source symbol (its fraction of the normalized total excitation – see below). Note that, as shown by the second equality above, for all target module symbols receiving the same number of knowledge links, the input excitation will vary directly and monotonically with the confabulation product. Thus, input excitation maximization can function as a surrogate for confabulation (maximization of the product $p(\alpha|\lambda) \cdot p(\beta|\lambda) \cdot p(\gamma|\lambda) \cdot p(\delta|\lambda)$). However, this mathematical formulation has many other benefits and is postulated to represent the neuronally implemented computational approach used in module confabulation (Hecht-Nielsen 2006). This additive knowledge combination law allows enabled knowledge links from active symbols to simply add their influence to that of others. This implies that symbols and knowledge links – as allowed by the genetically determined configuration of the brain – to freely combine available knowledge of all types (sensory, linguistic, abstract, planning, action, etc.). Thus knowledge links and symbols constitute a *universal internal language* of cognition, for which confabulation is the corresponding *universal information-processing operation.*

Once all of the symbols' initial input excitation levels are calculated, a "mild tightening" of the confabulation control signal is carried out. Effectively, this eliminates any symbol whose input excitation is more than B units below the input excitation of the most excited symbol [confabulation theory (Chap. 7) hypothesizes that a dynamic manipulation of module's thought-control signal can eliminate relatively-low-input-intensity symbols]. This yields the initial expectation list for this confabulation.

After the initial expectation list is formed, the final step of each individual confabulation tightening (in the event that the expectation contains any symbols) is to normalize the sum of the symbol excitations to 1.0 [this is hypothesized to occur in modules (Chap. 8)]. (Note: In these experiments there was one exception to this procedure; and that was the sentence meaning summary module of the continuation grouping. To simulate a much longer, sustained application of the knowledge link from the summary module of the context grouping, the expectation of this summary module, when it was used – namely, when a context sentence was present – was normalized to a sum of 8.0.) Thus, at the end of the confabulation tightening, we are left with an expectation list in which each symbol's excitation is set to its normalized value greater than or equal to 0.00 and less than or equal to 1.00 (a value of 1.00 implies that this is the only symbol in the expectation – i.e., this is the definitive conclusion). If the expectation is empty, then all symbols have excitations of 0.00. When a symbol in an existing expectation receives knowledge link inputs in a subsequent confabulation operation, its newly calculated input intensity is added to its pre-existing excitation before normalization.

When P4 (or the current leftmost unlocked phrase module) has still not been locked after Steps B, C, D, and E have been carried out, the knowledge links linking the most recent expectations of W4, W5, W6, and W7 are enabled to deliver excitation to the symbols of the latest P4 expectation. A strong confabulation contraction is then commanded; which selects that one symbol which is receiving the most input excitation (i.e., which has approximately the highest cogency). (Note: It is possible that no P4 symbol would be receiving any links and then this procedure would lead to a null conclusion. However, this never happened in our experiments.) This symbol is then used as the conclusion on P4. P4 is then locked with this symbol being expressed and, if the symbol represents a multi-word phrase with K words, the next (K-1) phrase modules are locked also, with no symbols expressed. The steps of the thought process are then applied again, starting with the first module following these locked modules. The process stops as soon as W4, W5, W6, and W7 become locked. These are then read out as the confabulation consensus.

Appendix 6.B: Complete List of Experiments and Results

The complete set of experimental examples that we investigated in this work, and the sentence continuations which resulted by applying the thought process, including the samples presented in Fig. 6.2, are provided below. These test starters / context sentences were formulated and frozen before the experiments began. We then carried out the thought process for each example. The same color scheme as in Figs. 6.1 and 6.2 is used. Changes of starter are indicated by lines of dots.

Because even more savvy enough investors usually …
In South Africa Mandela fought segregation and discrimination .
Because even more blacks weren't needed to …
A two-thirds majority won the vote .
Because even more Republicans vote Democratic Senate …
It was a technology training seminar for engineers
Because even more than enough jobs without …
The fed lowered interest rates again .
Because even more tight money market traders …
……………………………………

But the other semifinal match between fourth-seeded …
The state governor vetoed the bill .
But the other Republicans support abortion legislation …
The point guard made the free throws .
But the other game points out Mike …
The French foreign minister cried foul .
But the other EU leaders including France …
The school children study to improve math test scores .
But the other classes are students benefit …
Church priests counseled survivors of the earthquake .
But the other victims of the crash …
Our old machines were not powerful enough .
But the other programs have got us …
……………………………………

But the peace plan signed by former …
A neutral party helps to reach an agreement .
But the peace process fail to agree …
The volunteers helped refugees with basic needs .
But the peace forces began taking positions …
They were arrested for their beliefs .
But the peace activists gathered peacefully outside …
He was awarded for his lifetime achievement in science and technology .
But the peace prize awards ceremony Sunday …
The war in Gaza continues without ceasefire .
But the peace process still stalled further …
……………………………………

But the Roman Empire disintegrated during the ...
A two-thirds majority won the vote .
But the Roman Senate vote passed the ...
She learned the history of the saints .
But the Roman Catholic population aged 44 ...
Throughout history , wartime follows peace .
But the Roman period of stability particularly ...
She studied art history and classical architecture .
But the Roman Catholic church buildings dating ...
The athlete competed in several events .
But the Roman Empire split during games ...
....

For now this image has also uses ...
The space race stopped at the Moon .
For now this goal of an overwhelming ...
The Bulls also need a new guard .
For now this team isn't going to ...
The new model will add many features .
For now this technology provides users access ...
The coach was content with the season .
For now this game he's playing pretty ...
The returns were very low this year .
For now this money keeps getting better ...
It's a good time for businesses .
For now this year we're making more ...
Next year the Yankees will win .
For now this deal got four players ...
....

He liked to read them things better ...
Bill Clinton enjoyed his 2 terms in office .
He liked to stay around and keep ...
He brought us to his favorite restaurants .
He liked to eat anything he says ...
The guitarist was a fantastic musician .
He liked to dance around me in ...
Bill Clinton was a popular president .
He liked to play ; but kept ...
Patrick had a happy childhood .
He liked to draw people playing together ...
He sought the committee's advice .
He liked to show them that he ...
....

I was very nervous about my ability ...
The football quarterback fumbled the snap .
I was very upset with his team's ...
The state governor vetoed the bill .

I was very satisfied with his performance ...
The president addressed congress about taxes .
I was very difficult because Mrs. Clinton ...
Democratic citizens voted for their party's candidate .
I was very concerned that they chose ...
Restaurant diners ate meals that were served .
I was very hungry while knowing he ...
She quickly walked up the building stairs .
I was very nervous before she got ...
....

In spite of yesterday's agreement among analysts ...
The Mets were not expected to win .
In spite of the pitching performance of ...
The elections were deemed to be fair .
In spite of the dire results during ...
The President was certain to be reelected .
In spite of his statements toward the ...
The challenger was unlikely to win .
In spite of the fall season schedule ...
She had no clue about the answer .
In spite of her experience and her ...
Investor confidence was at an all-time low .
In spite of Japan's big business market ...
The price of oil fell further today .
In spite of the economic conditions facing ...
The president made a public apology .
In spite of political support in the ...
The mood was buoyant on Capitol Hill .
In spite of the government's budget appropriations ...
The mood was buoyant on Wall Street .
In spite of higher short-term rates lower ...
....

In the middle of the 5th century ...
Mike Piazza caught the foul ball .
In the middle of the season came ...
The frozen lake was still very dangerous .
In the middle of the lake is ...
Even the best students could not pay attention .
In the middle of one year term ...
The annual corn harvest was very abundant .
In the middle price range roughly $$ ...
....

It meant that customers could do away ...
The stock market had fallen consistently .
It meant that stocks could rebound later ...
I was not able to solve the problem .

It meant that we couldn't do much …
The company laid off half its staff .
It meant that if employees were through …
The president addressed congress about taxes .
It meant that we pay much more …
The salesman sold men's and women's shoes .
It meant that sales costs for increases …
She quickly walked up the building stairs .
It meant that she got me away …
Good quarterbacks run without fumbling the football .
It meant that offense against good team …
………………………………………

It must not be confused about what …
I can't make it work any better .
It must not be easy enough !
The news was a revelation to us .
It must not be true ... but …
The effects of alcohol can be dangerous .
It must not be used without supervision …
The subject was put to a vote .
It must not be required legislation to …
The opposition party has grown even stronger .
It must not be strong political opposition …
There's no way to solve the problem .
It must not be easy problem solution …
The Padres have not won a game yet .
It must not be better players thinking …
………………………………………

It was a gutsy performance by John …
The tennis player served for the match .
It was a match played on grass …
The state governor vetoed the bill .
It was a Republican bill in Congress …
The bank robber stole the money .
It was a dlrs 500 million project …
Desperate refugees arrived daily by the thousands .
It was a small city under Israeli …
Coastal homes were damaged by tropical storms .
It was a huge relief effort since …
The ship's sails swayed slowly in the breeze .
It was a long ride from the …
………………………………………

It was about twelve hundred men !
The football quarterback fumbled the snap .
It was about 20 yards this season …
The president addressed congress about taxes .

It was about 25 % 30 15 ...
The frozen lake was still very dangerous .
It was about 20 feet above flood ...
Stocks proved to be a wise investment .
It was about 20 percent cheaper than ...
The basketball player made the free throw .
It was about 10 players out there ...
International organizations are created nearly every day .
It was about 40 groups of people ...
....

She thought that would throw us away ...
The tennis player served for the match .
She thought that she played a good ...
Desperate refugees arrived daily by the thousands .
She thought that they live under close ...
At the library she studied many books .
She thought that I learned her words ...
Children played at the park all afternoon .
She thought that would put her kids ...
....

Shortly thereafter , she began singing lessons ...
The football quarterback fumbled the snap .
Shortly thereafter , Lewis got him to ...
The president addressed congress about taxes .
Shortly thereafter , Gore aides noted strong ...
The baseball pitcher threw at the batter .
Shortly thereafter , the Mets in Game ...
Democratic citizens voted for their party's candidate .
Shortly thereafter , Gore was elected vice ...
....

The following day she began her again ...
The football quarterback fumbled the snap .
The following day he took his first ...
The point guard made the free throws .
The following day was given another shot ...
The president addressed congress about taxes .
The following day he called for repeated ...
They hoped to improve election results .
The following day President and government released ...
....

The New York Times' computer model collapses ...
Stocks proved to be a wise investment .
The New York markets traded lower yesterday ...
Downtown events were interfering with local traffic .
The New York City Center area where ...

Library books were sold for almost nothing .
The New York library opens into three ...
The salesman sold men's and women's shoes .
The New York fashion show ran off ...
Coastal homes were damaged by tropical storms .
The New York City Emergency Service System ...
Church priests counseled survivors of the earthquake .
The New York Community Services Council voted ...
Metals such as silver and copper are mined .
The New York metals futures exchanges closed ...
Medical patients tried to see their doctors .
The New York University Medical Association reported ...
....

The president said he personally met French ...
Regarding the alleged abuse of Iraqi prisoners .
The president said he had met the ...
Free trade benefits the entire nation .
The president said that Pakistan would continue ...
The constitution guarantees freedom of religion .
The president said he had been assured ...
The flat tax is an interesting proposal .
The president said he promised Congress to ...
The commission has reported its findings .
The president said he appointed former Secretary ...
The court ruled yesterday on conflict of interest .
The president said he rejected the allegations ...
....

The San Francisco Redevelopment Authority officials announced ...
What makes fish and dolphins jump in the air ?
The San Francisco Water Department issued recently ...
Their star player caught the football and ran !
The San Francisco quarterback Joe Brown took ...
The pitcher threw a strike and won the game .
The San Francisco fans hurled the first ...
I listen to blues and classical music .
The San Francisco band draws praise from ...
Area citizens complained about pollution in local water .
The San Francisco residents witnessed the closest ...
Many survivors of the catastrophe were injured .
The San Francisco Police officials announced Tuesday ...
Annual crops of wheat and corn were subsidized .
The San Francisco biotechnology company ; cost ...
The wheat crops were genetically modified .
The San Francisco food sales rose 7.3 ...
....

There were several unconfirmed reports that Iraqi …
The football quarterback fumbled the snap .
There were several moments when he felt …
The bank robber stole the money .
There were several hundred dollars chasing fewer …
Stocks proved to be a wise investment .
There were several risks are investors regarding …
The state governor vetoed the bill .
There were several senators joined Republicans Tuesday …
....

This meant that instead of just how …
New Zealand won the America's Cup .
This meant that they lost face against …
The LA Dodgers defeated the Red Sox .
This meant that Park stadium where Democrats …
I got a large tax refund .
This meant that they get little money …
France opposed the US in front of the UN Security Council .
This meant that some government should not …
....

This resulted in a substantial performance increase …
The state governor vetoed the bill .
This resulted in both the state tax …
Oil prices rose on news of increased hostilities .
This resulted in cash payments of $ …
The United States veto blocked the security council resolution .
This resulted in both Britain and France …
....

Three or four persons who have killed …
The president addressed congress about taxes .
Three or four years because Clinton would …
The tennis player served for the match .
Three or four times in a row …
Stocks proved to be a wise investment .
Three or four years since Clinton opened …
....

We could see them again if we …
The point guard made the free throws .
We could see the ball forward toward …
The president addressed congress about taxes .
We could see additional spending money bills …
The number of car thefts was especially low .
We could see an increase in demand …
The view in Zion National Park was breathtaking .
We could see snow conditions for further …

We read the children's books out loud .
We could see the children who think …
The U.N. Security Council argued about sanctions .
We could see a decision must soon …
…………………………………………

What will occur during the darkest days …
Research is leading to new discoveries .
What will occur during training program testing …
Research scientists have made astounding breakthroughs .
What will occur within the industry itself …
The vacation should be very exciting .
What will occur during Christmas season when …
I would like to go skiing .
What will occur during my winter vacation …
There's no way to be certain .
What will occur if we do nothing …
…………………………………………

When the Union Bank launched another 100 …
The rebel leader fought until death .
When the Union flag was raised nearly …
He lived during the Civil War era .
When the Union Jack dips for the …
She loved her brother's Southern hospitality .
When the Union flag was raised again …
New York City theater is on Broadway .
When the Union Square Theater in Manhattan …
The railroad reached the Pacific ocean .
When the Union Pacific A fell 75 …
…………………………………………

When the United Center Party leader urged …
The airplane wreck worried anxious travelers .
When the United Airlines flight crashed off …
The car assembly lines halted due to labor strikes .
When the United Auto Workers union representation …
The price of oil in the Middle East escalated yesterday .
When the United Arab Emirates bought the …
The dictator's reign of torture and terror ended in trial .
When the United Nations opened in mid-May …
The American people felt quite patriotic .
When the United States' best friends ever …
…………………………………………

7 Mechanization of Confabulation[8]

7.1 Introduction

This chapter describes the state of the art in creating animal cognition in machines. It begins with a discussion of the two fundamental processes of cognitive knowledge acquisition – *training* and *education*. The subsequent sections then present some ideas for building key components of cognition (language, sound, and vision). The main point of this chapter is to illustrate how we can now proceed towards the mechanization of key elements of cognition. This chapter assumes that the reader is familiar with the concepts, terminology, and mathematics of elementary *confabulation* (as described in Chaps. 1–6 and the DVD video presentation) and its hypothesized biological implementation in the human cerebral cortex and thalamus (as described in Chaps. 3, 5, and 8).

7.1.1 Mechanized Cognition: The Most Important Piece of AI

As discussed in Chaps. 3, 5, and 8, human (and higher mammal) intelligence involves a number of strongly interacting, but functionally distinct, brain structures. Of these, significant progress has now been made on three: cerebral cortex and thalamus (the engine of cognition – and the focus of this chapter), basal ganglia (the *behavioral manager* of the brain – which manages action evaluation, action selection, and skill learning), and cerebellum (the *autopilot* of the brain – which implements detailed control of routine movement and thought processes with little or no need for ongoing cognitive involvement once a process has been launched and until it needs to be terminated). There are a number of other, smaller-scale, brain functions that are also critical for intelligence (e.g., ongoing drive and goal state determination by the limbic system), but these will not be discussed here.

Of all of the components of intelligence, cognition is, by far, the most important. It is also the one that has, until now, completely resisted explanation. This chapter provides the first sketch of how cognition can be mechanized. The approach is based upon the author's theory of vertebrate cognition described in Chaps. 3, 5, and 8. This chapter is not an historical description of "how cognition

[8] This chapter is based on the original publication Hecht-Nielsen R (2006) The mechanization of cognition. In: Bar-Cohen Y (ed) Biomimetics. CRC Press, Boca Raton, FL, pp 57–128, adapted here from the original with kind permission of the publisher.

was mechanized," rather, it is an "initial plan for mechanizing cognition." Initial progress in implementing this plan in areas such as language and hearing (the subjects of Sects. 7.3 and 7.4) has been encouraging.

7.1.2 Module Capabilities

This chapter considers some more sophisticated variants of confabulation that go beyond elementary confabulation. Each module used in our (technological) *confabulation architectures* (collections of modules and knowledge bases) will be assumed to possess the machinery for carrying out each of these confabulation variants (or information-processing *effects* – the term that will be used for them here), as described below. Thus, from now on, the term *module* implies a capability for implementing a finite set of symbols, maintaining a list of the excitation states of those symbols, and for executing the effects defined below. For the moment, module dynamics will be ignored. (However, in later sections, concepts such as *multiconfabulation* and *symbol interpolation*, which intrinsically require module dynamical behavior, will be briefly mentioned.)

One very important detail is that, while confabulation is underway, a module can have multiple highly excited symbols (as opposed to just one active symbol, or the null symbol). When this occurs, the neurons representing these symbols will be excited at various (high) levels for different symbols.

When such multiple highly excited but not active symbols are used as "assumed facts" transmitting through a knowledge base, their effects on symbols to which they link via this knowledge base will essentially be the product of their excitation and the link strength. In practice, such multi-symbol "assumed facts" are very important, as they are the key ingredient in *multiconfabulation*, which is also characterized as *consensus building*. This involves dynamically interacting confabulations taking place contemporaneously in multiple modules, which is the dominant mode of use of confabulation in human cognition. However, to keep this chapter at an elementary and introductory level, the mathematics of multiple-symbol "assumed fact sets" will not be discussed in detail. As needed, the qualitative properties of this mode of confabulation will be discussed, which will be sufficient for this introduction.

For the technological purposes of this chapter, confabulation will be taken to be dependent upon the *input excitation* sum $I(\lambda)$ of symbol λ, which is re-defined (from Chap. 3) to be:

$$\begin{aligned} I(\lambda) &\equiv [\ln(p(\alpha|\lambda)/p_0) + B] + [\ln(p(\beta|\lambda)/p_0) + B] \\ &\quad + [\ln(p(\gamma|\lambda)/p_0) + B] + [\ln\,(p(\delta|\lambda)/p_0) + B] \\ &= \ln[p(\alpha|\lambda) \cdot p(\beta|\lambda) \cdot p(\gamma|\lambda) \cdot p(\delta|\lambda)] - 4\ln(p_0) + 4B, \end{aligned} \tag{7.1}$$

where ln is the natural logarithm function, B is a positive global constant called the *bandgap* (a term coined by my colleague Robert W. Means), and p_0 is the smallest meaningful $p(\psi|\lambda)$ value. Clearly, for a symbol λ receiving N knowledge links, the value of $I(\lambda)$ ranges over the numerical *interval* from NB to $N[\ln(1/p_0) + B]$.

It will be assumed that the constant B is selected such that for $N = 1, 2, \ldots, N_{max}$ none of these intervals ever overlap. For example, if we take $p_0 = 0.0005$ and $N_{max} = 10$, then we can select $B = 100$. The intervals upon which $I(\lambda)$ can lie are then given by [100,107.6], [200,215.2], … , [1000,1076.0] for $N = 1, 2, \ldots, N_{max}$, respectively. The utility of this definition will be seen immediately below. Given these preliminaries, we can now discuss variants of confabulation.

The first effect considered is *erasing*, denoted by **E**. Erasing clears the current record of excitation states of the module and prepares the module for a new use. For example, before a module is used as the answer module of a confabulation operation it must be erased.

Elementary confabulation (as described in Chaps. 3 and 4), denoted by **W**, is carried out by *activating* a single symbol ε with the highest value of $I(\lambda)$ (ties are broken randomly). By activation it is meant that the *final excitation level* $I(\varepsilon)$ of that symbol is set to 1 and the final excitation levels of all other symbols are set to zero. There is also the effect **W**K, which is the same as **W** with the added requirement that the single winning symbol, if there is to be one, must have had at least K knowledge link inputs (i.e., the winning symbol must have its input intensity in the K^{th}, or higher, $I(\lambda)$ interval). The primary form of confabulation discussed in Chaps. 3 and 4 was **W**4.

The effect **C**K (*confabulation conclusions having* K *or more knowledge link inputs*), which will be needed for discussions, first zeros the excitation sum $I(\theta)$ of each symbol θ whose $I(\theta)$ is not in the K^{th} (or higher) $I(\lambda)$ interval(s) occupied by symbols of the module. The excitation levels of all the symbols are then summed. Finally, each remaining non-zero symbol excitation is then divided by this sum to yield its *final excitation level*. For example, in the above example with $p_0 = 0.0005$ and $N_{max} = 10$, and $B = 100$, if the four highest symbol input excitation sums are 346.8, 304.9, 225.0, and 146.8, then a **C**2 will yield only the top three symbols, with final excitation levels of 0.395, 0.3478, and 0.2566, respectively. Clearly, this effect yields the set of confabulation conclusion symbols that had K or more knowledge link inputs. Normalizing the sum of the "significant symbol's" excitation levels to 1.0 corresponds to the notions of "activation" and "high excitation" in cortex. Cognition must very often conduct a multi-stage process of gradually promoting hypotheses (expectation symbols) which gain significant support from incoming knowledge links and demoting those which fail to gain as much support. Thus, the effect **C**K is very important in cognition. In the brain, "**C**K processing" is continuous in time and happens very rapidly. In mechanizing cognition **C**K is important because in practical implementations time moves forward in discrete steps. **C**K is a transformation taking a module's excitation state from one time step to the next.

The set of symbols of a module having non-zero excitation levels $I(\lambda)$ following a **W**, **W**K, or **C**K effect is termed an *expectation*. Expectations are considered to have a short life (i.e., after a "short" time has elapsed after a confabulation the *state* of the module, its collection of $I(\lambda)$ values, becomes indeterminate). Note that this is a refinement of the term *expectation* used in Chap. 4.

The term *active* is still reserved for the case of a single confabulation conclusion; and *highly excited* will still mean that the expectation has multiple elements.

Another effect is *freezing*, denoted by the letter **F**. Freezing a module causes each symbol with positive final excitation (i.e., after a **W**, **WK**, or **CK**) to have its final excitation $I(\lambda)$ value preserved for a longer time. During this (still rather brief) period of time that follows **F**, only those symbols which are members of this expectation can receive further knowledge link inputs. In other words, the input excitations of symbols not in the expectation stay at zero during the frozen period. So, for example, if further new link inputs arrive shortly after an **F** has been invoked, and then a **W** is commanded, an expectation symbol (if there are any) which obtains the highest positive $I(\lambda)$ value will be made active.

As we will see later, building and using expectations is one of the most important elements of cognitive information-processing. By using sequences of confabulations to "whittle down" expectations, *constraint knowledge* of various kinds can be applied to rapidly home in on a final conclusion. In effect, each expectation represents the set of all "reasonable conclusions" that are worth considering further. When the expectation is finally reduced to one conclusion, via successive freezes and confabulations, the final, *decisive*, conclusion is found (or, if the final expectation is empty, then the answer is "I don't know"). Almost every aspect of cognition is implemented by such sequences of such "deductive" confabulation steps [although this is not deduction in any formal sense because it based on the undecidable (but usually reasonable) assumption of exhaustive knowledge].

Finally, consider a module which, when last used for confabulation (within the past few hours), yielded a decisive conclusion and which, subsequently, has not been erased. If this module now receives a **W**, **WK**, or **CK** but no knowledge link inputs, that symbol which was its last conclusion will, in isolation, become active. This is a sort of temporary symbol storage mechanism that the theory terms *working memory*.

7.1.3 Discussion

Technological cognition will be inherently limited without the other functionalities that brains provide (see Sect. 8.1). Further limitations arise because of the lack of on-line memory formation mechanisms (short-term, medium-term, and long-term memory processes) and the lack of a capability for goal-driven delayed reinforcement learning of thought and movement procedures (see Sect. 7.6). Yet, despite these limitations, there are probably many high-value early applications of *pure cognition* (i.e., cognition using contrived thought processes, as discussed in the DVD video presentation) that will be possible. Pure cognition is the focus of this chapter.

Language is almost surely the faculty which accounts for the dramatic increase in human mental capability in comparison with all other animals. It is in the language faculty, and in the language faculty's interfaces with the other cognitive faculties, that almost all distinctly human knowledge is centered. Thus, language is where the mechanization of cognition must start (see Sect. 7.3). However, before discussing language cognition, the next section discusses the currently available general methods of antecedent support knowledge acquisition: *training* and *education*.

7.2 Training and Education

As discussed above, current confabulation technology is limited to development of knowledge using some externally guided process; not via dynamic, autonomous goal and drive satisfaction-driven memory formation, as in brains. This section discusses the two main processes currently used in knowledge development: training and education. When dynamic memory formation eventually arrives, training and education will still be important learning processes (but no longer the only ones).

7.2.1 Training

Training is a knowledge acquisition process that is carried out in a batch mode without any significant active supervision or conditional intervention. It is a learning mode that can only be applied when the data set to be used has been carefully cleaned and prepared. For example, in learning proper English language structure it is possible to take a huge (multi-gigaword) proper text corpus and train knowledge bases between modules representing the words in English [Chap. 4]. The corpus used must be near-perfect. It must be purged of words, punctuation, and characters that are not within the selected word list and must not have any strange annotation text, embedded tables, meaningless text fragments, highly redundant content, or markup language fragments that will be inadvertently used for learning. Achieving this level of *cleanliness* in a huge training corpus which, necessarily – for diversity – is drawn from many sources, is expensive and time consuming.

Once a suitably clean text corpus has been created, each sentence is considered as a whole item (up to a chosen maximum allowed number of words – e.g., 20, as in Chap. 6 – after which the sentence is simply truncated). The confabulation architecture to be trained has as many word modules (in a linear sequence) as the maximum number of allowed words in a sentence). The words of the sentence are represented by active symbols on the corresponding word modules of the architecture (see Sect. 7.3 and Chap. 6 for more details). Co-occurrence counts are then recorded for each *causal* pair of symbols (i.e., between each symbol and each of the symbols on modules further down the sequence of modules). Once these counts are recorded, the process moves on to the next sentence of the training corpus.

A beautiful thing about training is that the result is knowledge that presumably has the same origin and legal standing as knowledge obtained from material that a human person has read; but which they do not remember in detail.

Namely, this knowledge is presumably not subject to source copyright restrictions or other source intellectual property restrictions. Use of raw data for training probably falls under the category of "fair use," which eliminates any need to pay royalties. Confabulation-based systems may thus be able to absorb whole libraries of knowledge without cost. This is fair use because the content of the work is not stored and cannot be recalled. (How much does your library charge you in royalties for reading a book? Answer: Absolutely nothing, because reading a library book is fair use.) This fortuitous loophole may allow cognitive machines to rapidly and efficiently accumulate almost all human knowledge, without having to pay any royalties and without the delays associated with working through legal and bureaucratic objections. Mechanizers of cognition may want to expose their systems to the available libraries of written knowledge at the first possible opportunity before legal innovators find ways of closing this loophole. It may not be long before intelligent machines are as unwelcome at libraries as blackjack card counters are at casinos.

In the short term, early confabulation entrepreneurs will probably use libraries, web scrapers, or informally obtained e-mail message examples (for text knowledge), informal public volunteer web portals for conversational data (for sound knowledge), public location video (for vision knowledge), and multi-camera video of moving humans with colored dots pasted to their bodies (for motor knowledge). Paying for training data will probably not be feasible for most confabulation startup companies.

The above comments also raise the technical legal question of whether the knowledge in confabulation-based systems can itself be copyrighted (this would seem reasonable); or must it be protected as a trade secret? Methods of training and education may be patentable. The legal implications and ramifications of confabulation are clearly going to be complicated, and probably contentious. An overriding consideration should be the irreplaceable value of the work output that intelligent machines (which can potentially produce prodigiously, but consume insignificantly) will quickly add to the world economic product. It will be fun to watch this saga unfold in the courts and in diplomacy over the coming decades.

Knowledge created by training is limited to situations such as that considered above; namely, where extensive, highly conditioned and prepared data sets exist. In more general situations, online, active, expert human supervision must be employed to carefully select meaningful symbol co-occurrences for use in learning. Such a carefully sequenced program of sophisticated and controlled exposure of the machine to meaningful examples is termed *education*, which is the subject of the next sub-section.

7.2.2 Education

A critical aspect of development, particularly in higher mammals, is the limited, deliberately controlled exposure to progressively more complicated stimuli that characterizes the early phases of an animal's life (in cats, this might occupy a few weeks, whereas in humans it occupies tens of years – which is often not enough!). During this development period, the sequence of exposure of the animal to information is in some manner controlled (often by confining the animal to a particular limited range, such as a nest, home, or school and its immediate surround).

For example, a human baby learning to see has eyes that are physically incapable of focusing much beyond its reach. Thus, most visual stimuli are the baby's own limbs or individual objects that the baby itself is holding and manipulating. During this period, the visual system develops its ability to segment individual objects in single views and also develops higher-level visual modules containing symbols that are pose-insensitive (see Sect. 7.5). Knowledge related to the integration of form, color, texture, and internal object motion is also developed during this initial phase. In order for this phase to properly complete, the baby must have spent a large amount of time holding and viewing a reasonably rich collection of objects.

Once the initial phase of human visual development is completed, the baby begins to acquire distant vision and begins to learn about a much richer visual environment. Again, parental provision of appropriate stimuli during this period is critical. Persons who are deprived of visual input during these early phases (e.g., due to disease that temporarily impairs visual function) are often never able to fully complete their visual development, even if their visual input is restored at some later point. Such persons can respond to light in some limited ways, but cannot see as well as normals. Some persons with restored sight actually voluntarily limit their exposure to visual input (Gregory 2004).

As with the initial stage of visual development, the most important source of educational input in the later stages of visual development is the child itself. By holding an object and examining it (e.g., in an exploration of its function or component parts), knowledge in the visual domain, as well as in the linkage of vision to the language (and other) faculties, is expanded. Unlike intellectual information (which is sometimes fallacious), visual knowledge is "safe" to rapidly gather and store because it is essentially never erroneous (except in cases where optical distortions exist – which, when corrected too late in the development process, often cause a permanent reduction in visual capability). Parents often endlessly admonish their children not to handle everything they fancy in stores, yet this is probably enriching. Perhaps the admonishment should be to take care not to soil or damage what they handle. If you are punctilious in this regard, have your children wear clean disposable latex gloves and force them to pay for any damage or breakage out of their allowance. But give them these valuable experiences.

Even more than basic sensory processing, learning to carry out important behavioral tasks requires deliberate provision of examples, and supervised rehearsal practice. This often includes feedback on performance, something that will be ignored here since using such feedback requires non-cognitive functions which, as yet, we do not understand sufficiently to build. Because of this current lack of a reinforcement learning adjunct to cognition [work is proceeding in this area – see Miyamoto et al. (2004), for example], for the moment, education of confabulation-based cognitive systems will probably be confined to strictly positive examples. In other words, examples where learning should definitely take place.

For example, consider a confabulation-based vision system viewing cars passing by on a busy road. The visual portion of the system segments each car it fixates on (see Sect. 7.5) and then re-represents its visual form, color, and internal motion using high-level symbols that have invariance properties (e.g., pose and illumination insensitivity). Thus, the final product of processing one such *look* is activation of a set of high-level symbols, each describing one visual *attribute* of the object.

Imagine that a human educator sitting at a computer screen where each *look* (eyeball snapshot image – see Sect. 7.5) to be processed by the confabulation-based vision system is being displayed (each subsequent look is processed only after the previous look's use for education has been completed). The human examines the visual object upon which the center (fixation point) of the eyeball image rests and describes it in terms of English phrases (spoken into a noise-canceling microphone connected to an accurate speech transcriber – see Sect. 7.4). For example, if the object is a green Toyota Tundra truck with a double cab, the educator might speak: "Toyota Tundra truck," "dark green," "two rows of seats; in other words, a full-sized back seat," "driving in the left lane of traffic." After accurate transcription, this text is represented by a set of active symbols in a language module (see Sect. 7.3). Knowledge links are then established between the active visual symbols representing the visual content of the look and the active language symbols representing the education-supplied language content of the look.

Note that the language description is not exhaustive; it is just a sample of descriptive terms for the visual object. For example, if a similar look of the same truck were presented on another occasion, the educator might add: "oh, and there are four dogs in the bed of the truck." This would add further links.

7.2.3 Discussion

One of the most exasperating things about confabulation theory is that it seems impossible that just forming links between co-occurring symbols and then using these links to approximately maximize cogency could ever yield anything resembling human cognition. The theory appears to be nothing but a giant mountain of wishful thinking! That such a simple construction can do all of cognition is indeed astounding. Yet that is precisely the claim of confabulation theory. Some reasons why confabulation may well be able to completely explain cognition are now discussed.

First is the fact that the number of links that get established (i.e., the number of individual items of knowledge that are employed) is enormous. Even in the narrow domain of single proper English text sentences (see Sect. 7.3 and Chap. 6), over a billion individual knowledge items are often employed (contrast this with the world's largest rule bases, which have about 2 million items of knowledge). Slightly more elaborate proper English text confabulation systems (able to deal with two successive sentences – see Sect. 7.3) often possess multiple billions of items of knowledge.

The value of having such huge quantities of such a simple form of knowledge is best seen in terms of how this knowledge is used in confabulation. The first use of knowledge is to excite those conclusion symbols which strongly support the truth of the set of assumed facts being considered.

However, an even more important aspect of cognition is the underlying assumption that the knowledge we possess is *exhaustive* (see Chap. 4). In other words, once the available knowledge has been used, we can be reasonably sure that no other viable conclusions have been excluded (dropped from the expectations involved). This effectively causes the known knowledge to act as an implied constraint. In particular, those possible conclusions identified as known to be supportive of the assumed facts are not just viable alternatives; they are probably the only viable alternatives. Thus, in non-Aristotelian information environments the *exhaustive knowledge assumption* leads to answers which are "almost logical deductions." Thus, confabulation can be viewed as a "strong" (although not logically rigorous) form of inductive reasoning.

Another factor that makes confabulation so powerful is its ability to support the construction and use of module hierarchies. In the simplest case, the symbols of a higher-level module each represent an ordered set of symbols that meaningfully co-occur on lower-level modules. But much more is possible. For example, in vision (see Sect. 7.5), individual higher-level symbols each represent several groups of lower-level symbols. These are symbol groups that are seen in successive "eyeball snapshots" of the same object at the same fixation point (but at slightly varied object poses). In this way, these higher-level symbols respond to the appearance of a localized portion of an object at a number of different poses. They are *pose-insensitive localized visual appearance descriptors*.

In language hierarchies, knowledge can be used to discern symbols which are highly similar in meaning and usage (in a particular given context) to a particular symbol. These are termed *semantically replaceable elements* (SREs). Knowledge possessed about an SRE) of a symbol can sometimes be used to augment knowledge possessed about that base symbol. This can significantly extend the "conceptual reach" of a system without requiring training material covering all possible combinations of all symbols. For example, what if a friend tells you about the food "guyap" that they had for breakfast. They poured the flakes of guyap from its cardboard box into a bowl, added milk and sweetener, and then ate it with a spoon. It was good. By now, you are fairly sure that "guyap" is a breakfast cereal of some kind and, at least in the "breakfast food" context, you can apply your knowledge about breakfast cereal to "guyap."

Hierarchies can work backwards too. For example, if you say you are looking for a ruler on your desk, then links from the word `ruler` (and perhaps some of its SREs) go to the visual system and provide input to high-level visual attribute ("holistic") representation symbols which, in the past, have meaningfully co-occurred with visual sightings of rulers. This is accomplished via a **CKF** effect, which leaves an expectation of all such symbols that have previously been significantly linked to the word `ruler`. During *perception*, which takes place immediately after this expectation symbol set has been generated, knowledge links from primary, and then secondary, visual modules arrive at this high-level visual module and a **W**, **CK**, or **WN** is issued at the same time. Only elements of the expectation can be activated by the visual input, and the net result is a set of symbols that are consistent with both the word `ruler` and with the current visual input. As discussed further in Sect. 7.5, the final step is a rapid bidirectional knowledge link interaction of the higher-level expected symbols with those of the lower-level modules to shut off any symbols that are not participating in "feeding" (i.e., are not consistent with) symbols of the high-level expectation. This, in effect, causes low-level symbols representing portions of the visual input that are not part of the ruler to be shut off. It is by this *visual object segmentation* mechanism that sensory objects are almost instantly isolated so that they can be analyzed without interference from surrounding objects (segmentation is also a key part of sound and somatosensory processing). Without an expectation, sensory processing cannot proceed. [This point, which seems to be to be widely unappreciated in the biologically-oriented neuroscience disciplines, is the subject of a wonderful book describing clever experiments that well illustrate this point (Mack and Rock 1998).]

Above all else, cognition works because of the huge hierarchical repertoire of learned and stored *action sequences* (programs of thought and/or movement – see Sect. 7.6). Appropriate actions, or action sequences, are triggered instantly each time a module confabulation operation yields a single active symbol (i.e., a *decisive conclusion*). As discussed in Chap. 8, this is the *conclusion → action principle* of the theory. The action(s) automatically triggered by the winning symbol can be of many characters. They can be immediate *postural goal* outputs that are sent down the spinal cord to motor nuclei and the cerebellum to launch a movement process, they can be immediate *module operation* commands, they can be immediate *knowledge base enablement* commands, or they can be *candidate actions* (cortically proposed thoughts and movements) which must first be sent to the basal ganglia for evaluation and approval before they are executed.

The main advantage of confabulation-based cognition over traditional *programmed computing* (formal computer programs, rule-based systems, etc.) is a much greater capacity for handling novel arrangements of individually familiar objects. Programmed computing must essentially have a pre-defined plan for dealing with every situation that is to be handled. For example, a plan for breaking up a complicated ensemble of problems into isolated, disconnected problems, so that each can be handled in a pre-defined way. Unfortunately, in most real-world situations this approach fails badly because complicated real-world situations inevitably have unanticipatable interrelations between their elements. By virtue of their huge stores of general-purpose, low-level knowledge, confabulation-based systems are inherently able to take novel external context into account as each individual conclusion (or ensemble of conclusions, if mutual solution constraints are to be honored – see the discussion of *multiconfabulation* in Sect. 7.3 below) is addressed. Confabulation-based systems can also adapt existing action plans (e.g., by replacing specific elements of a stored plan with similar substitutions which are relevant to the current situation) to fit novel circumstances. They do not typically run out of things to try and, instead, tend to press on and do the best they can, given what they know. If a particular approach yields no conclusion, other approaches are typically immediately launched. Yet, because actions are triggered each time a conclusion is reached, almost all behavioral sequences are dramatically novel. Also, each new experience having a positive outcome (as judged by the human educator) be added to the knowledge base to further enlarge the system's future repertoire.

More could be said regarding the benefits of the confabulation approach. However, the remaining sections of this chapter present more concrete examples of this. The nature of cognition is very different from that of computing. So much depends upon designing clever architectures of modules and knowledge bases and upon using clever, highly threaded, but very simple thought processes to control these architectures. Since the information-processing control which must be exerted at each stage of an action process is triggered by the current *cognitive world state* (the collection of all decisive confabulation conclusions that are active at that moment), cognition has no need for "computer programs" or "software." In effect, the conclusion of each "cognitive microaction" (lowest-level action sequence) is a GOTO statement. There is no overall program flow defined, just action sequences that complete and then trigger subsequent action sequences (in a pattern that almost never exactly repeats). Things happen as they happen, with no master program controller involved or needed (although a number of sub-cortical brain structures can execute "interrupts" when certain conditions occur). This brain operating system (or lack thereof, depending on your point of view) seems like an invitation to disaster. However, beyond possible conflicts between commands to the same action (movement or thought) resource (which are impossible by design! – see Sect. 8.2), very little can go wrong.

7.3 Language Cognition

This section discusses the use of confabulation for representing and generating language. This application arena is the most developed, and yet is transparently crude and primitive. An enormous amount of work needs to be done in language. The hope of this section is to illustrate how promising this research direction is.

7.3.1 Phrase Completion and Sentence Continuation

This discussion of language cognition begins with consideration of a class of *confabulation architectures* for dealing with single English sentences. These architectures address the problems of *phrase completion* and *sentence continuation*;

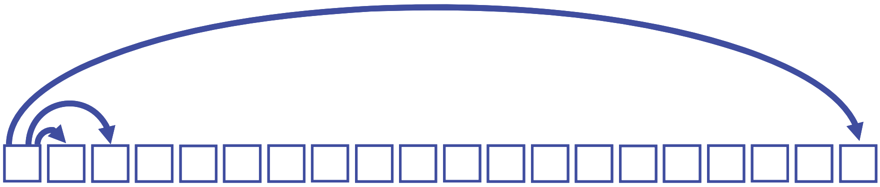


Fig. 7.1. Naïve single-sentence confabulation architecture for proper English phrase completion or sentence continuation. Knowledge bases link each of the first 19 of the 20 modules to all of the modules to their right. Sentences are represented with the first word in the first module on the left, and so on in sequence. This architecture has a total of 19 + 18 + ... + 1 = 190 knowledge bases

simple sub-cases of *language generation*. This sub-section expands upon on the brief introductions provided in Chaps. 4 and 6. These architectures provide a good introduction to the "look and feel" of cognitive information-processing – which is completely different from the familiar computer paradigm.

Figure 7.1 illustrates a confabulation architecture for phrase completion and sentence continuation in a single sentence of up to 20 words. Each module has about 63,000 symbols, including symbols for the 63,000 most common words in English (as reflected in the training corpus) and eight punctuations (period, comma, semicolon, etc.), which are treated as separate words. Capital letters are used when they appear in words in the training corpus selected for representation within the word modules (i.e., `mark` and `Mark` are different words, with different symbols). Thus, many of the words in the module are represented twice, once capitalized and once not; some have even more than two representations, e.g., `EXIT`, `Exit`, and `exit`; and some, such as `e.g.` and the punctuations, are never capitalized and only have one representation.

Once a suitably "clean" huge proper English text training corpus (typically containing billions of words) has been created, each successive sentence in the corpus is *entered*, in sequence, into the architecture of Fig. 7.1. The first word of the sentence is entered into the leftmost module (i.e., the symbol representing this word is made active) and the remaining words of the sentence (or punctuations – which, again, are treated as separate words) are entered successively until the ending period. If the sentence has more than 20 words, those words beyond the first 20 are discarded. Because of the positioning of the words of each sentence in order, this architecture is termed *position-dependent*.

It is also possible to use hierarchical *ring* architectures for representing strings of words, which I believe is probably how the human cortical language architecture is organized. As the words are loaded into the ring of modules, they are quickly removed in groups (phrases) and re-represented in modules at a higher conceptual level, leaving the lower-level modules free for capturing additional words. I believe that this is why humans can only instantly remember in working memory "about 7 things ± 2" (Miller 1956) – we physically only have about seven modules at the word level. When required to remember a sequence of things, we repeatedly rehearse the sequence (to firmly store it in short-term memory) by traversing the ring from the beginning module (which is always the same one for each sentence or word sequence) to the last item and then back to the beginning. However, given the lack of limitations of computer implementations of confabulation architectures (at least conceptually), there is no need for us to use these more complicated ring architectures for this chapter's introductory discussion.

The knowledge bases of the architecture of Fig. 7.1 are all *causal*, meaning that the symbols of each module are linked only to symbols of later modules (i.e., those that lie to the right of it), which represent words which occur later in the temporal sequence of the word string. The first (i.e., leftmost) module is connected to all of the 19 modules which follow it by 19 individual knowledge bases; the second module to the 18 modules to its right; and so forth. Thus, this architecture has a total of 20 modules and 190 knowledge bases.

The training process starts with the first sentence of the training corpus and marches one sentence at a time to the last sentence. As each sentence is encountered it is entered into the architecture of Fig. 7.1 (unless its first 20 words include a word not among the 63,000, in which case the sentence is discarded) and used for *training*. Sentences with more than 20 words are truncated to 20. The details of training are now discussed.

At the beginning of training, 190 63,000×63,000 single precision float matrices are created (one for each knowledge base) and all of their entries are set to zero. In each knowledge base's matrix, each row corresponds to a unique source module symbol and each column corresponds to a unique target module symbol. The indices of the symbols of each module are arbitrary, but once set, they are frozen forever. These matrices are used initially, during training on the text corpus, to store the (integer) co-occurrence counts for the (causally) ordered symbol pairs of each knowledge base. Then, once these counts are accumulated, the matrices are used to calculate and store the (floating point) $p(\psi|\lambda)$ antecedent support probabilities. In practice, various computer science storage schemes for sparse matrices are used (in both RAM and on hard disk) to keep the total memory cost low.

Given a training sentence, it is entered into the modules of the architecture by activating the symbol representing each word or punctuation of the sentence, in order. Unused trailing modules are left blank (*null*). Then, each causal symbol pair is recorded in the matrix of the corresponding knowledge base by incrementing the numeric entry for that particular source symbol (the index of which determines the row of the entry) and target symbol (the index of which determines the column of the entry) pair by one.

After all of the many tens of millions of sentences of the training corpus have been used for training (i.e., the entire training corpus has been traversed from the first sentence to the last), the entries (ordered symbol pair *co-occurrence counts*) in each knowledge base's matrix are then used to create the knowledge links of that knowledge base.

Given a knowledge base matrix, what we have traditionally done is to first set to zero any counts which are below some fixed threshold (e.g., in some experiments three, and in others 25 or even 50). In effect, such low counts are thereby deemed random and *not meaningful*. Then, after these low-frequency co-occurrences have been set to zero, we use the "column sum" of each count matrix to determine the *appearance count* $c(\lambda)$ of each target symbol λ for a particular knowledge base. Specifically, if the count of co-occurrences of source symbol ψ with target symbol λ is $c(\psi,\lambda)$ (i.e., the matrix entry in row ψ and column λ), then we set $c(\lambda)$ equal to the *column sum* of the quantities $c(\phi,\lambda)$ over all source module symbols ϕ. Finally, the knowledge link probability $p(\psi|\lambda)$ is set equal to $c(\psi,\lambda)$ / $c(\lambda)$, which approximates the ratio $p(\psi\lambda)$ / $p(\lambda)$, which, by Bayes' law, is equal to $p(\psi|\lambda)$.

Note that the values of $c(\psi,\lambda)$, $c(\lambda)$, and $p(\psi|\lambda)$ for the same two symbols can differ significantly for different pairs of source and target modules within the sentence. This is because the appearances of particular words at various positions within a sentence differ greatly. For example, essentially no sentences begin with the uncapitalized word `and`. Thus, the value of $c(\psi,\lambda)$ will be zero for every knowledge base matrix with the first module as its source module and the symbol $\psi =$ `and` as the source symbol. However, for many other pairs of modules and target symbols this value will be large. [A technical point: These disparities are greatest at the early words of a sentence. At later positions in a sentence the $p(\psi|\lambda)$ values tend to be very much the same for the same displacement between the modules – probably the underlying reason why language can be handled well by a ring architecture.]

After the $p(\psi|\lambda)$ knowledge link probabilities have been created for all 190 knowledge bases using the above procedure, we have then traditionally set any of these quantities which are below some small value (in some experiments 0.0001, in others 0.0002, or even 0.0005) to zero, on the basis that such weak links reflect random and meaningless symbol co-occurrences. It is important to state that this policy (and the policy of zeroing out co-occurrence counts below some set number) is arbitrary and definitely subject to refinement (e.g., in the case of high-frequency target symbols we sometimes accept values below 0.0001 because these low-probability links can still be quite meaningful). The final result of this *training process* is the formation of 190 knowledge bases, each containing an average of a million or so individual items of knowledge.

Given this architecture, with its 20 modules and 190 knowledge bases, we can now consider some thought processes using it. The simplest is *phrase completion*. First, we take a coherent, meaningful, contiguous string of fewer than 20 words and represent them on the modules of the architecture, beginning with the first module. The goal is to use these words as context for selecting the next word in the string (which might be a punctuation; since these are represented in each of the modules). To be concrete, consider a situation where the first three words of a sentence are provided.

The three words are considered to be the *assumed facts* of the confabulation. They must be coherent and "make sense," else the confabulation process will yield no answers. To find the phrase completion we use the knowledge bases from the first, second, and third modules to the fourth. The completion is obtained by carrying out confabulation on the fourth module using a **W**3. The answer, if there is one, is then the symbol expressed on the fourth module after confabulation.

With only three words of context (e.g., `The only acceptable`), the answer that is obtained will often be one of a huge number of viable possibilities (`alternative`, `person`, `solution`, `flight`, `car`, `seasoning`, `etc.` – which can be obtained as an expectation by simply performing a **C**3). Language generation usually involves invoking longer-range or abstract *context* (expressed in some manner as a set, or multiple sets, of assumed facts that act as constraints on the completions or continuations) to more precisely focus the *meaning content* of the language construction (which, by the inherent nature of confabulation, is generally automatically grammatical and syntactically consistent). This context can arise from the same sentence (e.g., by supplying more, or more specific, words as assumed facts) or from external bodies of language (e.g., from previous sentences, as considered below in Sect. 7.3.3).

If we supply more assumed facts, or more narrowly specific assumed facts, confabulation can then supply the best answer from a much more restricted expectation. For example, `Mickey` and `Minnie` yields only one answer: `mouse`.

However, using more words in phrase completion (or in *sentence continuation*, where multiple successive words are added onto a starting string – see Chap. 6) introduces some new dilemmas. In particular, beyond a range of two or three words, the string of words that emerges is likely to be novel in the sense that some of the early assumed facts may not have knowledge links to distant, newly selected words in the word string. The design of confabulation architectures and thought processes to handle this common situation is a key problem that my research group has solved, at least in a preliminary way. As always, there is no software involved, just proper sequences of thought actions (module confabulations and knowledge base enablements) that are invoked by the conclusions of previous confabulations.

For example, consider the assumed facts `The canoe trip was going smoothly when all of a sudden`. Such partial sentences will almost certainly not have a next word symbol that receives knowledge links from all of the preceding assumed fact symbols. So what procedure shall we use to select the next word? One answer is to simply go on the preponderance of evidence: select that 12th module symbol that has the highest input intensity among those symbols which have the maximum available number of knowledge links. This is accomplished by **W**. This approach can yield acceptable answers some of the time, but it does not work as well as one would like. If we were to attempt sentence completion with this approach (i.e., adding multiple words), the results are awful.

The solution is to invoke two new confabulation architecture elements: a *language hierarchy* and *multiconfabulation*. These are sketched next.

7.3.2 Language Hierarchies

The reasons why the architecture of Fig. 7.1 does not solve the long phrase completion and sentence continuation problems are many. First of all, this architecture disallows the learning and application of *standard language constructions* such as multi-word *conceptual units* (e.g., New York Stock Exchange, which we will refer to as *phrases*), variable element constructions (*VEC*s, e.g., ___ went to the ___), and pendent clauses (e.g., **the success of her daughter was**, except for ordinary daily distractions, **foremost on her mind**). Standard constructions are important elements of all human languages (although they differ, and can take on different forms, in different languages), and a comprehensive architecture must include provisions for representing them.

For the problems of phrase completion and sentence continuation, the architecture of Fig. 7.2 is much more capable than that of Fig. 7.1. For example, consider again the problem of finding the next word for the assumed fact phrase The canoe trip was going smoothly when all of a sudden. Now, the first thing that happens is that this phrase is *parsed*, meaning that the words are re-represented at the phrase level. This happens almost instantly by using the knowledge bases which proceed from the word level to the phrase level. The parsing process, which is described next, proceeds in a rapid "rippling wave" of thought processes running from the beginning of the assumed fact word string to the end.

The first word of the string, The, goes up first. These links (in accordance with the knowledge base design described in the caption of Fig. 7.2) only go to the first phrase module. A **C1F** on this first phrase module yields an expectation consisting of those symbols which represent phrases that begin with the word The. The second word module then sends links upward to the first and second phrase modules from the symbol for canoe. **C1F**s on phrase modules one and two then do two things: on phrase module one, only the symbol for The remains (since The canoe, or any further extension of it, is not in the phrase module – for brevity, the manner in which the phrase module itself, and the additional knowledge bases of the Fig. 7.2 architecture, are derived using word-level knowledge – this process too is totally confabulation-implemented and does not use any linguistic knowledge – is not described here). This parsing process continues down the phrase modules, quickly yielding the parse (with the phrase symbol numbers in parentheses): The(8) canoe(25085) trip(1509) {was going}(63957) smoothly(9723) when(64) all(56) {of a sudden}(69902). Thus, phrase modules 1, 2, 3, 4, 6, 7, 8, and 9 have symbols active on them (each phrase is represented on the module immediately above its first word). All the other phrase modules have no symbols active on them. Note that if the last word (sudden) of the assumed fact phrase were not present, that phrase module 9 would not have a single phrase active on it, but would have several (representing all of the phrases that begin with of a: e.g., of a, of a kind, of a sudden, of a sort, etc.). Thanks to my colleague Robert W. Means for implementing and providing the details of this example.

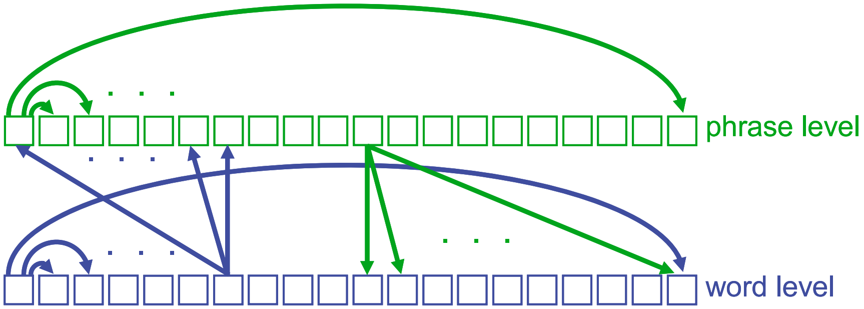


Fig. 7.2. Single-sentence hierarchical confabulation architecture for proper English phrase or sentence completion. The lower row of modules are used to represent words, as in the architecture of Fig. 7.1. Again, knowledge bases link each of the first 19 of these 20 *word-level* modules to all of the modules to their right. Positioned exactly above the word-level module row is a row of 20 phrase-level modules. These phrase *modules* represent word groups (and other standard language constructions, although these will not be discussed much in this chapter). Each phrase module has at least 126,000 symbols (63,000 single words and punctuations and the 63,000 most common multiple word groups). Knowledge bases connect each phrase module to each of the phrase modules which follow it. Knowledge bases also connect each phrase module to all of the word modules except those that lie to its left. Finally, knowledge bases connect each word module with all of the phrase modules except those that lie to its right. This architecture has a total of 800 knowledge bases. On average, each knowledge base contains roughly a million individual items of knowledge. The capability of this architecture is a practical demonstration of the main premise of the author's theory of vertebrate cognition; namely, that lots of simple knowledge, along with a single, simple information-processing operation can implement all of cognition

The above processing sounds like it would take a long time. Remember, however, that thinking is just like moving. When you throw a baseball many tens of muscles are being commanded in parallel in a precisely timed and coordinated way. The above thought process ("parse sentence") is stored, recalled, and executed just like a motor action such as throwing a baseball. The initiation of each involved knowledge base activation and confabulation happens in close succession in a "ripple" of processing that rapidly moves from the left end of the architecture to the right; terminating at the end of the assumed fact phrase. The entire parsing process is completed in just a small multiple of one knowledge base transmission time. Like some movement actions (e.g., dribbling a basketball); thought actions are often divided up into small "macro" segments which, depending upon their outcome (one active symbol, multiple highly excited symbols, or no symbol), trigger alternative next segments.

As discussed in Chap. 8, a key concept of hierarchical architecture design is the *precedence principle*. There, it is discussed in the context of the constitution of individual symbols within a single module. However, the same principle holds between lower and higher abstraction level modules within a hierarchy (such as that of Fig. 7.2). In this expanded form, what the precedence principle says is that as soon as content that is represented at a lower level of a hierarchy is re-represented at a higher level, the involved active lower-level module symbols must be shut off. This is implemented in human cerebral cortex by use of the *conclusion → action principle* (see Sect. 8.6).

In the case of the precedence principle, the action which is triggered by the expression of a phrase representation symbol is to shut off the modules which supplied words to the phrase that the symbol represents. For example, if the phrase that emerges from the parse has three words, the word module directly underneath the phrase module, and the next two modules to its right, are shut off (they stop expressing their word symbols). If the phrase has only one word, a different action is triggered; namely, only shutting off the module directly beneath. And so on. Note that these action commands are not issued until the choices have narrowed to a single symbol; since it is only then that the conclusion → action principle operates. This is a concrete example of how thought is not software. It is a series of sets of action commands, each set being immediately *originated* (issued to action nuclei) when a firm conclusion is reached (i.e., each conclusion has its own set of action commands that are permanently associated with it and which are originated every time that conclusion is expressed as the lone final result of a confabulation operation by its module).

This example illustrates a thought process that can be launched immediately with no further evaluation (e.g., by basal ganglia). It also illustrates how we will need to implement the action command output portion of cognition from the very outset of research. A great deal more could be said about action command generation and action symbol sequence learning and recall using confabulation architectures. However, this topic would take us beyond the introductory sketch being attempted in this chapter. Suffice it to say that quite a lot is known about how action sequences can be learned (by rehearsal), stored, and recalled using confabulation architectures (e.g., the UCSD graduate students in my course built a confabulation-based checker-playing system that learned to play by mimicking a skilled human). Confabulation architectures for appropriately modifying action sequences, in real time, in response to changes in the world state that occur during execution (a crucial capability if we are to perform in a complicated, real-world environment) have also been developed.

Obviously, when the conclusion → action principle "branching" capability is combined with an ability to store and retrieve data (e.g., using short-term, medium-term, long-term memory, or working memory), the cognitive brain passes the test of being, at least conceptually, capable of *universal computation* in the Turing sense. However, the very limited "RAM memory" or "tape memory" available for immediate reading and writing probably limits the value of this capability. Certainly, as demonstrated in Chap. 4, logical reasoning in Aristotelian information environments is carried out directly by confabulation (cogency maximization), without need for any recourse to computer principles. Nonetheless, a human with a paper and pencil (to supplement the extremely limited "RAM memory" available in the brain) can easily learn to carry out thought processes that will accurately simulate operation of a computer. However, such a "human-implemented computer" is to a modern desktop electronic computer as a unicycle is to a racing car.

Given the parse of an assumed fact phrase (say, the first few words of a sentence), we can then use the architecture of Fig. 7.2 to carry out "phrase completion" (as in Chap. 4). The first step is to build an expectation in the phrase module above the next word's module by activating the knowledge bases between the last active phrase module of the parse and that "target" phrase module and doing a **C1F**. This first step exploits the fact that adjacent phrases are usually highly coherent and it would be rare indeed for the next phrase to not receive knowledge links from the last known phrase of the parse. The result of this first step is an expectation on the target phrase module containing all of the reasonable next phrases. Note that the last-phrase module of the parse may itself have an expectation containing multiple symbols which themselves could contain the next word.

For example, as above, if `sudden` were not present in the starting word string phrase: `The canoe trip was going smoothly when all of a sudden`, the last-phrase module would have an expectation with multiple phrases, including `all of a sort`, `all of a sudden`, `all of a kind`, etc. If, for example, all of these symbols represent multi-word phrases then the target phrase module expectation will automatically be empty (since none of the phrases in the last-phrase module's expectation will have any knowledge links to symbols of that module). If this is not clear, using Fig. 7.2, work out some examples using a diagram on a piece of paper. This is a perfect example of how all thought processes are conclusion-driven.

The expectations established by the above process then send output through their knowledge links to the first unfilled word module; where an expectation is formed by a **C1F**. Since this word module is the next one after the last assumed fact word module, we can again assume that the symbols in this expectation represent all reasonable possibilities for the next word of the continuation. Then knowledge linking the rest of the parsed phrase symbols to the word module is used with a **W** to select that word symbol in the expectation which is most consistent with this additional context. Here again, there are many possible things that could go on (e.g., knowledge links may or may not exist from various phrase symbols to words of the expectation); yet, whatever the situation, this process works better than that using the architecture of Fig. 7.1. A bit of time spent thinking about this phrase completion process with some concrete examples will be most illuminating and compelling. Try to build some meaningful examples where this process won't work. You won't be able to.

Why would this phrase completion process (using the architecture of Fig. 7.2) be better than just using the word-level knowledge as described earlier in connection with Fig. 7.1? The answer is that knowledge links from phrases to words generally have two superior characteristics over links at the word level. First, the parse often removes a significant amount of ambiguity that can exist in word-level knowledge. For example, the word module symbol for word `New` will have strong links to the symbol for `Stock` two word modules later (independent of what word follows it). However, if the parse has activated the phrase `New Orleans` no such erroneous knowledge will be invoked. The other advantage of using the parsed representation is that the knowledge links tend to have a longer range of utility; since they represent originally extended conceptual collections that have been unitized.

If, as often occurs, we need to restore the words of a sentence to the word modules after a parse has occurred (and the involved word modules have been automatically shut off by the resulting action commands), all we need to do is activate all the relevant downward knowledge bases and simultaneously carry out confabulation on all of the word modules. This restores the word-level representation. If it is not clear why this will work, it may be useful to consider the details of Fig. 7.2 and the above description. The fact that "canned" thought processes (issued action commands), triggered by particular confabulation outcomes, can actually do the above information-processing, generally without mistakes, is rather impressive.

7.3.3 Multiconfabulation

For sentence continuation (adding more than just one word), we must introduce yet another new concept: *multiconfabulation* (sometimes termed *consensus building*). Multiconfabulation is simply a set of brief, but not instantaneous, temporally overlapping, mutually interacting confabulation operations that are conducted in such a way that the outcomes of each of the involved operations are consistent with one another in terms of the knowledge possessed by the system. Multiconfabulation involves *constraint satisfaction*, a classic topic introduced into neurocomputing in the early 1980s by studies of Boltzmann machines (Ackley et al. 1985).

For example, consider the problem of adding two more sensible words onto the following sentence-starting word string (or, simply *starter*): `The hyperactive puppy`. One approach would be to simply do a **W** simultaneously on the fourth and fifth word modules. This might yield: `The hyperactive puppy was water`; because `was` is the strongest fourth word choice and, based upon the first three words alone, `water` (as in `drank water`) is the strongest fifth word choice. The final result doesn't make sense.

But what if the given three-word starter were first used to create expectations on both the fourth and fifth modules (e.g., using **C3F**s). These would contain all the words consistent with this set of assumed facts. Then, what if **W**s on word modules four and five were carried out simultaneously, with a requirement that the only symbols on module five that will be considered are those which receive inputs from module four. Further, the knowledge links back to phrase modules having unresolved expectations from word modules four and five, and those in the opposite directions, are used as well to incrementally enhance the excitation of symbols that are consistent. Expectation symbols which do not receive incremental enhancement have their excitation levels incrementally decreased (to keep the total excitation of each expectation constant at 1.0). This multiple, mutually interacting, confabulation process is called *multiconfabulation* and it yields a *consensus* of conclusions (see below and Chap. 6).

Applying multiconfabulation yields sensible continuations of starters. For example, the starter `I was very` continues to: `I was very pleased with my team's`, and the starter `There was little` continues to: `There was little disagreement about what importance`. Thanks to my colleague Robert W. Means for these examples.

7.3.4 Multi-sentence Language Units

The ability to exploit long-range context using accumulated knowledge is one of the hallmarks of human cognition (and one of the glaring missing capabilities in today's computer and AI systems).

This section presents a simple example of how confabulation architectures can use long-range context and accumulated knowledge. The particular example

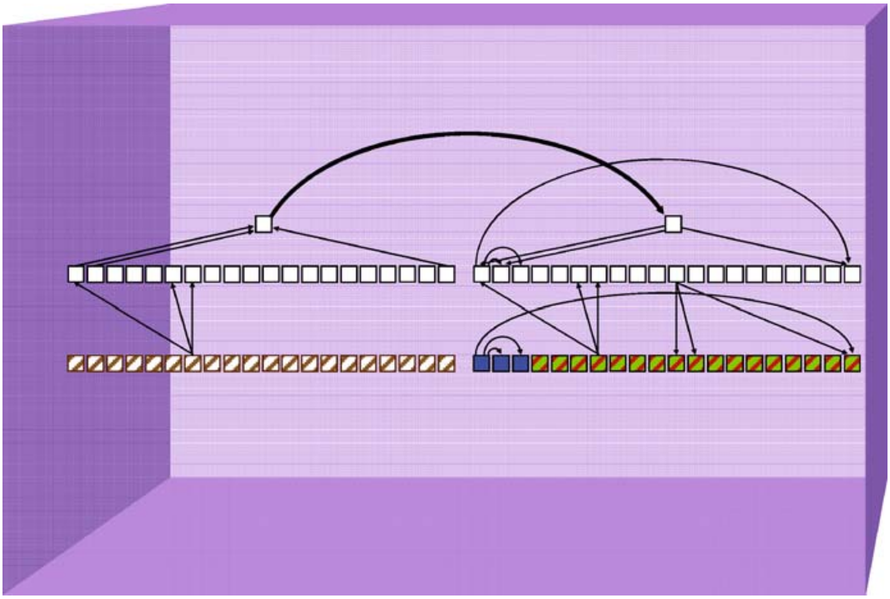
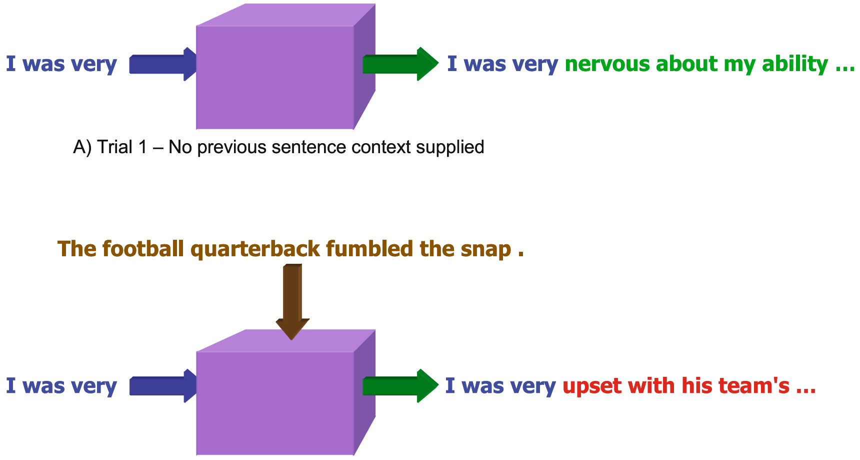


Fig. 7.3. Two-sentence hierarchical confabulation architecture for English text analysis or generation illustrated as the functional machinery of a "purple box." The sub-architectures for representing the first sentence (illustrated on the left) and that for the second sentence – the one to be continued – illustrated on the right) are each essentially the same as the architecture of Fig. 7.2, along with one new module and 20 new knowledge bases. The one additional module is shown above the *phrase layer* of modules of each sub-architecture. This *sentence meaning content summary module* contains symbols representing all of the 126,000 words and word groups of the phrase-level modules (and can also have additional symbols representing various other standard language constructions). Once the first sentence has been parsed its summary module has an expectation containing each phrase level module symbol (or construction subsuming a combination of phrase symbols) that is active. The (causal) *long-range context* knowledge base connects the summary module of the first sentence to the summary module of the second sentence


I was very
I was very nervous about my ability ...
The football quarterback fumbled the snap .
I was very
I was very upset with his team's ...

A) Trial 1 – No previous sentence context supplied

B) Trial 2 – Previous sentence context supplied

Fig. 7.4. Use of the "purple box" confabulation architecture of Fig. 7.3 for sentence continuation. Following knowledge acquisition (see text), the architecture's capabilities are evaluated by a series of testing events (each consisting of two *trials*). In Trial 1 (part A of the figure), three words, termed a sentence *starter* (shown in blue entering the architecture from the left) are entered into the architecture; without a previous sentence being provided. The architecture then uses its acquired knowledge and a simple, fixed, thought process to add some words; which are shown on the right in green appended to the starting words. In Trial 2 (part B of the figure), a previous context sentence (shown in brown being entered into the top of the architecture) is also provided. This alters the architecture's continuation output (shown in red). The context sentence (if one is being used on this trial) is entered into the left-hand sentence representation module of Fig. 7.3 and the starter is entered into the first three words of the right-hand module. A simple, fixed, "swirling" multiconfabulation thought process then proceeds to generate the continuation

considered is an extension of the architecture of Fig. 7.2 namely, the one explored in detail in Chap. 6. The remainder of this sub-section is largely a recapitulation of Chap. 6. However, the reader may find this useful because some of the involved issues are discussed from a somewhat different viewpoint.

The confabulation architecture illustrated in Fig. 7.3 allows the meaning content of a previous sentence to be brought to bear on the continuation, by multiconfabulation, of a starter (shown in green in Fig. 7.3) for the second sentence. The use of this architecture, following knowledge acquisition is illustrated in Fig. 7.4 (where, for simplicity, the architecture of Fig. 7.3 is represented as a "purple box"). This architecture, its education, and its use are now briefly explained.

The sentence continuation architecture shown in Fig. 7.3 contains two of the architectures of Fig. 7.2 along with two new sentence meaning content *summary* modules (one above each sentence architecture). The left-hand modules are used to represent the context sentence, when it is present. The right-hand modules represent the sentence to be continued.

To prepare this architecture for use, it is educated by selecting pairs of topically coherent successive sentences, belonging to the same paragraph, from a general coverage, multi-billion-word proper English text corpus. This sentence pair selection process can be done by hand by a human or using a simple computational linguistics algorithm. Before beginning education, each individual sentence module was trained in isolation on the sentences of the corpus.

During education of the architecture of Fig. 7.3, each selected sentence pair (of which roughly 50 million were used in the experiment described here) was loaded into the architecture, completely parsed (including the summary module), and then counts were accumulated for all ordered pairs of symbols on the summary modules. The long-term context knowledge base linking the first sentence to the second was then constructed in the usual way, using these counts. This education process takes about two weeks on a PC-type computer (see Chap. 6 for details).

Figure 7.5 (spanning the next few pages) illustrates the architecture evaluation process. During each testing episode, two *evaluation trials* are conducted: one with no previous sentence (to establish baseline continuation) and one with a previous sentence (to illustrate the changes in the continuation that the availability of context elicited). For example, if no previous sentence was provided, and the first three words of the sentence to be continued were **The New York**, then the architecture constructed: **The New York Times' computer model collapses ...** (where the words added by this sentence continuation process without context are shown in green). However, if the previous context sentence **Stocks proved to be a wise investment .** was provided, then, again beginning the next sentence with **The New York**, the architecture constructed **The New York markets traded lower yesterday ...** (where, as in Fig. 7.4, the words added by the sentence continuation process are shown in red). Changing the context sentence to **Downtown events were interfering with local traffic .**, the architecture then constructs **The New York City Center area where ...**. Changing the context sentence to **Coastal homes were damaged by tropical storms .** yields **The New York City Emergency Service System ...**. And so on. Below are some other examples (first line: continuation without context, second line: previous sentence supplied to the architecture, third line: continuation with the previous sentence context):

The New York Times' computer model collapses ...
Medical patients tried to see their doctors .
The New York University Medical Association reported ...

But the other semifinal match between fourth-seeded ...
Chile has a beautiful capital city .
But the other cities have their size ...

But the other semifinal match between fourth-seeded ...
Japan manufactures many consumer products .
But the other executives included well-known companies ...

When the United Center Party leader urged …
The car assembly lines halted due to labor strikes .
When the United Auto Workers union representation …

When the United Center Party leader urged …
The price of oil in the Middle East escalated yesterday .
When the United Arab Emirates bought the shares …

But the Roman Empire disintegrated during the fifth …
She learned the history of the saints .
But the Roman Catholic population aged 44 …

But the Roman Empire disintegrated during the fifth …
She studied art history and classical architecture .
But the Roman Catholic church buildings dating …

The San Francisco Redevelopment Authority officials announced …
Their star player caught the football and ran !
The San Francisco quarterback Joe Brown took …

The San Francisco Redevelopment Authority officials announced …
The pitcher threw a strike and won the game .
The San Francisco fans hurled the first …

The San Francisco Redevelopment Authority officials announced …
I listen to blues and classical music .
The San Francisco band draws praise from …

The San Francisco Redevelopment Authority officials announced …
Many survivors of the catastrophe were injured .
The San Francisco Police officials announced Tuesday …

The San Francisco Redevelopment Authority officials announced …
The wheat crops were genetically modified .
The San Francisco food sales rose 7.3 …

I was very nervous about my ability …
The football quarterback fumbled the snap .
I was very upset with his team's …

I was very nervous about my ability …
Democratic citizens voted for their party's candidate .
I was very concerned that they chose …

I was very nervous about my ability …
Restaurant diners ate meals that were served .
I was very hungry while knowing he had …

In spite of yesterday's agreement among analysts …
The Mets were not expected to win .
In spite of the pitching performance of some …

In spite of yesterday's agreement among analysts …
The President was certain to be reelected .
In spite of his statements toward the government …

In spite of yesterday's agreement among analysts …
She had no clue about the answer .
In spite of her experience and her …

In the middle of the 5th century BC …
Mike Piazza caught the foul ball .
In the middle of the season came …

In the middle of the 5th century BC …
The frozen lake was still very dangerous .
In the middle of the lake is a …

It meant that customers could do away …
The stock market had fallen consistently .
It meant that stocks could rebound later …

It meant that customers could do away …
I was not able to solve the problem .
It meant that we couldn't do much better …

It meant that customers could do away …
The company laid off half its staff .
It meant that if employees were through …

It meant that customers could do away …
The salesman sold men's and women's shoes .
It meant that sales costs for increases …

It must not be confused about what …
The effects of alcohol can be dangerous .
It must not be used without supervision …

It must not be confused about what …
The subject was put to a vote .
It must not be required legislation to allow …

It was a gutsy performance by John …
The tennis player served for the match .
It was a match played on grass …

It was a gutsy performance by John …
Coastal homes were damaged by tropical storms .
It was a huge relief effort since …

It was a gutsy performance by John …
The ship's sails swayed slowly in the breeze .
It was a long ride from the storm …

She thought that would throw us away …
The tennis player served for the match .
She thought that she played a good …

Shortly thereafter , she began singing lessons …
The baseball pitcher threw at the batter .
Shortly thereafter , the Mets in Game …

Shortly thereafter , she began singing lessons …
Democratic citizens voted for their party's candidate .
Shortly thereafter , Gore was elected vice president …

The president said he personally met French …
The flat tax is an interesting proposal .
The president said he promised Congress to let …

The president said he personally met French …
The commission has reported its findings .
The president said he appointed former Secretary …

The president said he personally met French …
The court ruled yesterday on conflict of interest .
The president said he rejected the allegations …

This resulted in a substantial performance increase …
The state governor vetoed the bill .
This resulted in both the state tax …

This resulted in a substantial performance increase …
Oil prices rose on news of increased hostilities .
This resulted in cash payments of $ …

This resulted in a substantial performance increase …
The United States veto blocked the security council resolution .
This resulted in both Britain and France …

Three or four persons who have killed …
The tennis player served for the match .
Three or four times in a row …

We could see them again if we …
The president addressed congress about taxes .
We could see additional spending money bills …

We could see them again if we …
The view in Zion National Park was breathtaking .
We could see snow conditions for further …

We could see them again if we …
We read the children's books out loud .
We could see the children who think …

We could see them again if we …
The U.N. Security Council argued about sanctions .
We could see a decision must soon …

What will occur during the darkest days …
Research scientists have made astounding breakthroughs .
What will occur within the industry itself …

What will occur during the darkest days …
The vacation should be very exciting .
What will occur during Christmas season when …

What will occur during the darkest days …
I would like to go skiing .
What will occur during my winter vacation …

What will occur during the darkest days …
There's no way to be certain .
What will occur if we do nothing …

When the Union Bank launched another 100 …
She loved her brother's Southern hospitality .
When the Union flag was raised again …

When the Union Bank launched another 100 …
New York City theater is on Broadway .
When the Union Square Theater in Manhattan …

Fig. 7.5. Confabulation architecture evaluation process (see text)

A good analogy for this system is a child learning a human language. Young children need not have any formal knowledge of language or its structure in order to generate it effectively. Consider what this architecture must "know" about the objects of the world (e.g., their attributes and relationships) in order to generate these continuations, and what it must "know" about English grammar and composition. Is this the world's first AI system? You decide.

Note that in the above examples the continuation of the second sentence in context was conducted using an (inter-sentence, long-range context) knowledge base educated via exposure to meaning-coherent sentence pairs selected by an external agent. When tested with context, using completely novel examples, it then produced continuations that are meaning-coherent with the previous sentence (i.e., the continuations are rarely unrelated in meaning to the context sentence). Think about this for a moment. This is a valuable general principle with endless implications. For example, we might ask: How can a system learn to carry on a conversation? Answer: simply educate it on the conversations of a master human conversationalist! There is no need or use for a "conversation algorithm." Confabulation architectures work on this *monkey-see/monkey-do* principle. If these statements upset you, then you are one of those exquisite few who actually delve into details. Your reward is to now understand how profoundly alien confabulation theory is in the context of the panorama of classical information-processing and systems neuroscience. Again: Conversation involves no algorithms whatsoever.

This sentence continuation example reveals the true nature of cognition: it is based on ensembles of properly phased confabulation processes mutually interacting via knowledge links. Completed confabulations provide assumed facts for confabulations newly underway. Contemporaneous confabulations achieve mutual "consensus" via rapid interaction through knowledge links as they progress (thus the term *multiconfabulation*). There are no algorithms anywhere in cognition; only such ensembles of confabulations. This illustrates the starkly

alien nature of cognition in comparison with existing neuroscience, computer science, and AI concepts.

In speech cognition (see Sect. 7.4), elaborations of the architecture of Fig. 7.3 can be used to define expectations for the next word that might be received (which can be used by the acoustic components of a speech understanding system), based upon the context established by the previous sentences and previous words of the current sentence which have been previously transcribed. For *text generation* (a generalization of sentence continuation, in which the entire sentence is completed with no starter), the choices of words in the second sentence can now be influenced by the context established by the previous sentence. The architecture of Fig. 7.3 generalizes to using larger bodies of context for a variety of cognition processes.

Even more abstract levels of representation of language meaning are possible. For example, after years of exposure to language and co-occurring sensory and action representations, modules can form that represent sets of commonly encountered lower-abstraction-level symbols. Via the SRE mechanism (a type of thought process), such symbols take on a high level of abstraction, as they become linked (directly, or via equivalent symbols) to a wide variety of similar-meaning symbol sets. Such symbol sets need not be complete to be able to (via confabulation) trigger activation of such high-abstraction representations. In language, these highest-abstraction-level symbols often represent words! For example, when you activate the symbol for the word joy, this can mean joy as a word, or joy as a highly abstract concept. This is why in human thought the most exalted abstract concepts are made specific by identifying them with words or phrases. It is also common for these most abstract symbols to belong to a foreign language. For example, in English speaking lands, the most sublime abstract concepts in language are often assigned to French, or sometimes German, words or phrases. In Japanese, English or French words or phrases typically serve in this capacity.

High-abstraction modules are used to represent the meaning content of *objects* of the mental world of many types (language, sound, vision, tactile, etc.). However, outside of the language faculty, such symbols do not typically have names (although they are often strongly linked with language symbols). For example, there is probably a module in your head with a symbol that abstractly encodes the combined taste, smell, surface texture, and masticational feel of a macaroon cookie. This symbol has no name, but you will surely know when it is being expressed!

7.3.5 Discussion

A key observation is that confabulation architectures automatically learn and apply grammar, and honor syntax, without any in-built linguistic structures, rules, or algorithms. This strongly suggests that grammar and syntax are fictions dreamed up by linguists to explain an orderly structure that is actually a by-product of the mechanism of cognition. Otherwise put, for cognition to be able, given the limitations of its native machinery, to efficiently deal with language, that language must have a structure which is compatible with the mathematics of confabulation and multiconfabulation. In this view, every functionally usable human language must be structured this way. Ergo, the universality of grammar and syntactic structure in all human languages.

Thus, Noam Chomsky's (Chomsky 1980) famous long search for a universal grammar (which must now be declared over) was both correct and incorrect. Correct, because if you are going to have a language that cognition can deal with at a speed suitable for survival, grammar and syntactic structure are absolute requirements (i.e., languages that don't meet these requirements will either adapt to do so, or will be extincted with their speakers). Thus, grammar is indeed universal. Incorrect, because grammar itself is a fiction. It does not exist. It is merely the visible spoor of the hidden underlying native machinery of cognition: confabulation and antecedent support knowledge.

7.4 Sound Cognition

Unlike language, which is the centerpiece and masterpiece of human cognition, many of the other functions of cognition (e.g., sensation and action) must interact directly with the outside world. Sensation requires conversion of externally supplied sensory representations into symbolic representations and vice versa for actions. This section, and the next (discussing vision), must therefore discuss not only the confabulation architectures used, but also cover the implementation of this *transduction* process, which is necessarily different for each of these cognitive modalities. Readers are expected to have a solid understanding of traditional speech signal processing and speech recognition.

7.4.1 Representation of Multi-source Soundstreams

Figure 7.6 illustrates an "audio front end" for transduction of a soundstream into a string of "multi-symbols," with a goal of carrying out ultra-high-accuracy speech transcription for a single speaker embedded in multiple interfering sound sources (often including other speakers). The description of this design does not concern itself with computational efficiency. Given a concrete design for such a system, there are many well-known signal-processing techniques for implementing approximately the same function, often orders of magnitude more efficiently. For the purpose of this introductory treatment (which, again, is aimed at illustrating the universality of confabulation as the mechanization of cognition), this audio front end design does not incorporate embellishments such as binaural audio imaging.

Referring to Fig. 7.6, the first step in processing is analog speech lowpass filtering (say, with a flat, zero-phase-distortion response from DC to 4 kHz, with a steep rolloff thereafter) of the high-quality (say, over 110 dB dynamic range) analog microphone input. Following bandpass filtering, the microphone signal

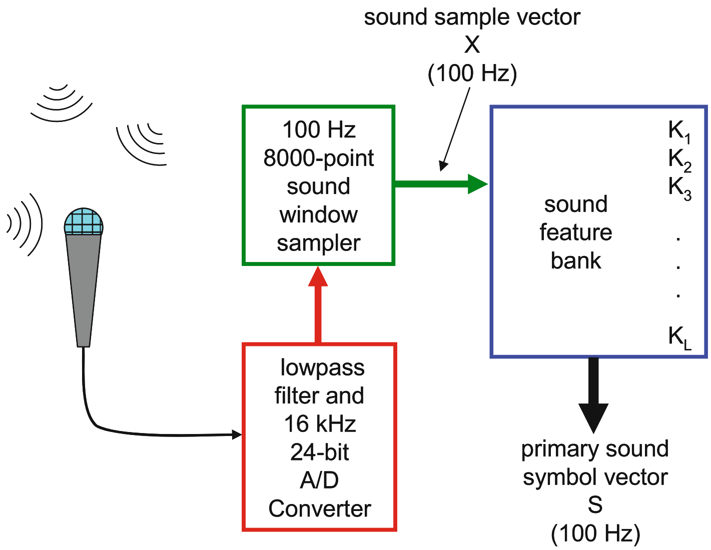


Fig. 7.6. An audio front end for representation of a multi-source soundstream. See text for details

is sampled with an (e.g., 24-bit) analog to digital converter operating at a 16 kHz sample rate. The combination of high-quality analog filtering, sufficient sample rate (well above the Nyquist rate of 8 kHz) and high dynamic range yields a digital output stream with almost no artifacts (and low information loss). Note that digitizing to 24 bits supports exploitation of the wide dynamic ranges of modern high-quality microphones. In other words, this dynamic range will make it possible to accurately understand the speech of the attended speaker, even if there are much higher amplitude interferers present in the soundstream.

The 16 kHz stream of 24-bit signed integer samples generated by the above preprocessing (see Fig. 7.6) is next converted to floating point numbers and blocked up in time sequence into 8,000-sample windows (8,000-dimensional floating point vectors), at a rate of one window every 10 ms. Each such *sound sample vector* X thus overlaps the previous such vector by 98% of its length (7,840 samples). In other words, each X vector contains 160 new samples that were not in the previous X vector (and the "oldest" 160 samples in that previous vector have "dropped off the left end").

As shown in Fig. 7.6, the 100 Hz stream of sound sample vectors then proceeds to a *sound feature bank*. This device is based upon a collection of L fixed, 8,000-dimensional floating point *feature vectors*: K_1, K_2, … , K_L (where L is typically a few tens of thousands). These feature vectors represent a variety of *sound detection correlation kernels*. For example: gammatone wavelets with a wide variety of frequencies, phases, and gamma envelope lengths; broadband impulse detectors; fricative detectors; etc. When a sound sample vector X arrives at the feature bank the first step is to take the inner product of X with each of the

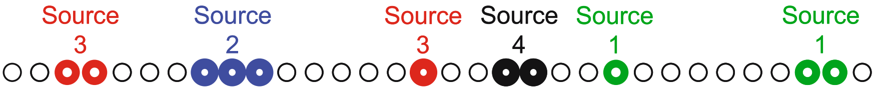


Fig. 7.7. Illustration of the properties of a primary sound symbol excitation vector S (only a few of the L components of S are shown). Excited symbols have thicker circles. Each of the four sound sources present (at the moment illustrated) in the auditory scene being monitored is causing a relatively small subset of feature symbols to be excited. Note that the symbols excited by sources 1 and 3 are not contiguous. That is typical. Keep in mind that the number of symbols, L (which is equal to the number of feature vectors) is typically tens of thousands, of which only a small fraction are meaningfully excited. This is because each sound source only excites a relatively small number of sound features at each moment and typical audio scenes contain only a relatively small number of sound sources (typically fewer than 20 monaurally distinguishable sources, with all other sources merging into an unresolved "audio background noise")

L feature vectors, yielding L real numbers: $(X \cdot K_1)$, $(X \cdot K_2)$, … , $(X \cdot K_L)$. These L values form the *raw feature response vector*. The individual components of the raw feature response vector are then each subjected to further processing (e.g., discrete time linear or quasi-linear filtering), which is customized for each of the L components. Finally, the logarithm of the square of each component of this vector is taken. The net output of the sound feature bank is an L-component non-negative *primary sound symbol excitation vector* S (see Fig. 7.6). A new S vector is issued every 10 ms.

The criteria used in selection of the feature vectors are low information loss, sparse representation (a relatively small percentage of S components meaningfully above zero at any time due to any single sound source), and low rate of individual feature response to multiple sources. By this latter it is meant that, given a typical application mix of sources, the probability of any feature which is meaningfully responding to the incoming soundstream at a particular time being stimulated (at that moment) by sounds from more than one source in the auditory scene is low. The net result of these properties is that S vectors tend to have few meaningfully non-zero components per source, and each sound symbol with a significant excitation is responding to only one sound source [see (Sagi et al. 2001) for a concrete example of a sound feature bank].

Figure 7.7 illustrates a typical primary sound symbol excitation vector S. This is the mechanism of analog sound input transduction into the world of symbols. One hundred times per second a new S vector is created. S describes the content of the sound scene being monitored by the microphone at that moment. Each of the L components of S (again, L is typically tens of thousands) represents the response of one sound feature detector (as described above) to this current sonic scene.

S is composed of small, mostly disjoint (but usually not contiguous), subsets of excited sound symbol components – one subset for each sound source in the current auditory scene. Again, each excited symbol is typically responding to the sound emanating from only one of the sound sources in the audio scene being monitored by the microphone. While this single-source-per-excited-symbol rule is not strictly true all the time, it is almost always true (which, as we will see, is all that matters). Thus, if, at each moment, we could somehow decide which subset of excited symbols of the symbol excitation vector to *pay attention* to, we could ignore the other symbols and thereby focus our attention on one source. That is the essence of all initial cortical sensory processing (auditory, visual, gustatory, olfactory, and somatosensory): figuring out, in real time, which primary sensor input representation symbols to pay attention to, and ignoring the rest. This ubiquitous cognitive process is termed *attended object segmentation*.

7.4.2 Segmenting the Attended Speaker and Recognizing Words

Figure 7.8 shows a confabulation architecture for directing attention to a particular speaker in a soundstream containing multiple sound sources and also recognizing the next word they speak. For a concrete example of a simplified version of this architecture (which nonetheless can competently carry out these kinds of functions), see Sagi et al. (2001). This architecture will suffice for the purposes of this introduction, but would need to be further augmented (and streamlined for computational efficiency) for practical use.

Each 10 ms a new S vector is supplied to the architecture of Fig. 7.8. This S vector is directed to one of the primary sound modules; namely, the next one (moving from left to right) in sequence after the one which received the last S vector. It is assumed that there are a sufficient number of modules so that all of the S vectors of an individual word have their own module. Of course, this

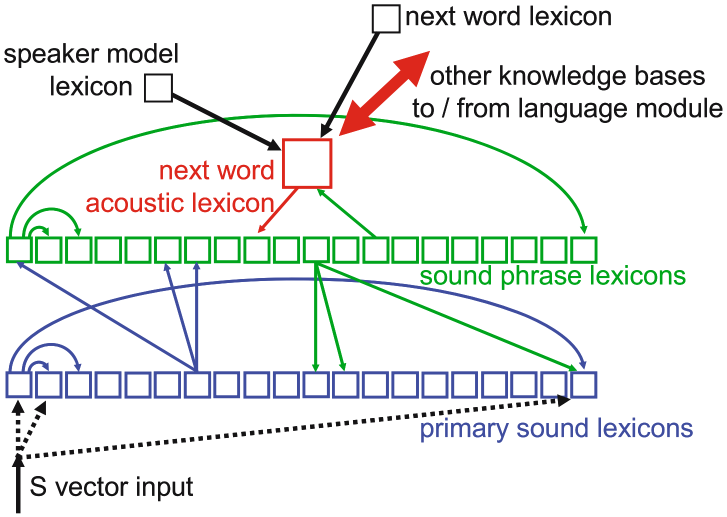


Fig. 7.8. Speech transcription architecture. The key components are the *primary sound modules*, the *sound phrase modules*, and the *next-word acoustic module*. See text for explanation

requires 100 modules for each second of word sound input, so a word like antidisestablishmentarianism will require hundreds of modules. For illustrative purposes, only 20 primary sound modules are shown in Fig. 7.8. Here again, in an operational system, one would simply use a ring of modules [which is probably what the cortical "auditory strip" common to many mammals, including humans (Paxinos and Mai 2004), probably is – a linear sequence of modules which functionally "wraps around" from its physical end to its beginning to form a ring].

The architecture of Fig. 7.8 presumes that we know approximately when the last word ended. At that time, a thought process is executed to erase all of the modules of the architecture, feed in expectation-forming links from external modules to the next-word acoustic module (and form the next-word expectation), and redirect S vector input to the first primary sound module (the one on the far left). [Note: As is clearly seen in mammalian auditory neurophysiology, the S vector is wired to all portions (modules) of the strip in parallel. The process of "connecting" this input to one selected module (and no other) is carried out by manipulating the operating command of that one module. Without this operate command input manipulation, which only one module receives at each moment, the external sound input is ignored.]

The primary sound modules have symbols representing a statistically complete coverage of the space of momentary sound vectors S that occur in connection with auditory sources of interest, when they are presented in isolation. So, if there are, say, 12 sound sources contributing to S, then we would nominally expect that there would be 12 sets of primary sound module symbols responding to S (this follows because of the "quasi-orthogonalized" nature of S, e.g., as depicted in Fig. 7.7). Mathematically, the symbols of each primary sound module are a vector quantizer (Zador 1963) for the set of S vectors that arise, from all sound sources that are likely to occur, when each source is presented in isolation (i.e., no mixtures). Among the symbol sets that are responding to S are some that represent the sounds coming from the attended speaker. This illustrates the critically important need to design the acoustic front end so as to achieve this sort of *quasi-orthogonalization of sources*. By confining each sound feature to a properly selected time interval (a sub-interval of the 8,000 samples available at each moment, ending at the most recent 16 kHz sample), and by using the proper post-filtering (after the dot product with the feature vector has been computed), this quasi-orthogonalization can be accomplished. [Note: This scheme answers the question of how brains carry out "independent component analysis" (Hyvärinen et al. 2001). They don't need to. Properly designed quasi-orthogonalizing features, adapted to the pure sound sources that the critter encounters in the real world, map each source of an arbitrary mixture of sources into its own separate components of the S vector. In effect, this is essentially a sort of "one-time ICA" feature development process carried out during development and then essentially frozen (or perhaps adaptively maintained) for life. Given the stream of S vectors, the confabulation processing which follows (as described below) can then, at each moment, ignore all but the attended-source-related subset of components, independent of how many, or few, interfering sources are present. Of course, this is exactly what is observed in mammalian audition – effortless segmentation of the attended source at the very first stage of auditory (or visual, or somatosensory, etc.) perception.]

The expectation formed on the next-word acoustic module of Fig. 7.8 (which is a huge structure, almost surely implemented in the human brain by a number of physically separate modules) is created by successive **C1F**s. The first is based on input from the speaker model module. The only symbols (each representing a stored acoustic model for a single word – see below) that then remain available for further use are those connected with the speaker currently being attended to.

The second **C1F** is executed in connection with input from the language module word module that has an expectation on it representing possible predictions of the next word that the speaker will produce (this next-word module expectation is produced using essentially the same process as was described in Sect. 7.3 above in connection with sentence continuation with context). (NOTE: This is an example of the situation mentioned above and in Chap. 8, where an expectation is allowed to transmit through a knowledge base.) After this operation, the only symbols left available for use on the next-word acoustic module are those representing expected words spoken by the attended speaker. This expectation is then used for the processing involved in recognizing the attended speaker's next word.

As shown in Fig. 7.8, knowledge bases have previously been established (using pure source, or well-segmented source, examples) to and from the primary sound symbol modules with the sound phrase modules and to and from these with the next-word acoustic module. Using these knowledge bases, the expectation on the next-word acoustic module is *transferred* (as described immediately above), via the appropriate knowledge bases, to the sound phrase modules, where expectations are formed; and from these to the primary sound modules, where additional expectations are formed. It is easy to imagine that, since each of these transferred expectations is typically much larger than the one from which it came, that by the time this process gets to the primary sound modules the expectations will encompass almost every symbol. This is not so! While these primary module expectations are indeed large (they may encompass many hundreds of symbols), they are still only a small fraction of the total set of tens of thousands of symbols. Given these transfers, which actually occur as soon as the recognition of the previous word is completed – which is often long before its full acoustic content has arrived, the architecture is prepared for detecting the next word spoken by the attended speaker.

As each S vector arrives at the architecture of Fig. 7.8, it is sent to the proper module in sequence. For simplicity, let us assume that the first S vector associated with the initial sound content of the next word is sent to the first primary sound module (if it goes to the "wrong" module, or is missed altogether, it doesn't matter much – as will be explained below). Given that the first primary sound module has an expectation, and that the only symbols in this expectation are those that represent sounds that a speaker of this type (we each have hundreds of "canonical models" of speakers having different accents and vocal apparati, and most of us add to this store throughout life) speaking early parts of one of the words we are expecting. Again note that, because of the orthogonalized nature of the S vector and the pure-signal nature of the primary feature symbols, each of the symbols in this expectation will typically represent sounds having only a tiny number of S vector components that are non-zero. Each symbol in a primary sound module is expressed as a unit vector having these small number of components with coefficients near 1, and all other components at zero. The module takes the inner product of each symbol's vector expression with S and this is then used as that symbol's *initial input excitation* (this is how symbols get excited by sensory input signals, in contrast to how symbols get excited by knowledge links from other symbols, which was discussed in Sect. 7.1). We have now completed the transition from acoustic space to symbol space.

Notice that the issue of signal level of the attended source has not been discussed. As described in Sect. 7.3.1, each S vector component has its amplitude expressed on a logarithmic scale (based on "sound power amplitudes" ranging across many orders of magnitude). Thus, on this scale, the inner product of S with a particular symbol's unit vector will still (because of the linear nature of the inner product) be substantial, even if the attended source sounds are tens of decibels below those of some individual interferers. Thus, with this design, attending to weak, but distinct, sources is generally possible. These are, of course, the characteristics we as humans experience in our own hearing. Further, in auditory neuroscience, such logarithmic coding of sound feature response signals (in particular, those from the brainstem auditory nuclei to the medial geniculate nucleus, which are the auditory signals analogous to the components of S) is well established (Oertel et al. 2002).

During the entire time of the word detection processes, all of the modules of the Fig. 7.8 architecture are operated in a multiconfabulation mode. Thus, as soon as the S-input excitations are established on the expectation element symbols of the first primary sound module, only those symbols which received these expectations remain on the expectation (the multiconfabulation is run faster on the primary sound modules, somewhat slower on the sound phrase modules, and even slower on the next-word acoustic module). This process of expectation refinement that occurs during multiconfabulation is termed *honing*.

After acoustic input has arrived at each subsequent primary sound module (the pace of the switching is set by a separate part of the auditory system, which will not be discussed further here, which synchronizes the pace of S vector formation – no, it is not always exactly every 10 ms – to the recent pace of speech production of the attended speaker), that module's expectation is thereby honed and this revised expectation is then automatically transferred to all of the sound phrase modules that are not on its right (during multiconfabulation, all of the involved knowledge bases remain operational). This has the effect of honing some of the sound phrase module expectations, which then are transferred to the next-word acoustic module; honing its expectation.

This process works in reverse also. As higher-level module expectations are honed, these are transferred to lower levels, thereby refining those lower-level expectations. Note that if occasional erroneous symbols are transferred up to the sound phrase modules, or even from the phrase modules to the next-word acoustic module, this will not have much effect. That is because the process of multiconfabulation effectively "integrates" the impact of all of the incoming transfers on the symbols of the original expectation. Only when a phrase module has honed its symbol list down to one symbol (which then becomes active) is a final decision made at that level. Similarly, only at the point where the expected word duration has been reached does the next-word acoustic module make a decision (or, it can even transfer a small expectation back to the language module – which is one way robust operation is aided).

Figure 7.9 illustrates the multiconfabulation process on the next-word acoustic module as the S vector is directed to each subsequent primary sound module in turn. As honed expectations are transferred upward, the expectation of the next-word acoustic module is itself honed. This honed expectation is then transferred downward to refine the expectations of the as-yet-unresolved sound phrase and primary sound layer modules. Unlike ordinary confabulations, these multiconfabulation interactions happen dynamically in continuous time as the involved operation commands are slowly tightened. This again illustrates the almost exact analogy between thought and movement. As with a movement, these smoothly changing, precisely controlled, multiconfabulation module operate commands are generated by a set of modules (in frontal cortex) that specialize in storing and recalling action symbol sequences.

A common objection about this kind of system is that as long as the expectations keep being met, the process will keep working. However, if even one glitch occurs, it looks like the whole process will fall apart and stop working. Then, it will somehow have to be restarted (which is not easy – for example, it may require the listener to somehow get a high enough signal-to-noise ratio to allow a much cruder trick to work). Well, this objection is quite wrong. Even if the next word and the next-after-that word are not one of the expected ones, this architecture will often recover and ongoing speechstream word recognition will continue, as we already proved with our crude initial version (Sagi et al. 2001). A problem that can reliably make this architecture fail is a sudden major change in the pace of delivery, or a significant brief interruption in delivery. For example, if the speaker suddenly starts speaking much faster or much slower the mentioned sub-system that monitors and sets the pace of the architecture's operation will cause the timing of the multiconfabulation and word-boundary segmentation to be too far off. Another problem is if the speaker gets momentarily tongue-tied and inserts a small unexpected sequence of sounds in a word (try this yourself by smoothly inserting the brief meaningless sound "bryka" in the middle of a word at a cocktail party – the listener's Fig. 7.8 architecture will fail and they will be forced to move closer to get clean recognitions to get it going again).

A strong tradition in speech recognition technology is an insistence that speech recognizers be "time-warp insensitive" (i.e., insensitive to changes in the pace of word delivery). Well, Fig. 7.8 certainly is not strongly "time-warp insensitive," and, as pointed out immediately above, neither are humans! However,

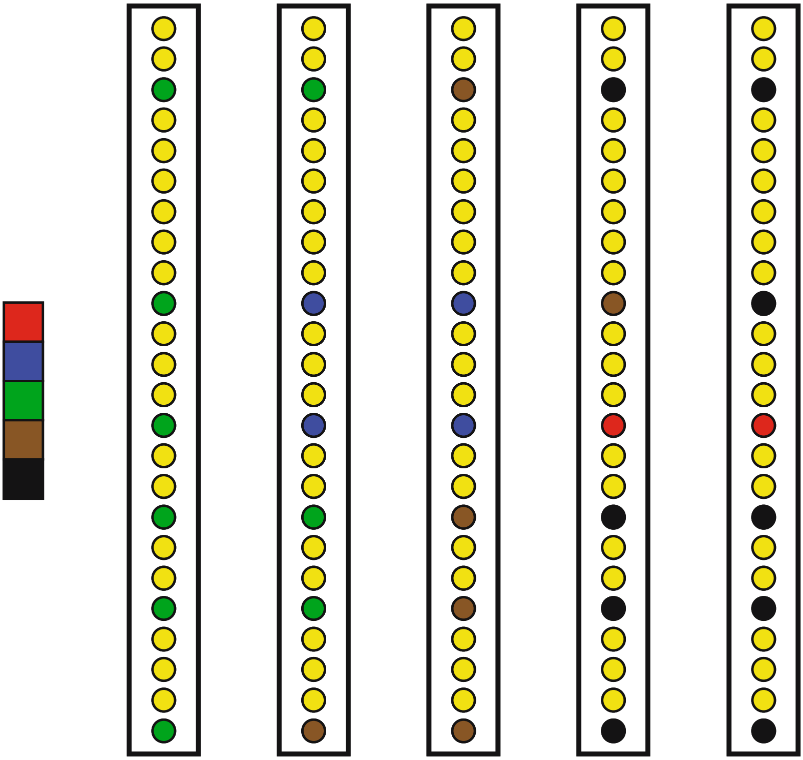


Fig. 7.9. Multiconfabulation process on the next-word acoustic module (see Fig. 7.8). Initially, symbols in the next-word expectation (*green* dots in the *left-most* representation of the module state) are established by knowledge link inputs from the speaker model module and from the language module. As multiconfabulation progresses, transfers of honed expectations from sound phrase modules (which themselves are receiving transfers from primary sound modules) hone this initial expectation, as illustrated here moving from left to right. Yellow-filled circles represent symbols that were not part of the initial expectation. These are locked at zero excitation. The color chart on the left shows the positive excitation scale from lowest on the bottom to highest on top. Some of the initial expectation symbols become progressively promoted to higher levels of excitation (the sum of all symbol excitations is roughly constant during multiconfabulation). Others go down in excitation (it is possible for a symbol to change non-monotonically, but that is not illustrated here. In the end state of the module (*far right*) one symbol (*red*) has become active – this symbol represents the word that has been detected. Keep in mind that in a real architecture there would typically be tens of thousands of symbols and that only a few percent, at most, would be part of the initial expectation

modest levels of time warp have no impact, since this just changes the location of the phrase module (moves it slightly left or right of its nominal position) where a particular phrase gets detected. Also note that since honed phrase expectations are transferred, it is not necessary for all of the primary sound symbols of a phrase to be present in order for that phrase to contribute significantly to the “promotion” of the next-word acoustic module symbols that receive links from it. Thus, many primary symbols can be missed with no effect on correct word recognition. This is one of the things which happens when we speak more quickly: some intermediate sounds are left out. For example, say **Worcestershire sauce** at different speeds from slow to fast and consider the changes in the sounds you issue.

7.4.3 Discussion

This section has outlined how sound input can be transduced into a symbol stream (actually, an expectation stream) and how that stream can, through a multiconfabulation process, be interpreted as a sequence of words being emitted by an attended speaker.

One of the many Achilles’ heels of past speech transcription systems has been the use of a vector quantizer in the sound-processing front end. This is a device that is roughly the same as the sound feature bank described in this section, except that its output is one and only one symbol at each time step (10 ms). This makes it impossible for such systems to deal with multi-source audio scenes, since their “attention” is always focused on whichever feature happens to be responding most strongly at the moment (a response that could be elicited by any one of multiple sound sources, or from superpositions of those sources, from moment to moment).

The sound-processing design described in this section also overcomes the inability of past speech recognition systems to exploit long-range context. Even the best of today’s speech recognizers, operating in a totally noise-free environment with a highly cooperative speaker, cannot achieve much better than 96% sustained accuracy with vocabularies over 60,000 words. This is primarily because of the lack of a way to exploit long-range context from previous words in the current sentence and from previous sentences. In contrast, the system described here has full access to the context-exploitation methods discussed in Sect. 7.3, which can be extended to arbitrarily large bodies of context.

Building a speech recognizer for colloquial speech is much more difficult than for proper language. As is well known, children essentially cannot learn to understand speech unless they can also produce it (in some way). Undoubtedly, this will hold for systems of the type considered in this section. Thus, to solve the whole speech language understanding problem we must also solve the speech language production problem.

In summary, the confabulation theory of vertebrate cognition seems to provide the basis for mechanizing sound cognition in a manner that has the familiar characteristics of human sound cognition.

7.5 Visual Cognition

As with sound, the key challenge of monocular (cycloptic) vision is to usefully transduce incoming image information into symbolic form. Another key part of vision is to build, at higher representation levels, symbolic representations of individual visual objects that are invariant to certain transformations of selected visual attributes such as pose, lighting, color, and form. These are the main topics of this section. Ancillary subjects, such as the highly specialized visual human face recognition system and binocular vision, are not discussed. Readers are expected to have a solid understanding of traditional machine vision.

7.5.1 Building an Eyeball Vision Sensor and Its Gaze Controller

Vertebrate vision is characterized by the use of eyeballs. A *gaze controller* is used to direct the eye(s) to (roughly repeatable) key points on objects of interest. In this section we will consider only monocular, panchromatic visual cognition in detail.

Figure 7.10 illustrates the basic elements of the confabulation-based vision architecture that will be discussed in this section. For simplicity, the subject of how pointing of the video camera sensor will be controlled is ignored. It is assumed that the wide-angle large image camera is fixed and that everything we want to see and visually analyze is within this sensor's fixed visual field of view and is of sufficient size (number of pixels) to make its attributes visible at the sensor's resolution. For example, imagine a wide-angle, high-resolution video camera positioned about 8 feet above the pavement near a busy downtown street intersection, pointed diagonally across the intersection, viewing the people on the sidewalks and the vehicles on the streets.

Assume that the visual sensor (i.e., video camera) gathers digital image frames, each with many millions, or tens of millions, of pixels, at a rate of 30 frames per second. For simplicity, each pixel will be assumed to have its panchromatic (gray scale) brightness measured on a 16-bit linear digital scale.

The *gaze controller* (sometimes also called a *gaze director* or *saliency detector*) of this visual system (see Fig. 7.10) is provided with all of the pixels of each individual frame of imagery. Using this input, it decides whether to select a *fixation point* (a particular pixel of the frame) for that frame (it can select at most one). The manner in which a gaze controller can be built [my laboratory has built one (Hecht-Nielsen and Zhou 1995) and so have a number of others] is described next. To make the discussion which follows concrete, consider a situation where our video camera sensor is monitoring a street scene in a busy downtown area. Each still frame of video contains tens of people and a number of cars driving by.

The basic idea of designing a gaze controller is to mimic human performance. Let an attentive human visual observer watch the output of the video sensor on a computer screen. Attach an eye tracker to the screen to monitor the human's eye movements. These movements will typically be *saccades* – jumps of the eye

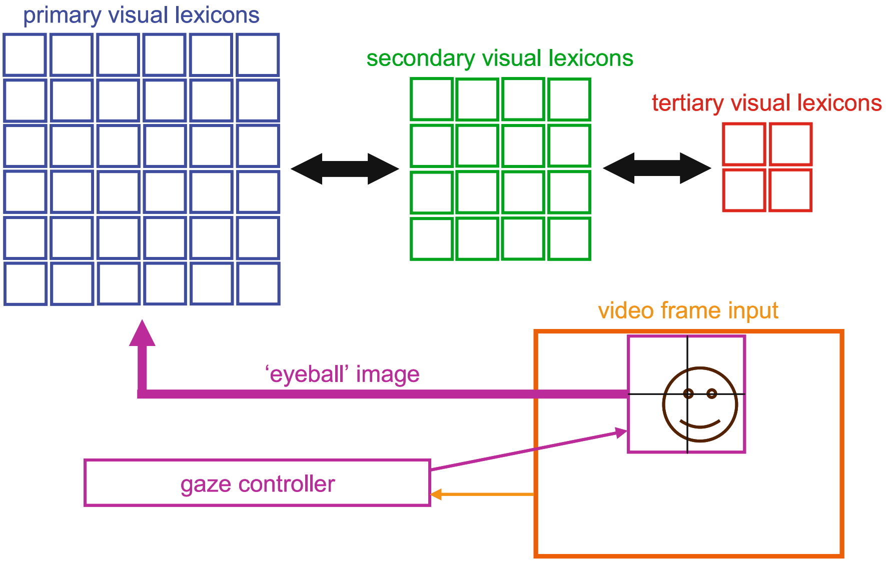


Fig. 7.10. Vision cognition architecture. The raw input to the visual system is a wide-angle high-resolution video camera (large frame shown in the lower right of the figure). A sub-image, of a permanently fixed size (say 1024 × 1024 pixels) of a single video frame (shown as a square within the large frame), termed the *eyeball image*, is determined by the location of its center (depicted by the intersection of crosshairs), known as the *fixation point*. The gaze controller uses the entire large frame to select a single fixation point, if it deems that such a selection is warranted for this large frame (it only attempts to select a fixation point when processing of the last eyeball image has been completed). For simplicity, it is assumed that the video camera is fixed and is able to see the entire visual scene of interest (e.g., a camera viewing a busy downtown intersection). The confabulation architecture used for visual processing is described in the text

position between one *fixation point* and the next. At each fixation point, the human eye gathers image data from a region surrounding the fixation point. This can be viewed as taking a "snapshot" or "eyeball" image centered at that fixation point. The human visual system then processes that eyeball image and jumps to the next fixation point selected by its gaze director (a function which is implemented, in part, by the *superior colliculus* of the brainstem). Visual processing is not carried out during these eyeball jumps. While the human observer is viewing the video it is important that they be carrying out whatever specific task or tasks that the automated vision system will be asked to carry out (e.g., spotting people, pets, bicycles, and cars).

After many tens of hours of video have been viewed by the human observer carrying out the function that the machine visual cognition system will later perform, and their eye movements have been recorded, this provides a record of their fixation point choices for each still frame of specific scene content when that choice was made. This record is then used to train a multilayer perceptron

(Hecht-Nielsen 1989, 2004b) to carry out the gaze control function. The basic idea is simple. Each frame of high-resolution video is described by an *image feature vector* V. This feature vector is produced by first taking the inner product of each of a collection of Gabor logons with the image frame (both considered as vectors of the same dimension). The specific Gabor logons used in forming V (each logon is defined by the constants E, F, and G, and by its position and angle of plane rotation in the image – see Fig. 7.11) are now described.

First, we create a fixed rectangular set of gridpoints located at equal pixel spacings across the entire high-resolution video camera frame (Caid and Hecht-Nielsen 2001, 2004; Hecht-Nielsen and Zhou 1995; Daugman 1985, 1987, 1988a, 1988b; Daugman and Kammen 1987). For example, if each video camera image frame were a 8,192 × 8,192 pixel digital image, with a 16-bit panchromatic grayscale, or, equivalently, a 67,108,864-dimensional floating point real vector with integer components between 0 and 65,535, then we might have gridpoints spaced every 16 pixels vertically and horizontally, with gridpoints on the image edges, for a total of $513 \times 513 = 263{,}169$ gridpoints.

At each gridpoint we create a set of Gabor logons centered at that position, each having a specified rotation angle and E, F, and G values. The set of logons at each gridpoint are exactly the same, save for their translated position. This set, which is now described, is termed a *jet* (von der Malsburg 1990). In the vision work done in my lab we have typically set the ratio E/F to 5/8 and the G/E ratio to $3\pi/2$ for every logon in every jet [these are the values that seem to be used by domestic cats (Hecht-Nielsen 1989)].

Each jet consists, for example, of pairs of a sine logon and a cosine logon at each of seven scales ($E = 2$, 3, 5, 9, 15, 20, and 35 pixel units) and 16 regularly spaced angular orientations, including having the major ellipse axis of one logon pair vertical. Thus, each jet at each gridpoint has 224 logons. Again, each individual logon is viewed as a 67,108,864-dimensional floating point real vector, with each component value given by the evaluation of its formula (Fig. 7.11, properly translated and rotated) evaluated at the pixel location corresponding to that component (obviously, with most of its values at pixels distant from the gridpoint very close to zero). Thus, there are a total of $224 \times 263{,}169 = 58{,}949{,}856$ logons in all of the gridpoint jets; almost as many as there are pixels in the high-resolution camera image.

The image feature vector V of a single camera (assumed to employ a progressive scan) frame is defined to be the 29,474,928-dimensional vector obtained by first calculating the inner product of each logon of each jet with the image vector, and then, to get each component of V, adding the squares of the sine and cosine inner products of the logons of the same scale and rotational orientation in each jet (which reduces the total dimensionality of V to half that of the total number of logons). [Note: Other mathematical transformations are then applied to each of these sums to make their values insensitive to average absolute brightness within a region of the image and insensitive to lighting gradient slopes – but these details go beyond the scope of this sketch and so are left out – see Hecht-Nielsen and Zhou (1995) for examples of such transformations.]

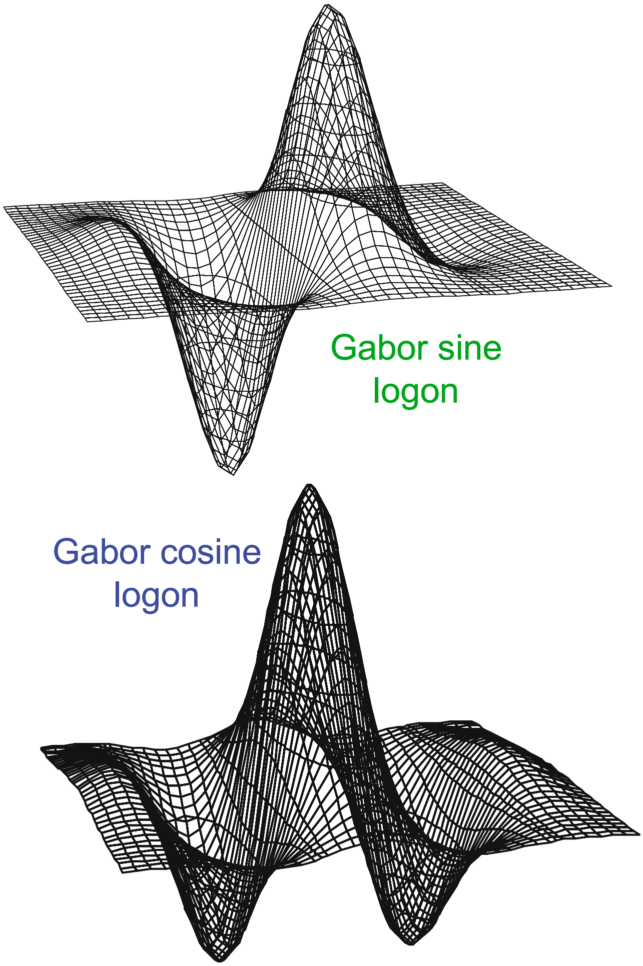


Fig. 7.11. *Gabor logon* local image features. Logons are defined as images with real-valued pixel brightnesses (i.e., both positive and negative values are allowed) defined by geometrical plane rotations and translations (in the image plane) of the canonical two-dimensional functions $\sin(G\,x)/\exp(E\,x^2 + F\,y^2)$ (called a *sine logon*) and $\cos(G\,x)/\exp(E\,x^2 + F\,y^2)$ (termed a *cosine logon*); where E, F, and G are positive constants and x and y are image plane coordinates in the translated and rotated coordinates. Note that E and F define the principal axis lengths of a two-dimensional Gaussian-type ellipsoid and G defines the spatial frequency of a plane grating (with oscillations along the x-axis). The ratio E/F is fixed for all logons used. Each individual logon is considered as a (real-valued) digital image; i.e., as image vectors of the same dimension as the wide-angle video camera frames

Each component of V essentially represents an estimate of the localized spatial frequency content of the camera image (at the position of the associated gridpoint) at the spatial frequency of the involved logon pair, in the direction of oscillation of that pair. It is on the basis of local spatial frequency structure (which V accurately defines) that fixation points are chosen by the gaze controller.

The job of the gaze controller is to learn to mimic the performance of a skilled human observer performing the visual task that is to be mechanized. The manner in which the gaze controller works and the method used to train it are now described.

The gaze controller [a perceptron (Hecht-Nielsen 2004b)] has 224 inputs and two outputs. The inputs represent the components of V corresponding to the jet at a particular image gridpoint (the current *position of regard* of the gaze controller). The outputs of the gaze controller are estimates of the *a posteriori* probability of this gridpoint being chosen by the skilled human as a fixation point along with the *a posteriori* probability of this gridpoint not being chosen by the skilled human as a fixation point. Training of the gaze controller is discussed below; but, to set the stage, the manner in which the gaze controller is used operationally is described first.

Once trained, the gaze controller is used to select a fixation point in a newly acquired video frame by evaluating each of the V component sets from each of the 263,169 gridpoints of the frame. If the first output of the controller is above a fixed threshold (say, 0.8), and the second output is below a fixed threshold (say, 0.2), then that gridpoint is selected as a *candidate fixation point*. If there are no candidate fixation points for the frame, then that frame is skipped. If there are one or more, the one with the highest first output value is selected as the fixation point. The gaze controller also has provisions for creating multiple successive "looks" at the same object during visual training, to facilitate learning of pose insensitivity (see below). In operational use, when a visual object of interest has been fixated on and described, the gaze controller tracks that object's fixation point and prevents return to it until the other visual objects of interest in the scene have been described.

To train the gaze controller, each fixation point example (for which a *reference frame* is selected as the definitive "image input" that the human used – by taking a frame that is a fixed time increment right before the beginning of the saccade) has its pixel coordinates (supplied by the frequently recalibrated eye tracker) stored with its reference frame. Eventually, many thousands of such fixation point–reference frame pairs are produced, randomly scrambled to remove possible content correlations between them, and stored. The V vector for each reference frame is also calculated and stored with it.

The gaze controller perceptron is trained by marching through the fixation point–reference frame examples, in sequence, many times. At each training episode, the next fixation point–reference frame example in sequence is selected and the gridpoint nearest to the fixation point is located. The jet components of the reference frame V vector for that gridpoint are then extracted and provided to the perceptron, along with desired outputs 1 and 0, and one backpropagation training episode using these specified inputs and outputs is carried out. Another gridpoint, distant from any fixation point, is then selected and its jet V components are provided to the perceptron, along with desired outputs 0 and 1, and a second perceptron training episode is carried out using these inputs and outputs. The training process then moves on to the next fixation point–reference image example. Thus, this training procedure beneficially utilizes *oversampling* of the examples of the class of human-supplied fixation points (Hecht-Nielsen 2004).

Training is continued until the perceptron learning curve (as calculated by considering the performance of the perceptron when tested on, say, the last 1000 training examples) reaches a sufficiently low value (say, 80% of the training example pairs would be declared as fixation points and not fixation points, respectively). Final testing is carried out on hundreds of fresh examples not used in testing. If, say, 70% of the final testing examples are classified correctly, then the gaze controller is frozen and ready for service. If not, then additional training is called for. After training, the outputs are scaled to reflect operational class *a priori* probabilities (Hecht-Nielsen 2004b).

It is natural to doubt that the above procedure would produce a functional gaze controller that would mimic the performance of a skilled human. But it can! The reason is probably that the human superior colliculus (and its various input nuclei) is essentially a fixed neuronal machine (at least in autonomous operational mode where no external control is exerted – there are several brain nuclei that can send “commands” to the superior colliculus which override its indigenous decisions) that is not all that “smart” (it operates very fast, in what looks like a “flow through” processing mode). Thus, its natural internal function is capable of being fairly accurately mimicked by a perceptron.

7.5.2 Building the Primary Visual Modules and Knowledge Bases

After the gaze controller has finished its training, it is time to build the rest of the visual system (and link it up with the language module). The first step is to set up the camera and start feeding frames to the gaze controller. Every time it chooses a fixation point (which is, of necessity, a grid point), the V components of the gridpoints lying within the eyeball image centered at that fixation gridpoint are gathered to form the eyeball image description vector (or just *eyeball vector*) U.

Just as in the design of mammalian primary visual cortex, each of the primary visual modules is responsible for monitoring a small local neighborhood of the eyeball image (these neighborhoods are all regularly spaced, they overlap somewhat, and they completely cover the eyeball image). For instance, using the example numbers provided above, each primary visual module (of which, for illustration purposes, Fig. 7.10 shows 36, but there might actually be, say, 81) would monitor the U components from, say, 4,900 gridpoints within and adjacent to its neighborhood of the eyeball image. The vector formed by these selected U components constitutes the *input vector* to that module.

Now comes the tricky part! In order to train the primary visual modules, it is essential that, while this training is underway, the images being gathered by the high-resolution video camera have only <u>one visual object</u> (an object of operational interest) in them; and nothing else. Further, all visual objects that will ever be of interest to the system must be presented in this manner during this training phase. As mentioned in Chap. 8, in humans, this requirement is met by physically altering the characteristics of the baby's eyes after it passes through this stage (during which its vision is limited in range to about arm length). Similarly in other mammals. For artificial visual cognition, a way must be found to meet this critical requirement. For many applications, motion segmentation, and rejecting eyeball images with more than one object fragment in them (as determined by a human educator supervising visual knowledge acquisition), will work.

The symbols of each module are built by collecting input vectors from a huge collection of eyeball images selected by the gaze controller from images gathered in the operational visual environment, but where each eyeball image contains only one object (as described in the previous paragraph). These input vectors for each module are then used to build a VQ codebook for that module (Zador 1963) which is sufficiently large so that, as training progresses, very few input vectors are relatively far (more than the local intra-codebook vector distance) from a codebook vector. Once this criterion is met, the codebook is frozen and one symbol is created for, and uniquely associated with, each codebook vector. This is how the primary visual module symbol sets are developed. As discussed in Chap. 8, it can also be useful (but it is not essential) to develop "complex feature detector" symbols and invoke the precedence principle, as in mammalian primary visual cortex. However, this possibility will be largely ignored here.

Once the primary module symbol sets are developed, the next step is to develop the knowledge bases between these modules. For simplicity, we can assume that every primary visual module is connected to every other by a knowledge base.

The *primary visual layer* (i.e., the primary visual modules and the knowledge bases linking them) knowledge bases are trained using large quantities of new video gathered from the operational source, with the gaze controller selecting fixation points. Again, it is somehow arranged that each eyeball image contains only an object of operational interest at the fixation point and no visual elements of other objects (i.e., the rest of the eyeball image is blank).

As each eyeball image vector U is created and its selected subsidiary components (making up the 81 primary visual module input vectors) are sent to the primary visual modules, each module expresses an expectation with the, say, 10 symbols whose associated codebook vectors lie closest to its input vector. Count accumulation then takes place for all (unidirectional) links between pairs of these expectation symbols lying on d.

The idea of using the 10 closest symbols is based upon the discovery (Caid and Hecht-Nielsen 2001, 2004) that jet correlation vectors which are near to one another in the Euclidean metric (i.e., in the VQ space of a module) represent local visual appearances that are (to a human observer) visually similar to each other; and vice versa. This valuable fact was pointed out in the 1980s by John Daugman (Daugman 1985, 1987, 1988a, 1988b; Daugman and Kammen 1987);

(Daugman also invented the "iris scan" biometric signature). This way, symbols which could reasonably occur together meaningfully within the same object become linked. This is much more efficient and effective than if each module simply expressed the one closest symbol; and yet, because of Daugman's important principle, no harm can come of this expansion to multiple symbols. The key point is that counts are kept between each of the combinatorially-many ordered excited symbol pairs (of symbols on different modules) involved. The process of deriving the $p(\psi|\lambda)$ knowledge link strengths from the counts ensures that only the meaningful links are retained in the end.

As training progresses, the $p(\psi|\lambda)$ knowledge link strengths are periodically calculated from the symbol co-occurrence count matrices (of which there is one for each knowledge base). When the meaningful $p(\psi|\lambda)$ values stop changing much, training is ended. The primary visual layer is now complete.

7.5.3 Building the Secondary and Tertiary Visual Layers

After completion of the primary visual layer, it is time to build the secondary and tertiary visual layers. However, this process requires that the primary visual layer representation of each eyeball image pertain to only one object – which can be accomplished using the primary layer's knowledge bases, as described next.

Figure 7.12 shows a portion of a frame from the wide-angle high-resolution panchromatic video camera containing an eyeball image that has been selected by the gaze controller. Each of the 81 primary visual modules shown is receiving its input vector from this eyeball image. The first thing that happens is that each module expresses an expectation consisting of those (again, say, 10) symbols which were closest to that module's input vector. (Note: This is similar to a **C1F** effect, except that the inputs are not coming from knowledge links, but from "extra-cortical sensory afferents." This illustrates, as does the handling of the S vector by primary sound modules discussed in Sect. 7.4, how the handling of these special external sensory inputs is very similar to the handling of knowledge link inputs.)

Once the primary visual module expectations are established, knowledge links proceeding from the central module of the primary layer, and its immediately neighboring modules, outward are enabled (allowing all symbols of all expectations of those modules to transmit) and the *distal modules* that these links target receive **C1F** commands. Those distal modules that do not receive links to symbols of their (previously established and frozen) expectations describing their portion of the eyeball image have all of their symbols shut down and thereby become null (this follows from the fact, discussed in Sect. 7.1, that the only symbols of a module with a frozen expectation which can receive input excitation from a knowledge link are those which belong to the expectation). In general, the only way that the expectation of an outlying module can have any symbols retained is if one or more of its expectation symbols codes a local appearance that has been meaningfully seen before in conjunction with one or more of those expectations of the modules proximal to the fixation point.

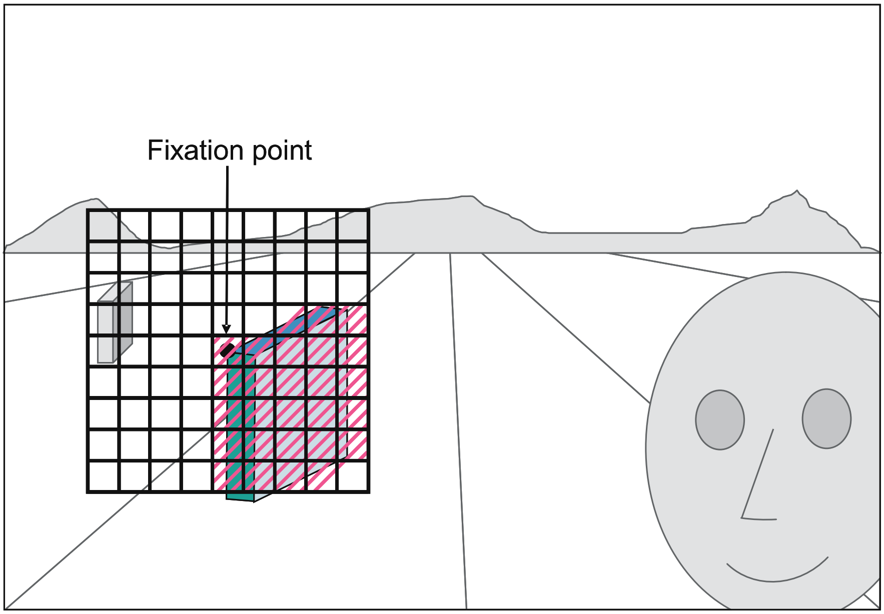


Fig. 7.12. Object segmentation example. A portion of a wide-angle camera frame is shown. The gaze controller has fixated upon the upper left back corner of a rectangular solid (shown here in colour for clarity). The eyeball image is shown surrounding the fixation point with the 81 (overlapping – they actually overlap a bit more than is shown here) fields of view of the 81 primary visual modules shown. As explained in the text, the primary visual layer knowledge bases can be used to eliminate the module responses to all visual objects except the one upon which the fixation point lies

A more elaborate version of this process can also be used, in which a "wave" of confabulations moves outward from the middle of the primary module array to the periphery; with only knowledge bases that span one or two inter-module distances being enabled as the wave progresses. This improves performance because closer-distance-related appearances are more likely to have appeared enough during training to be considered meaningful and be retained.

The astounding thing about this process (which is very fast because all of the distal module confabulations happen in parallel) is that it effectively segments the object upon which the fixation point lies from all the other image content of the eyeball image. In other words, ideally, after this segmentation procedure, which is virtually instantaneous, only symbols describing local appearance of the *attended object* (the one selected by the gaze controller, having the fixation point sitting on it) remain in the expectations of the primary visual layer modules. In Fig. 7.12, these non-null modules (representing the rectangular solid shown) are illustrated as diagonally hatched in magenta. In other words, the only visual appearance data left on the primary visual layer is that describing the attended object, which has thereby effectively been *segmented* and *isolated* from the surrounding objects (as if cut out by scissors).

Note that, given the reasonably long reach of the knowledge bases projecting radially outward from the center of the primary visual layer, even objects which are interrupted by an occluding foreground object will, in principle, have all of their visible components represented by primary modules [and those coding the interrupting object(s) will be nulled]. Also note that the smaller each primary visual module is (in terms of the fraction of the eyeball image it covers), the better this process will work. Thus, segmentation might work even better if we had 625 primary visual modules (a 25×25 array). The use of more "complex" features, based upon more localized "simple" features (see Chap. 8), and the precedence principle is one design approach to achieving some of the benefits of more, smaller modules without actually building them.

Multiple natural questions arise at this point. First, how well does this design actually work in practice? In other words, how thoroughly does this segmentation process null modules coding other objects and how reliably are the modules that code the attended object retained? The short answer is that I don't know. The only evidence I have is based upon experiments done in my lab with a very simple image environment (images of capital Latin alphabetical characters moving about, on a plane, with slowly randomly changing spatial and angular velocities). In this case, a segmentation scheme of this basic type worked very well.

In reality, probably not all fixation point objects will segment cleanly. Sometimes irrelevant modules will not be nulled, and relevant modules will be. However, because of the nature of development process for the secondary and tertiary visual module symbol sets, which is described next, such errors will not matter; as long as these lapses occur randomly and as long as the general quality of the segmentation is fairly good. We will proceed on the assumption that these conditions are satisfied.

The goal for the secondary module symbols is twofold. First, each such symbol should be somewhat *pose insensitive* (i.e., if it responds strongly to an object at one pose it will respond strongly to the same object at nearby poses). Also, each secondary module symbol should represent a larger spatial "chunk" of an object than any primary symbol. Such symbols are said to be more *holistic* than primary module symbols. Tertiary layer module symbols are to be even more holistic than secondary layer symbols.

For secondary and tertiary layer development, sequences of camera images containing the same (operationally relevant) visual object are used. At the beginning of each sequence, we assume that the gaze director has selected a fixation point on the object. In the subsequent frames of the sequence, we check to see that, in each, one point near the initial fixation point is also given a high score by the gaze director. If this is true for a significant sequence of frames (say, 10–20 or more), then these nearby points on the subsequent frames are designated as the fixation points for those frames and this sequence of eyeball images is added to our training set for layers two and three. It is assumed that this set of sequences provides good statistical coverage of the set of all operationally relevant objects, and that each object is seen in many different operationally characteristic poses in the sequences. It is also assumed that the poses of the fixated object in each sequence are dynamically changing. (NOTE: This dynamic variation in pose is needed for training, but is not a requirement for operational use, where objects can be stationary, and yet can still usually be described with a single look).

As shown in Fig. 7.10, the secondary layer modules receive knowledge links from primary layer modules. The arrangement of these links is that a secondary module symbol can only receive a link from symbols lying on primary modules surrounding the position of the secondary module in the second layer module array (i.e., like the primary module array, the secondary array is envisioned as also representing, with a regular "tiling," the eyeball image content of the attended object, but with each secondary module representing a larger "chunk" of this object than a primary layer module – since the secondary layer has fewer modules than the primary layer). These knowledge links connect every symbol belonging to each primary module within the "field of view" of a secondary module to every symbol of that secondary module. For each such forward knowledge link, a link between the same two symbols in the reverse (secondary to primary) direction is also created. All of these links start out with zero strength.

As mentioned in Chap. 8, not all knowledge bases need to have graded $p(\psi|\lambda)$ strengths. For many purposes in cognition, it is sufficient for knowledge links to simply be present (essentially with strength 1) or absent (with effective "strength" 0). These inter-visual-level knowledge links are of this *binary* character.

During each secondary layer training episode, sequences of, say, four to six consecutive eyeball images of the same fixation point on a dynamically pose-changing object of interest (extracted at random from one of the training set sequences) is used in order. As described above, as each eyeball image in the sequence is entered, the above segmenting process is applied to it – yielding expectations on a subset of the primary layer modules. After the first eyeball image of the first training episode sequence has been so represented, a symbol is formed in each secondary module and that symbol is bidirectionally connected from and to each of the primary layer symbols to which it is connected (by setting the relevant knowledge link strengths to 1). The second eyeball image of the sequence is then entered and segmented. The same secondary module symbols created using the first eyeball image of the sequence are then connected from and to all of the primary symbols of this processed second eyeball image for which connections exist. And so forth for the remaining eyeball images of the sequence used on this first training episode. On subsequent training episodes we proceed in exactly the same manner.

Clearly, one symbol is typically going to be added to each secondary module on each image sequence training trial. We stop training when the vast majority of new secondary symbols turn out to be equivalent to existing symbols – as measured by noting that, of those secondary modules which are receiving knowledge link inputs from primary symbols, each such secondary module already has a symbol that simultaneously receives links from at least one expectation symbol of each non-null primary module from which that secondary module receives a knowledge base. In other words, training is stopped when the vast majority of segmented eyeball images can be well represented by secondary symbols which have already been created.

Once training has been stopped, we then use the same training set again for *consolidating* the symbols. This involves using the primary symbols representing each eyeball image as inputs to the secondary modules to which they connect by strengthened connections. Whenever multiple symbols of a secondary module receive links from one primary symbol of each primary module from which that secondary module receives links, then those secondary module symbols are *merged*. Merging simply means that all of the primary to secondary links that went to each of the symbols being merged now go to the merged symbol (and vice versa for the secondary to primary links). What merging does is combine symbols which represent intersecting pose-space trajectories for the same object; thus increasing the pose insensitivity of the merged symbol.

Once the secondary layer modules are built and merged (and the knowledge bases between the primary and secondary layers frozen), the last step is to train the knowledge bases between the secondary layer modules. This is done by entering single eyeball images from the training set, segmenting and representing each image using the primary layer (as during training), carrying out a **W** on each secondary module, and recording the symbol co-occurrence counts for each secondary layer knowledge base.

When all of this is done, the secondary to tertiary knowledge bases (and their inverses) are built using the same method as described above for the primary to secondary knowledge bases, except that this time, each training episode uses the entire set of eyeball images of each training set sequence. The resulting tertiary module symbols are then merged and the tertiary layer inter-module knowledge bases are built. This completes development of the vision module.

7.5.4 How Is the Visual Module Used?

After all of the modules and knowledge bases of the visual module of Fig. 7.10 are built, the module is ready for use. This sub-section briefly sketches how it can be used.

Given a new frame of imagery in which the gaze controller has found a fixation point, the primary layer of the visual module segments and represents the attended object with expectations, just as during the later phases of training and education. The symbols of the non-null expectations of primary modules then transmit to other primary modules and to secondary layer modules via the established knowledge bases. The other primary layer and the secondary layer modules then create expectations in response to **C1F**s. The secondary visual layer expectation symbols then transmit to other secondary modules without expectations (if any there be) and to tertiary modules, again using the knowledge links established during training, and **C1F**s establish expectations on all relevant modules. Finally, the knowledge links of the third layer are used to transmit from the tertiary expectations to any modules without expectations, followed by a final round of **C1F**s.

The expectations formed by this initial “feedforward” interaction represent all of the symbols that are known (i.e., established by the knowledge) to be compatible with the combinations of the symbols in the primary module expectations. At this point, a multiconfabulation process is launched involving all non-nulled modules on all layers and all knowledge bases linking those modules. This multiconfabulation process hones all the expectations until each of the involved secondary and tertiary modules has at most one symbol left (which is, of necessity, active). This collection of symbols is the vision module’s representation of the attended visual object.

This tertiary visual object representation has three important properties. First, it has significant pose insensitivity. With high probability, if you changed the pose of the object somewhat, almost the same set of symbols would be obtained as the object’s representation.

Second, the object has been *completed*, meaning that the representation has removed the effects of occluding objects that blocked the view of some portions of the object (of course, the visible portions of the object must be sufficient for completion by this method).

Third, the representation of the object at the lower levels contains details. For example, if the object is a truck being viewed from the front, the front grille and headlamps will typically be visible and will be represented at the primary level, whereas the representation of the object at the tertiary level will not have these details. It will be more abstract (many more specific truck images would invoke this same, or a very similar, representation).

7.5.5 Linking the Visual Module with the Language Module

Once the visual module is built, what good is it? By itself, not much. It only becomes useful when it is linked by knowledge with other cognitive modules. This sub-section presents a brief sketch of an example of how, via instruction by a human educator, a vision module could be usefully linked with a language architecture.

A problem that has been widely considered is the automated text annotation of video describing objects within video scenes and some of those object’s attributes. For example, such annotations might be useful for blind people if the images being annotated were taken by a camera mounted on a pair of glasses (and the annotations were synthesized into speech provided by the glasses to the wearer’s ears via small tubes issuing from the temples of the glasses near the ears).

Figure 7.13 illustrates a simple concept for such a text annotation system. Video input from the eyeglasses-mounted camera are operated upon by the gaze controller and objects that it selects are segmented and represented by the already-developed visual module, as described in the previous sub-section. The objects that were used in the visual module development process were those that a blind person would want to be informed of (curbs, roads, cars, people, etc.). Thus, by virtue of its development, the visual module will search each new frame of video for an object of operational interest (because these were the objects

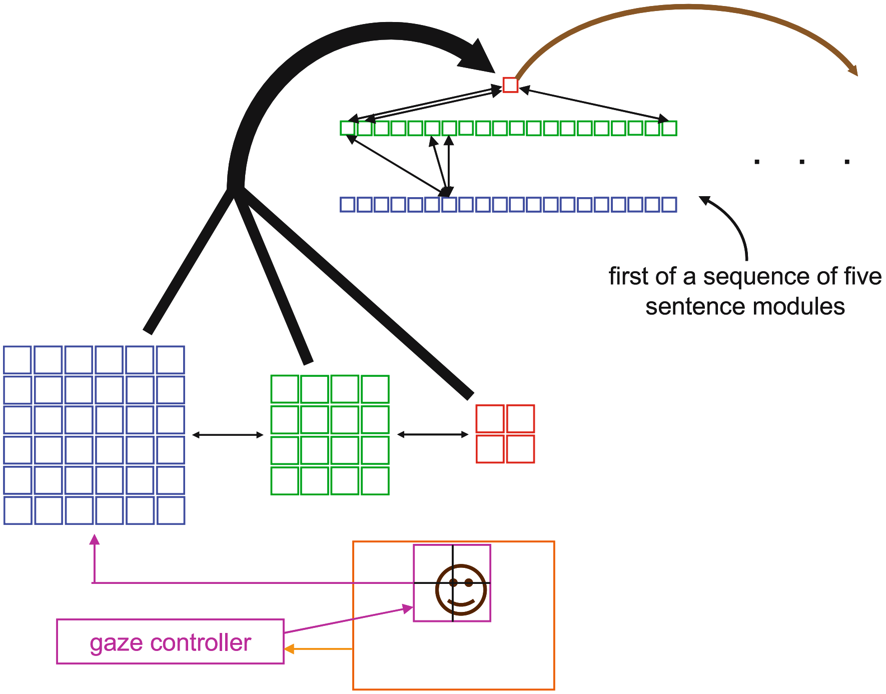


Fig. 7.13. Image text annotation. A simple example of linking a visual architecture with a (text) language architecture. See text for description

sought out by the human educator whose examples were used to train the gaze controller perceptron) and then that object will be segmented and, after multiconfabulation, represented by the architecture on all of its three layers.

To build the knowledge links from the visual architecture to the text architecture, another human educator is used. This educator looks at each fixation point object selected by the vision architecture (while it is being used out on the street in an operationally realistic manner), and, if this is indeed an object that would be of interest to a blind person, enters a few sentences describing that object. These sentences are designed to convey to the blind person useful information about the nature of the object and its visual attributes (information that can be extracted by the human educator just by looking at the visual representation of the object).

To train the links from the vision architecture to the language architecture (every visual module is afforded a knowledge base to every phrase module), the educator's sentences are entered, in order, into the word modules of the sentence architecture (each of which represents one sentence – see Fig. 7.13); each sentence is parsed into phrases (see Sect. 7.4); and these phrases are represented on the sentence summary module of each sentence. Counts are accumulated between the symbols active on the visual architecture's tertiary modules and those active on the summary modules. If the educator wishes to describe specific visual sub-components of the object, they may designate a local window in the eyeball image for each sub-component and supply the sentence(s) describing each such sub-component. The secondary and tertiary module symbols representing the sub-components within each image are then linked to the summary modules of the associated sentences. Before being used in this application, all of the internal knowledge bases of the language architecture have already been trained using a huge text training corpus.

After a sufficient number of education examples have been accumulated (as determined by final performance – described below), the link use counts are converted into $p(\psi|\lambda)$ probabilities and frozen. The knowledge bases from the visual architecture's modules to all of the sentence summary modules are then combined (so that the available long-range context can be exploited by a sentence in any position in the sequence of sentences to be generated). The annotation system is now ready for testing.

The testing phase is carried out by having a sighted evaluator walk down the street wearing the system (yes, the idea is that the entire system is in the form of a pair of glasses!). As the visual module selects and describes each object, knowledge link inputs are sent to the language module. These inputs are used, much as in the example of Sect. 7.3: as context that drives formation of complete sentences. Using multiconfabulation, the language architecture composes one or more grammatical sentences that describe the object and its attributes (see Chap. 2 and the DVD video presentation for examples of whole-sentence generation).

The number of sentences is determined by a *meaning content critic subsystem* (not shown in Fig. 7.13) which stops sentence generation when all of the distinctive, excited, sentence summary module symbols have been "used" in one or more of the generated sentences.

This sketch illustrates the monkey-see/monkey-do principle of cognition: there is never any complicated algorithm or software. No deeply principled system of rules or mathematical constraints. Just confabulation and multiconfabulation. It is a lot like that famous cartoon where scientists are working at a blackboard, attempting, unsuccessfully, to connect up a set of facts on the left with a desired conclusion on the right, via a complicated scientific argument spanning the gap between them. In frustration, one of the scientists erases a band in the middle of the argument and puts in a box (equipped with input and output arrows) labeled "And Then a Miracle Occurs." That is the nature of cognition.

7.6 Discussion

This chapter has reviewed a "unified theory of cognition" which purports to explain all aspects of this vast subject with one type of knowledge and one information-processing operation. The hope is that this discussion has convinced you that this approach to cognition is worthy of more extensive investigation.

Only after language, sound, and vision systems such as those described here have been built, and widely evaluated and criticized, will a sense begin to emerge that the mechanization of cognition is truly possible. I am hopeful that the arguments and discussion presented here are sufficiently compelling to make such a research program sensible.

8 Confabulation Neuroscience II[9]

8.1 Introduction

This chapter sketches the author's confabulation theory of animal cognition. The discussion is focused on the biological implementation of cognition in human cerebral cortex and thalamus (hereinafter often referred to jointly as thalamocortex).

The enormous diversity of animal life, currently ranging in size from single cells (the smallest animals which have ever lived) to blue whales (the largest), and ranging in adaptation across a huge range of biomes, obfuscates its unity. All animal cells function using very similar basic biochemical mechanisms. These mechanisms were developed once and have been genetically conserved across essentially all species. Mentation is similar. The basic mechanism of cognition is, in the view of this theory, the same across all vertebrates (and possibly invertebrates, such as octopi and bees, as well).

The term cognition, as used in this chapter, is not meant to encompass all aspects of mentation. It is restricted to (roughly) those functions carried out by the human cerebral cortex and thalamus. Cognition is a big part of mentation for certain vertebrate species (primates, cats, dogs, parrots, ravens, etc.), but only a minor part for others (fish, reptiles, etc.). Frog cognition exists, but is a minor part of frog mentation. In humans, cognition is the part of mentation of which we are, generally, most proud, and most want to imitate in machines.

An important concept in defining cognition is to consider function; not detailed physiology. In humans, the enormous expansion of cerebral cortex and thalamus has allowed a marked segregation of cognitive function to those organs. Birds can exhibit impressive cognitive functions (Pepperberg 1999; Weir et al. 2002). However, unlike the case in humans, these cognitive functions are probably not entirely confined to a single, neatly delimited, laminar brain nucleus. Even so, confabulation theory hypothesizes that the underlying mathematics of cognition is exactly the same in all vertebrate species (and probably in invertebrates), even though the neuronal implementation varies considerably.[10]

[9] This chapter is based on the original publication Hecht-Nielsen R (2006) The mechanization of cognition. In: Bar-Cohen Y (ed) Biomimetics. CRC Press, Boca Raton, FL, pp 57–128, and reprinted from the original with kind permission of the publisher, CRC Press.

[10] This is much as in electronics: the same logic circuit can be implemented with electromechanical relays, in silicon CMOS circuits, using vacuum tubes, or even using fluidics devices. While these implementations are physically dissimilar, their functions are mathematically identical.

Although not enough is known to create a definitive list of specific human cognitive functions, the following items would certainly be on such a list:

- Language
- Hearing
- Seeing
- Somatosensation
- Action (movement process and thought process) origination

This chapter focuses upon the implementation of cognitive knowledge links, confabulation, and action command origination by the human cerebral cortex and thalamus. It is assumed that the reader is familiar with the concepts, terminology, and mathematics of elementary confabulation (e.g., as discussed in Chaps. 3 and 4) and with elementary human neuroanatomy and neurophysiology [e.g., as presented in (Mai et al. 2004; Mountcastle 1998; Nicholls et al. 2001; Nolte 1999; Paxinos and Mai 2004; Steward 2000)]. The theory hypothesizes that all human cognitive functions, including those listed above, are implemented using the basic confabulation machinery sketched in this chapter. To keep this chapter focused, the manner in which confabulation can be used to carry out specific cognitive functions (such as those listed above) will not be discussed here, as this is essentially the material covered in Chap. 7.

To keep the size of this chapter reasonable, and to avoid speculations about fine details, the treatment will avoid extensive discussion at the level of individual neurons, synapses, and axonal signals. For example, only simplified gross or summary aspects of interneuronal signaling processes and neurodynamic processes will be discussed. Yet the theory contends that the fine details jibe with these slightly larger-scale functional descriptions. *Multiconfabulation* (dynamically interacting confabulations taking place contemporaneously in multiple modules), which the theory hypothesizes is the dominant mode of use of confabulation in human cognition (see Chaps. 3, 5, and 6), will only be briefly mentioned, as a detailed treatment would go beyond the introductory scope of this sketch of the theory's biological implementation. At the current time, the theory presented in this chapter is the only existing detailed explanation of the operation of cerebral cortex/thalamus, and of human cognition.

8.2 Summary of the Theory

The fundamental hypotheses of the theory are summarized in this section. Subsequent sections elaborate (see also Chaps. 3 and 5).

All information-processing involved in human cognition is hypothesized to be carried out by many thousands of separate thalamocortical modules, each consisting of a particular small localized patch of cortex (possibly consisting of disjoint, non-adjacent, sub-patches) and a particular, uniquely paired, small localized zone of thalamus which are reciprocally connected axonally. These *thalamocortical modules* (of which human thalamocortex has thousands) are hypothesized to each implement a list of (typically thousands of) discrete symbols (which is stable over time, and can be added to) and to carry out a single symbolic information-processing operation called confabulation. Each symbol is represented by a specific (bipartite) collection of neurons within the module. These collections are all about the same size within any single module; but this size varies considerably, from tens to hundreds of neurons per symbol – a genetically determined value, between modules located in different parts of the cortex. Any pair of such neuron collections of the same module, representing two different symbols, typically have a few neurons in common. Each neuron which participates in such a collection typically participates in many others as well.

Each module is hypothesized to be controlled by a single graded (i.e., analog-valued) excitatory control input, exactly in analogy with an individual muscle (each muscle has a single graded excitatory control input that, by its value across time, specifies the muscle's contraction force history). The theory hypothesizes that properly phased and timed sequences of such thought-control inputs to each member of an ensemble of cortical modules causes them to carry out a thought process. These thought processes are "data-independent," much like computer operations such as numerical addition and Boolean XOR. It is hypothesized that vast numbers of such thought processes (and movement processes) are learned by rehearsal training and stored in a hierarchical organization within knowledge bases between modules of cerebral cortex.

Confabulation is implemented in parallel by the neurons of a module and is often completed in a few tens of milliseconds. This is a "winners-take-all" style of dynamical parallel competitive interaction between symbols that does not require a "referee" or "controller" to be in charge. The states of the involved neurons evolve dynamically and autonomously during confabulation via the massively parallel mutual interactions of the involved neurons. The state of each involved neuron is either *excited* or *active* (a small minority of neurons), or almost completely *inactive* (the vast majority). The term *active* (implying a momentary, maximally communicating state) is deliberately undefined as it involves neuronal signaling details which are not yet known, as does the term *excited* (implying a highly, but not maximally, communicating state).

If the outcome of a confabulation is a single symbol, the neurons representing that symbol will automatically be made active and all other neurons of the module are *inactive* (not communicating). However, if multiple symbols result from a confabulation (the outcome is dependent upon multiple factors, including the time profile of the module control signal – see below), these will be at different levels of excitation (but not active) and all other symbols will be inactive. Those few neurons which end up in the excited or active state represent the symbol(s) which "won" the confabulation competition. These symbols are termed the *conclusions* of that confabulation operation. Confabulations frequently end with no excited or active neurons – a conclusion termed the *null symbol* – which signifies that no viable conclusion was reached. This ability to decide that "I don't know" is one of the great strengths of cognition.

The theory hypothesizes that the only knowledge stored and used for cognition within thalamocortex takes the form of (indirect, parallel) unidirectional axonal connections between the population of neurons within one module used to represent one symbol and neurons used to represent a symbol in another module. Each such *link* between a pair of symbols is termed an *item of knowledge*. The average human is hypothesized to possess billions of such items of knowledge.

The theory hypothesizes that items of knowledge are immediately established on a temporary basis when a novel, meaningful, co-occurrence of symbol pair activity occurs during a period of wakefulness (assuming that those symbols are equipped with the necessary axonal paths with which a link can be established). If this *short-term memory* link is selected for deliberate rehearsal during the next period of sleep, it gets promoted to the status of a *medium-term memory*. If this medium-term memory link is revisited on near-term subsequent sleep periods it then gets promoted to a *long-term memory*, which will typically last as long as the involved tissue remains patent and not re-deployed.

It is hypothesized that each time a thalamocortical module carries out a confabulation which concludes with the expression of a single active symbol (as opposed to no symbols or multiple excited symbols), an *action command* associated with that symbol is immediately issued by a specialized cortical component of that module (this is the theory's *conclusion → action principle*). Action commands cause muscle and thalamocortical thought module control signals to be sent. In other words, every time a thought process successfully reaches a single conclusion, a new movement process and/or thought process is launched (some of which may undergo additional evaluation before being finally *executed*). This is the theory's explanation for the origin of all non-autonomic animal behavior.

As with almost all cognitive functions, actions are organized into a hierarchy; where individual symbols belonging to higher-level modules typically each represent a time-ordered sequence of multiple lower-level symbols.

Evolution has seen to it that symbols which, when expressed alone, launch action commands which could conflict with one another (e.g., carrying out a throwing motion at the same time as trying to answer the telephone) are grouped together and collected into the same module (usually at a high level in the action hierarchy). That way, when one such *action symbol* wins a confabulation (and has its associated lower-level action commands launched), the others are silent – thereby automatically *deconflicting* all actions. This is why all aspects of animal behavior are so remarkably *focused* in character. Each complement of our moving and thinking "hardware" is, by this mechanism, automatically restricted to doing one thing at a time. *Dithering* [rapidly switching from one decisive action (behavioral program) to another, and then back again] illustrates this perfectly.

The thought processes at the lowest level of the action hierarchy are typically carried out unconditionally at high speed. If single symbol states result from confabulations which take place as part of a thought process, these symbols then decide which actions will be carried out next (this happens both by the action commands the expression of these symbols launch, and by the influence of these symbols – acting through knowledge links – on the outcomes of subsequent confabulations, for which these symbols act as assumed facts). Similarly for movements, as ongoing movements bring about changes in the winning symbols in confabulations in somatosensory cortex – which then alter the selections of the next action symbols in modules in motor and pre-motor cortex. This ongoing, high-speed, dynamic contingent control of movement and thought helps account for the astounding reliability, and comprehensive, moment-by-moment adaptability, of animal action.

All of cognition is built from the above-discussed elements: modules, knowledge bases, and the action commands associated with the individual symbols of each module. The following sections of this chapter discuss more details of how these elements are implemented in the human brain. See Chaps. 3 and 5 for some citations of past research that influenced this theory's development.

8.3 Implementation of Modules

Figure 8.1 illustrates the physiology of thalamocortical modules. In reality, these modules are not entirely disjoint, nor entirely functionally independent, from their physically neighboring modules. However, as a first approximation, they can be treated as such, which is the view which will be adopted here.

Figure 8.2 shows more details of the functional character of an individual human module. The cortical patch of the module probably uses certain neurons in layers II, III and IVa to represent the symbols of the module. (Note: layer IVa is also sometimes identified as part of layer III in the neuroscience literature – thanks to Soren Solari for this observation.) Each symbol (of which there are typically thousands) is represented by a roughly equal number of neurons, ranging from tens to hundreds (this number deliberately varies, by genetic command, with the position of the cortical patch of the module on the surface of cortex). The union of the cortical patches of all modules is the entire cortex, whereas the union of the thalamic zones of all modules constitutes only a portion of thalamus.

Symbol-representing neurons of the module's cortical patch can send signals to the glomeruli of the paired thalamic zone via neurons of layer VI of the patch (as illustrated on the left side of Fig. 8.2). These downward connections each synapse with a few neurons of the thalamic reticular nucleus (NRT) and with a few glomeruli. The NRT neurons themselves (which are inhibitory) send axons to a few glomeruli. The right side of Fig. 8.2 illustrates the connections back to the cortical patch from the thalamic zone glomeruli (each of which also synapses with a few neurons of the NRT). These axons synapse primarily with neurons in layers III and IVa – some of which are, presumably, members of the *symbol-representing neuron* population. As mentioned above, no attempt to discuss the details of this module design will be made, as these details are not yet adequately established and, anyway, are irrelevant for this introductory sketch. Instead,

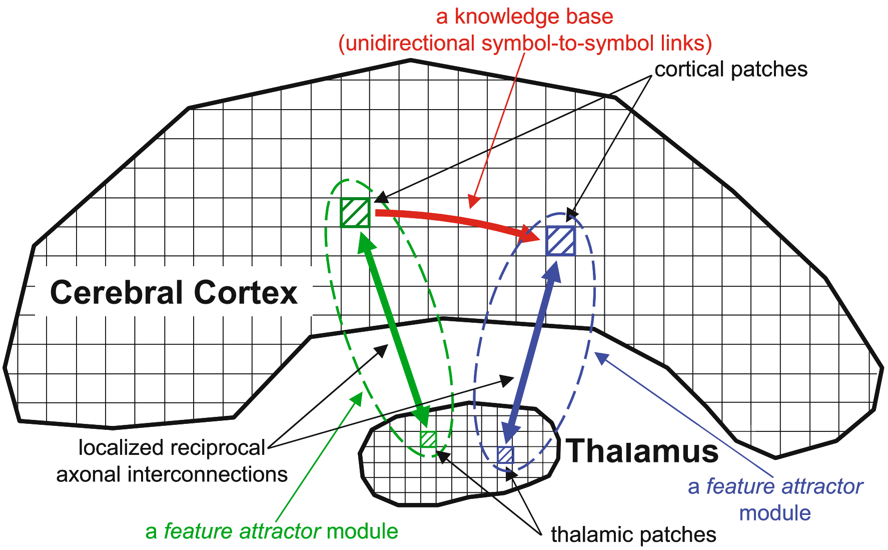


Fig. 8.1. Thalamocortical modules. All cognitive information-processing is carried out by distinct, modular, thalamocortical circuits termed *neuronal attractor networks*, of which two are shown here. Each module (of which human cortex has many thousands) consists of a small localized patch of cortex (which may be comprised of disjoint, physically separated, sub-patches), a small localized zone of thalamus, and the reciprocal axonal connections linking the two. Each module implements a large stable set of attractive states called *symbols*, each represented by a specific collection of neurons (all such collections within a module are of approximately the same size). Neuron overlap between each pair of symbols is small, and each neuron involved in representing one symbol typically participates in representing many symbols. One *item of knowledge* is a (parallel, two-stage synfire) set of unidirectional axonal connections collectively forming a *link* between the neurons representing one symbol within one module (e.g., the green one shown here) and neurons representing one symbol on a second module (e.g., the blue one shown here). The collection of all such links between the symbols of one module (here the green one), termed the *source module*, and that of a second (here the blue one), termed the *target module*, are termed a *knowledge base* (here represented by a red arrow spanning the cortical portions of the green and blue modules)

a discussion is now presented of a simple mathematical model of an attractor network to illustrate the hypothesized dynamical behavior of a thalamocortical model in response to proper knowledge link and operation command inputs.

The theory hypothesizes that each thalamocortical module carries out a single information-processing operation: confabulation. This occurs whenever appropriate knowledge link inputs and the operation command input arrive at the module at the same time. The total time required for the module to carry out one confabulation operation is roughly 100 ms. Ensembles of mutually interacting confabulations (instances of *multiconfabulation* – see Chaps. 6 and 7) can often be highly overlapped in time. By this means, the "total processing time"

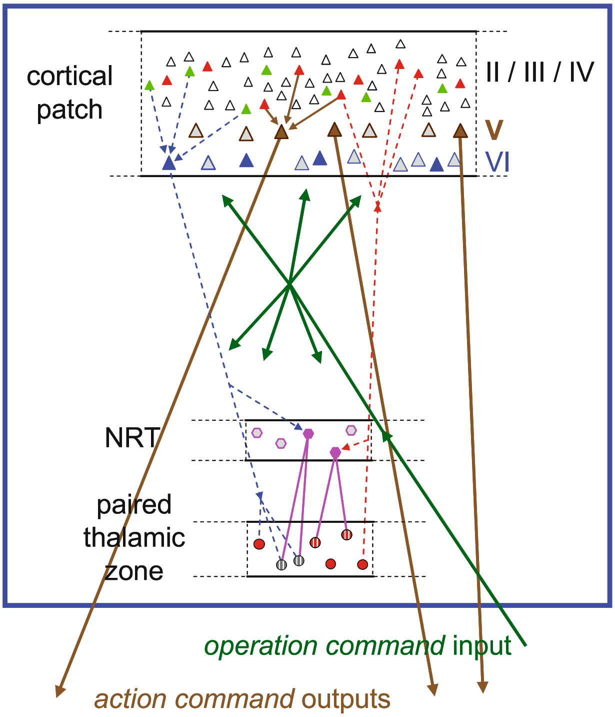


Fig. 8.2. A single thalamocortical module: side view. The module consists of a full-depth patch of cortex (possibly comprised of multiple separate full-depth disjoint sub-patches – not illustrated here), as well as a paired zone of thalamus. The green and red neurons in cortical layers II / III / IV illustrate the two collections of neurons representing two symbols of the module (common neurons shared by the two collections are not shown, nor are the axons involved in the neuronal attractor network function used to implement confabulation). The complete pool of neurons within the module used to represent symbols contains many tens, or even hundreds, of thousands of neurons. Each symbol-representing neuron collection has tens to hundreds of neurons in it. Axons from cortical layer VI to NRT (thalamic reticular nucleus) and thalamus are shown in dashed blue. Axons from thalamic glomeruli to NRT and cortical layers III / IVa are shown in dashed red. Axons from NRT neurons to glomeruli are shown in pink. An axon of the operation command input, which affects a large subset of the neurons of the module, and which arrives from an external sub-cortical nucleus, is shown in green. The theory only specifies the overall information-processing function of each cortical module (implementation of the list of symbols, confabulation, and origination/termination of knowledge links). Details of module operation at the cellular level are not known

exhibited by such an ensemble of confabulations can be astoundingly short – often a small multiple of the involved axonal and synaptic delays involved, and not much longer than a small number of individual confabulations. This accounts for the almost impossibly short "reaction times" often seen in various psychological tests.

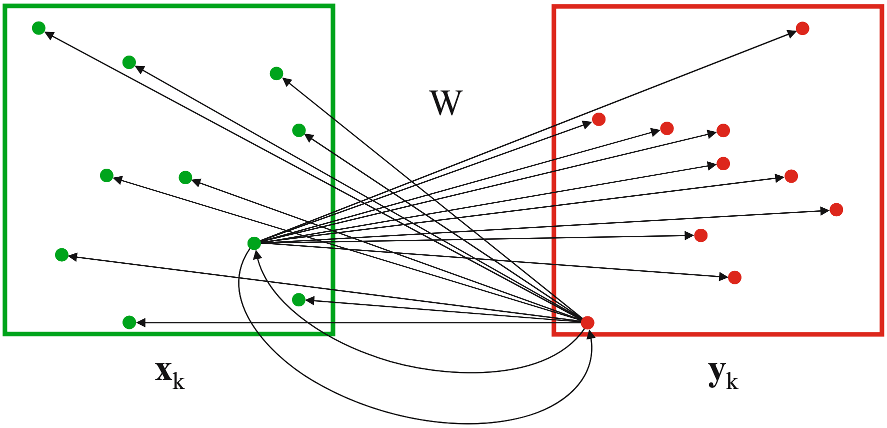


Fig. 8.3. Simple attractor network example. The left, **x**, neural field has N neurons, as does the right, **y**, neural field. One Willshaw stable state pair, $\mathbf{x}_k$ and $\mathbf{y}_k$, is shown here (actually, each $\mathbf{x}_k$ and $\mathbf{y}_k$ typically has many tens of neurons – e.g., Np = 60 for the parameter set described in the text – of which only 10 are shown here). Each neuron of each state sends connections to all of the neurons of the other (only the connections from one neuron in $\mathbf{x}_k$ and one neuron in $\mathbf{y}_k$ are shown here). Together, the set of all such connections for all L stable pairs is recorded in the connection matrix W. Notice that these connections are not knowledge links; rather, they are internal connections between $\mathbf{x}_k$ and $\mathbf{y}_k$, the two parts of the neuron population of symbol k within a single module. Also, unlike knowledge link connections (which, as discussed in Sect. 8.4, are unidirectional and for which the second stage is typically very sparse), these interpopulation connections must be reciprocal and dense (although they need not be 100% dense – a fact that you can easily establish experimentally with your model)

The mathematical model discussed below illustrates the dynamical process involved in carrying out one confabulation. Keep in mind that this model might represent module neurodynamics between the cortical and thalamic portions of the module, strictly cortical neuron dynamics, or even the overall dynamics of a group of smaller attractor networks [e.g., a localized version of the “network of networks” hypothesis of Sutton and Anderson (Sutton and Anderson 1995; Sutton and Anderson in: Hecht-Nielsen and McKenna 2003)].

In 1969 Willshaw and his colleagues (Willshaw et al. 1969) introduced the “non-holographic” associative memory. This “one-way” device (“retrieval key” represented on one “field” of neurons and “retrieved pattern” on a second), based on Hebbian learning, is a major departure in concept from the previous (linear algebra-based) associative memory concepts (Anderson 1968, 1972; Gabor 1969; Kohonen 1972). The brilliant Willshaw design (an absolutely essential step towards the theory presented in this chapter) is a generalization of the pioneering Steinbuch *learnmatrix* (Steinbuch 1961a, 1961b, 1963, 1965; Steinbuch and Piske 1963; Steinbuch and Widrow 1965; Widrow et al. 2005), although Willshaw and his colleagues did not seem to be aware of this earlier development. For efficiency, it is assumed that the reader is familiar with the Willshaw network and its theory (Amari 1989; Kosko 1988; Palm 1980; Sommer and Palm 1989). A related, centrally important, idea is the "Brain State in a Box" attractor network architecture of Anderson (Anderson et al. 1977; and Chaps. 3 and 5).

In 1987 I conceived a hybrid of the Willshaw network and the Amari–Hopfield "energy function" attractor network (Amari 1974; Amit 1989; Hopfield 1982, 1984). In effect, this hybrid network was two reciprocally connected Willshaw networks; however, it also had an energy function. Karen Haines and I theoretically investigated the dynamics of this network (Haines and Hecht-Nielsen 1988) (in 1988 computer exploration of the dynamics of such networks, at scales sufficiently large to explore their utility for information-processing, was not feasible). We were able to show theoretically that this hybrid had four important (and unique) characteristics. First, it would, with very high probability, converge to one of the Willshaw stable states. Second, it would converge in a finite number of steps. Third, there were no "spurious" stable states. Fourth, it could carry out a "winner take all" kind of information-processing. This hybrid network could thus serve as the functional implementation of (in the parlance of this chapter) some easy cases of confabulation. However (see Chaps. 3 and 5), its convergence capabilities turned out to be too limited to make it a general solution. This was the first result on the trail to the theory presented here. It took another 16 years to discover that, by having antecedent support knowledge links deliver excitation to symbols (i.e., stable states) of such a module, this simple one-winner-takes-all information-processing operation (*confabulation*) is sufficient to carry out all of cognition.

By 1992 it had become possible to carry out computer simulations of reciprocal Willshaw networks of interesting size. This immediately led to the rather startling discovery that, even without an energy function (i.e., carrying out neuron updating on a completely local basis, as in Willshaw's original work), even significantly "damaged" (the parlance at that stage of discovery) starting states (Willshaw stable states with a significant fraction of added and deleted neurons) would almost always converge in one "round-trip" or "out-and-back cycle." This made it likely that this is the functional design of cortical module circuits.

As this work progressed, it became clear that large networks of this type were even more robust and would converge in one cycle even from a small incomplete fragment of a Willshaw stable state. It was also at this point that the issue of "threshold control" (Willshaw's original neurons all had the same fixed "firing" threshold – equal to the number of neurons in each stable state) came to the fore. If such networks were operated by a threshold control signal that rose monotonically from a minimum level, it could automatically carry out a global "most excited neurons win" competition without need for communication between the neurons. The subset of neurons which becomes active first then inhibits others from becoming so (at least in modules in the brain; but not in these simple mathematical models, which typically lack inhibition). From this came the idea that each module must be actively controlled by a graded command signal, much like an individual muscle. This eventually led to the realization that the control of movement and the control of thought are implemented in essentially the same manner, using the same cortical and sub-cortical structures (indeed, the theory postulates that there are many combined movement and thought processes which are represented as unitized symbols at higher levels in the action hierarchy – e.g., a back dive action routine in which visual perception must feed corrections to the movement control in order to enter the water vertically).

To see what attractor networks of this unusual type are all about, the reader is invited to pause in their reading and build (e.g., using C, LabVIEW, MATLAB, etc.) a simple working example using the following prescription. If you accept this invitation, you will see first-hand the amazing capabilities of these networks (which will help you appreciate and accept the theory). While simple, this network possesses many of the important behavioral characteristics of the hypothesized design of biological modules.

We will use two N-dimensional real column vectors, **x** and **y**, to represent the states of N neurons in each of two "neural fields." For good results, N should be at least 10,000 (even better results are obtained for N above 30,000). Using a good random number generator, create L pairs of **x** and **y** vectors $\{(\mathbf{x}_1,\mathbf{y}_1), (\mathbf{x}_2,\mathbf{y}_2), \ldots, (\mathbf{x}_L,\mathbf{y}_L)\}$ with each $\mathbf{x}_i$ vector and each $\mathbf{y}_i$ vector having binary (0 and 1) entries selected independently at random; where the probability of each component being 1 is p. Use, for example, $p = 0.003$ and $L = 5{,}000$ for $N = 20{,}000$. As you will see, these $\mathbf{x}_i$ and $\mathbf{y}_i$ pairs turn out to be *stable states* of the network. Each $\mathbf{x}_k$ and $\mathbf{y}_k$ vector pair, $k = 1, 2, \ldots, L$ *represents* one of the L *symbols* of the network. For simplicity, we will concentrate on the $\mathbf{x}_k$ vector as the representation of symbol k. Thus, each symbol is represented by a collection of about Np "active" neurons. The random selection of the symbol neuron sets and the deliberate processes of neuronal interconnection between the sets correspond to the development and refinement processes in each thalamocortical module that are described later in this section.

During development of the bipartite stable states $\{(\mathbf{x}_1,\mathbf{y}_1), (\mathbf{x}_2,\mathbf{y}_2), \ldots, (\mathbf{x}_L,\mathbf{y}_L)\}$ (which happens gradually over time in biology, but all at once in this simple model), connections between the neurons of the **x** and **y** fields are also established. These connections are very simple: each neuron of $\mathbf{x}_k$ (i.e., the neurons of the **x** field whose indices within $\mathbf{x}_k$ have a 1 assigned to them) sends a connection to each neuron of $\mathbf{y}_k$, and vice versa. This yields a connection matrix W given by:

$$W = U\left(\sum_{i=1}^{N} \mathbf{y}_k \mathbf{x}_k^T \right) , \tag{8.1}$$

where the matrix function U sets every positive component of a matrix to 1 and every other component to zero. Given these simple constructions, you are now ready to experiment with your network.

First, choose one of the $\mathbf{x}_k$ vectors and modify it. For example, eliminate a few neurons (by converting entries that are 1 to 0s) or add a few neurons (by converting 0s to 1s). Let this modified $\mathbf{x}_k$ vector be called **u**. Now, "run" the network using **u** as the initial **x** field state. To do this, first calculate the *input excitation* I_j of each **y** field neuron j using the formula $\mathbf{I} = W\mathbf{u}$.; where **I** is the column vector containing the input excitation values I_j, $j = 1, 2, \dots, N$. In effect, each active neuron of the **x** field (i.e., those neurons whose indices have a 1 entry in **u**) sends output to neurons of the **y** field to which it has connections (as determined by W). Each neuron j of the **y** field sums up the number of connections it has received from active **x** field neurons (the ones designated by the 1 entries in **u**) and this is I_j.

After the I_j values have been calculated, those neurons of the **y** field which have the largest I_j values (or very close to the largest – say within 3 or 4 – this is a parameter you can experiment with) are made *active*. As mentioned above, this procedure is a simple, but roughly equivalent, surrogate for active centralized control of the network. Code the set of active **y** field neurons using the vector **v** (which has a 1 in the index of each active **y** field neuron and zeros everywhere else). Then calculate the input intensity vector $W^T v$ for the **x** field (this is the "reverse transmission" phase of the operation of the network) and again make active those neurons with largest, or near-largest, values of input intensity. This completes one cycle of operation of the network. Astoundingly, the state of the **x** field of the network will be very close to $\mathbf{x}_k$, the vector used as the dominant base for the construction of **u** (as long as the number of modifications made to $\mathbf{x}_k$ when forming **u** was not too large).

Now expand your experiments by letting each **u** be equal to one of the **x** field stable states $\mathbf{x}_k$ with many (say half) of its neurons made inactive plus the union of many (say, 1–10) small fragments (say, 3–8 neurons each) of other stable **x** field vectors, along with a small number (say, 5–10) of active "noise" (randomly selected) neurons (see Fig. 8.4). Now, when operated, the network will converge rapidly (again, often in one cycle) to the $\mathbf{x}_k$ symbol whose fragment was the largest. When you do your experiments you will see that this works even if that largest fragment contains only a third of the neurons in the original $\mathbf{x}_k$.

Again, notice that to achieve the "neurons with the largest, or near-largest, input excitation win" information-processing effect, all that is needed is to have an excitatory operation control input to the network which uniformly raises all of the involved neurons' excitation levels (towards a constant fixed "firing" threshold that each neuron uses) at the same time. By ramping up this input, eventually a group of neurons will "fire," and these will be exactly those with the largest or near-largest input intensity. Localized mutual inhibition between cortical neurons (which is known to exist, but is not included in the above simplified model) then sees to it that there are no additional winners, even if the control input keeps rising. Note also that the rate of rise of the control signal can control the width of the band of input excitations (below maximum) for which neurons are allowed to win the competition: a fast rate allows more neurons (with slightly less input intensity than the first winners) to become active before inhibition has time to kick in. A slow rate of rise restricts the winners to just one symbol. Finally, the operation control input to the network can be limited to be less than some deliberately chosen maximum value, which will leave no symbols active if the sum of the all neuron's input excitation, plus the control signal, are below the fixed threshold level. Thus, an attractor network confabulation can yield a null conclusion when there are no sufficiently strong answers. Sect. 7.1 of the previous chapter discusses some of these information-processing *effects*; which can be achieved by judicious control of a module's operation command input signal.

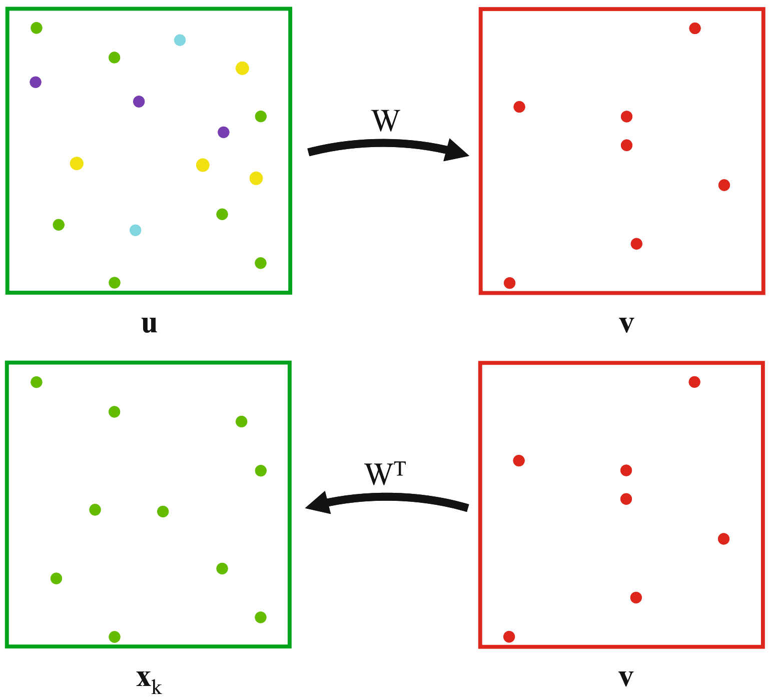


Fig. 8.4. Weak symbol convergence property of the simple attractor network example. The initial state (top portion) of the **x** neural field is a vector **u** consisting of a large portion (say, half of its neurons) of one particular $\mathbf{x}_k$ (the neurons of this $\mathbf{x}_k$ are shown in green), along with small subsets of neurons of many other **x** field stable states. The network is then operated in the **x** to **y** direction (top diagram). Each neuron of **u** sends output to those neurons of the **y** field to which it is connected (as determined by the connection matrix W). The **y** field neurons which receive the most, or close to the most, connections from active neurons of **u** are then made active. These active neurons are represented by the vector **v**. The network is then operated in the **y** to **x** direction (bottom diagram), where the **x** field neurons receiving the most, or close to the most, connections from active neurons of **v** are made active. The astounding thing is that this set of active **x** field neurons is typically very close to $\mathbf{x}_k$, the dominant component of the initial **u** input. Yet all of the processing is completely local and parallel. As will be seen below, this is all that is needed to carry out confabulation. In thalamocortical modules this entire cycle of operation (which is controlled by a rising operation command input supplied to all of the involved neurons of the module) is probably often completed in roughly 100 ms. The hypothesis of the theory is that some sort of *attractor network* behavior of this kind implements *confabulation* in human thalamocortical modules – the universal information-processing operation of cognition

An important difference between the behavior of this simple attractor network model and that of thalamocortical modules is that, by involving inhibition (and some other design improvements), the biological attractor network can somehow successfully deal with situations where even hundreds of stable **x** field vector fragments (as opposed to only a few in the simple attractor network) can be suppressed to yield a fully expressed dominant fragment $\mathbf{x}_k$.

The development process of thalamocortical modules is hypothesized by the theory to take place in steps (which are usually completed in childhood; although under some conditions adults may be able to develop new modules).

Each module's set of symbols is used to describe one *attribute* of objects in the mental universe. Symbol development starts as soon as meaningful (i.e., not random) inputs to the module start arriving. For "lower level" attributes, this *self-organization* process sometimes starts before birth. For "higher level" attributes (modules), the necessary inputs do not arrive (and module organization does not start) until after the requisite lower-level modules have organized and started producing assumed fact outputs.

The hypothesized process by which a module is developed is now sketched. At the beginning of development, a sizable subset of the neurons of cortical layers II, III, and IV of the module happen by chance to preferentially receive extra-modular inputs and are stimulated repeatedly by these inputs. These neurons develop, through various mutually competitive and cooperative interactions, responses which collectively cover the range of signal ensembles the module's input channels are providing. In effect, each such feature detector neuron is simultaneously driven to respond strongly to one of the input signal ensembles it happens to repeatedly receive, while at the same time, through competition between feature detector neurons within the module, it is discouraged from becoming tuned to the same ensemble of inputs as other feature detector neurons of that module. This is the classic insight that arose originally in connection with the mathematical concepts of *vector quantization* (*VQ*) and *k-means*. These competitive and cooperative VQ feature set development ideas have been extensively studied in various forms by many researchers from the 1960s through today [e.g., see (Grossberg 1976; Carpenter and Grossberg 1991; Kohonen 1984, 1995; Nilsson 1965, 1998; Tsypkin 1973; Zador 1963)]. The net result of this first stage of attractor network circuit development is a large set of feature detector neurons (which, after this brief initial plastic period, become largely frozen in their responses – unless severe trauma later in life causes recapitulation of this early development phase) that have responses with moderate local redundancy and high input range coverage (i.e., low information loss). These might be called the *simple* feature detector neurons.

Once the simple feature detector neurons of a module have been formed and frozen, additional *secondary* (or "complex") feature detector neurons within the module then organize. These are neurons which just happen (the wiring of cortex is locally random and is essentially formed first, during early organization and learning, and then is soon frozen for life) to receive most of their input from nearby simple feature detector neurons (as opposed to primarily from extra-modular inputs, as with the simple feature detector neurons themselves).

In certain areas of cortex (e.g., primary visual cortex) secondary feature detector neurons can receive inputs from primary feature detector neurons "belonging" to other nearby modules. This is an example of why it is not correct to say that modules are disjoint and non-interacting (which, nonetheless, is exactly how we will treat them here).

Just as with the primary neurons, the secondary feature detector neurons also self-organize along the lines of a VQ codebook – except that this codebook sits to some degree "on top" of the simple cell codebook. The net result is that secondary feature neurons tend to learn statistically common <u>combinations</u> of multiple co-excited simple feature detector neurons; again, with only modest redundancy and with little information loss.

A new key principle postulated by the theory relative to these populations of feature detector neurons, is that secondary (and tertiary – see below) feature detector neurons also develop inhibitory connections (via growth of axons of properly interposed inhibitory interneurons that receive input from the secondary feature detector neurons) that target the simple feature detector neurons which feed them. Thus, when a secondary feature detector neuron becomes highly excited (partly) by simple feature detector neuron inputs, it then immediately shuts off these simple neurons. This is the theory's *precedence principle*. In effect, it causes groups of inputs that are statistically "coherent" to be re-represented as a whole ensemble, rather than as a collection of "unassembled" pieces. For example, in a visual input, an ensemble of simple feature detector neurons together representing a straight line segment might be re-represented by some secondary feature detector neurons which together represent the whole segment. Once activated by these primary neurons, these secondary neurons then, by the precedence principle, immediately shut off (via learned connections to local inhibitory interneurons) the primary neurons that caused their activation.

Once the secondary feature detectors of a module have stabilized they too are then frozen and (at least in certain areas of cortex) tertiary feature detectors (often coding even larger complexes of statistically meaningful inputs) form their codebook. They too obey the precedence principle. For example, in primary visual cortical modules, there are probably tertiary feature detectors which code long line segments (probably both curved and straight) spanning multiple modules. Again, this is one example of how nearby modules might interact – such tertiary feature detectors might well inhibit and shut off lower-level feature detector neurons in other nearby modules. Of course, other inhibitory interactions also develop, such as the line "end stopping" that inhibits reactions of line continuation feature detectors beyond its end. In essence, the interactions within cortex during the short time span of its reaction to external input (20–40 ms) are envisioned by this theory as similar to the "competitive and cooperative neural field interactions" postulated by Stephen Grossberg and Gail Carpenter and their colleagues in their visual processing theories (Carpenter and Grossberg 1991; Grossberg 1976, 1987; Grossberg et al. 1997), without their concept's problem of impossibly slow "field" interactions. When external input (along with an operate command) is provided to a developed module, the above brief interactions ensue and then a single symbol (or a small set of symbols, depending upon the manner in which the operate command to the module is manipulated) representing that input is expressed. The process by which the symbols are developed from the feature detector neuron responses is now briefly discussed.

Once the feature detector neurons (of all orders) have had their responses frozen, the next step is to consider the sets of feature detector neurons which become highly excited together across the cortical module due to external inputs. Because the input wiring of the feature detector neurons is random and sparse, the feature detector neurons function somewhat like VQ codebook vectors with many of their components randomly zeroed out (i.e., like ordinary VQ codebook vectors projected into randomly selected low-dimensional sub-spaces defined by the relatively sparse random axonal wiring feeding the feature detector neurons of the module). In general, under these circumstances, it can be established that any input to the module (again, whether from thalamus, from other cortical modules, or from other extracortical sources) will cause a roughly equal number of feature detector neurons to become highly excited. This is easy to see for an ordinary VQ codebook. Imagine a probability density function in a high-dimensional input space (the raw input to the module). The feature detector responses can be represented as points spread out in a roughly equiprobable manner within this data cloud (at least before projection into their low-dimensional sub-spaces), (Kohonen 1995). Thus, given any specific input, we can choose to highly excite a roughly uniform number of highest appropriate precedence feature detector points that are closest to that input vector.

In effect, if we imagine a rising externally supplied operation control signal (effectively supplied to all of the feature detector neurons that have not been shut down by the precedence principle) as the sum of the control signal and each neuron's excitation level (due to the external inputs) climbs, the most highly excited neurons will cross their fixed thresholds first and "fire" (there are many more details than this, but this general idea is hypothesized to be correct). If the rate of rise of the operate signal is constant, a roughly fixed number of not-inhibited feature detector neurons will begin "firing" before local inhibition from these "early winners" prevents any more winners from arising. This leaves a fixed set of active neurons of roughly a fixed size. The theory presumes that such fixed sets will, by means of their co-activity, and the mutually excitatory connections that develop between them, tend to become established and stabilized as the module's internal attractor network circuit connections gradually form and stabilize. Each such neuron group, as adjusted and stabilized as an attractor state of the module over many such trials, becomes one of the symbols in the module.

Each final symbol can be viewed as being a localized "cloud" in the VQ external input representation space composed of a uniform number of close-by co-active feature detector responses (imagine a VQ where there is not one winning vector, but many). Together, these clouds cover the entire portion of the space in which external inputs are seen. Portions of the VQ space with higher input vector probability density values automatically have smaller clouds. Portions with lower density have larger clouds. Yet each cloud is represented by roughly the same number of vectors (neurons). These clouds are the symbols. In effect, the symbols form a Voronoi-like partitioning of the occupied portion of the external input representation space (Kohonen 1984, 1995); except that the symbol cloud partitions are not disjoint, but overlap somewhat.

Information theorists have not spent much time considering the notion of having a cloud of "winning vectors" (i.e., what this theory would term a *symbol*) as the outcome of the operation of a vector quantizer. The idea has always been to only allow the single VQ codebook vector that is closest to the "input" win (that deviations from this tradition have not been extensively studied is an example of the stifling influence of Shannon worship, in this instance his tendentious source coding theory, on the information theory community). From a theoretical perspective, the reason clouds of points are needed in the brain is that the connections which define the "input" to the module (whether they be sensory inputs arriving via thalamus, knowledge links arriving from other portions of cortex, or yet other inputs) only connect (randomly) to a sparse sampling of each symbol's neurons. As mentioned above, this causes the feature detector neurons' vectors to essentially lie in relatively low-dimensional random subspaces of the VQ codebook space. Thus, to comprehensively characterize the input (i.e., to avoid significant information loss) a number of such "individually incomplete," but mutually complementary, feature representations are needed. So, only a cloud will do. Of course, the beauty of a cloud is that this is exactly what the stable states of a thalamocortical module must be, in order to achieve the necessary confabulation "winner-takes-all" dynamics.

A subtle point the theory makes is that the organization of a module is dependent upon which input data source is available first. This first-available source (whether from sensory inputs supplied through thalamus or active symbol inputs from other modules) drives development of the symbols. Once development has finished, the symbols are largely frozen (although they sometimes can change later due to symbol disuse and new symbols can be added in response to persistent changes in the input information environment). Since almost all aspects of cognition are hierarchical, once a module is frozen, other modules begin using its assumed fact outputs to drive their development. So, in general, development is a one-shot process (which illustrates the importance of getting it right the first time in childhood). Once the symbols have been frozen, the only synaptic modifications which occur are those connected with knowledge acquisition, which is the topic discussed next.

8.4 Implementation of Knowledge

As discussed in the previous chapters, all of the knowledge used in cognition (e.g., for vision, hearing, somatosensation, contemplation, thinking, and moving)

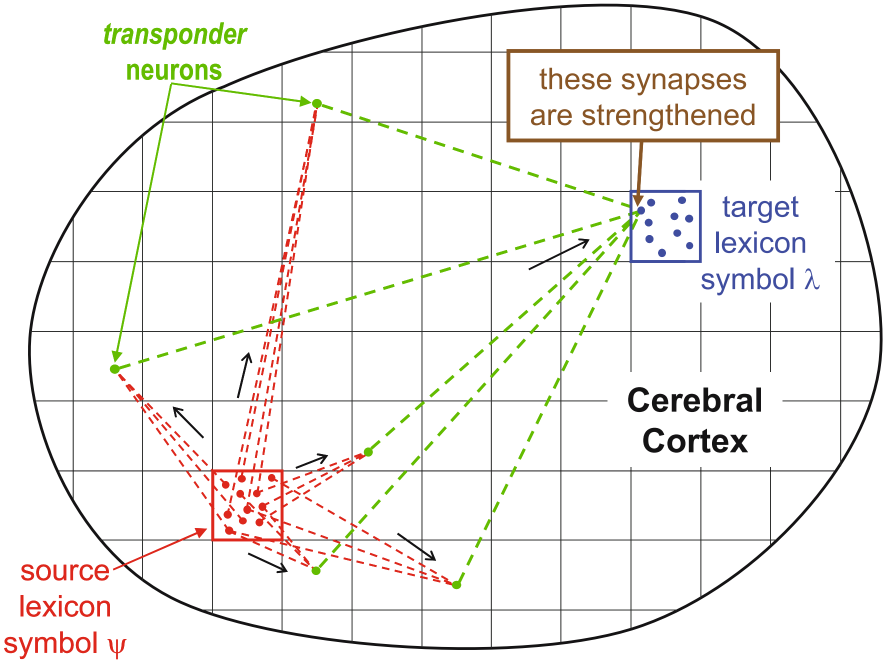


Fig. 8.5. A single knowledge link in the human cerebral cortex. See text for discussion

takes the form of unidirectional weighted links between pairs of symbols (typically, but not necessarily, symbols residing within different modules). This section sketches how these links are hypothesized to be implemented in human cortex (all knowledge links used in human cognition reside entirely within the gray and white matter of cortex).

Figure 8.5 considers a single knowledge link from symbol ψ in a particular cortical *source* module (module) to symbol λ in a particular *target* or *answer* module. The set of all knowledge links from symbols of one particular source module to symbols of one particular target module are called a *knowledge base*. The single knowledge link considered in Fig. 8.5 belongs to the knowledge base linking the particular source module shown to the particular target module shown.

When the neurons of Fig. 8.5 representing symbol ψ are active (or highly excited if multiple symbols are being expressed, but this case will be ignored here), these *ψ neurons* send their action potential outputs to millions of neurons residing in cortical modules to which the neurons of this source module sends axons (the gross statistics of this axon distribution pattern are determined genetically, but the local details are random). Each such active symbol-representing neuron sends action potential signals via its axon collaterals to tens of thousands of neurons. Of the millions of neurons which receive these signals from the ψ neurons, a few thousand receive not just one such axon collateral, but many. These are termed *transponder* neurons. They are strongly excited by this simultaneous input from the ψ neurons, causing them to send strong output to all of the neurons to which they in turn send axons. In effect, the first step of the link transmission starts with the tens to hundreds of active neurons representing symbol ψ and ends with many thousands of excited transponder neurons, which also (collectively) uniquely represent the symbol ψ. In effect, transponder neurons momentarily *amplify* the size of the ψ symbol representation. It is hypothesized by the theory that this *synfire chain* (Abeles 1991) of activation does not propagate further because only active (or highly excited) neurons can launch such a process, and while the transponder neurons are *excited*, they are not active or highly excited (i.e., active, or highly excited, neurons – a rare state that can only exist during and following a confabulation information-processing operation – are the only ones that can unconditionally excite other neurons). However, as with transponder neurons, if a neuron receives a high-enough number of simultaneous inputs from active neurons – even through unstrengthened synapses, and in the absence of any operation command input – it will become excited. Finally, excited neurons *can* excite other neurons if those other neurons reside in a module which is simultaneously also receiving operation command signal input (this is what happens when knowledge is used and when short-term memory learning takes place, as will be discussed below).

The wiring of the cortical knowledge axons is (largely) completed in childhood and then remains (at least for our purposes here) essentially fixed for life. Again, the gross statistics of this wiring are genetically determined, but the local details are random.

A relatively small number (say, 1–25% – a genetically controlled percentage that deliberately varies across cortex) of the target module neurons representing symbol λ will just happen to each receive many synaptic inputs from a subset of the transponder neurons (Fig. 8.5 illustrates the axonal connections from ψ transponder neurons for only one of these few λ neurons). These particular λ neurons *complete* the knowledge link. If all of the neurons representing symbol λ are already active at the moment these synaptic inputs arrive, then (in the event that they have not been previously permanently strengthened) the transponder neuron synapses that land on this subset of them will be temporarily strengthened (this is called *short-term memory*). During the next sleep period, if this causal pairing of symbols ψ and λ is again deliberately rehearsed, these temporarily strengthened synapses may be more lastingly strengthened (this is *medium-term memory*). If this link is subsequently rehearsed more over the next few days, these synapses may be permanently strengthened (this is *long-term memory*). It is important to note that the synapses from the ψ neurons to the ψ transponder neurons are generally not strengthened. This is because the transponder neurons are not meaningfully active at the time when these inputs arrive. Only deliberate usage of a link with immediately prior co-occurrence of both source symbol and target symbol activity causes learning. This was, roughly, the learning hypothesis that Donald Hebb advanced 58 years ago (Hebb 1949).

Note again that the transponder neurons that represent a symbol ψ will always be the same, independent of which target module(s) are to be linked to. Thus, ψ transponder neurons must send a sufficiently large number of axons to all of the modules containing symbols to which symbol ψ might need to connect. The theory posits that genetic control of the distribution of axons (nominally) ensures that all of the potentially necessary knowledge links can be formed. Obviously, this postulated design could be analyzed, since the rough anatomy and statistics of cortical axon fascicles are known. Such an analysis might well be able to support this hypothesis, or raise doubts that it is capable of explaining cortical knowledge.

Cognitive functions where confabulations always yield zero or one candidate conclusions (because at most one symbol has anything close to enough knowledge links from the assumed facts) do not need precisely weighted knowledge links. In cortical modules which only require such confabulations, knowledge links terminating within that module are hypothesized by the theory to be essentially binary in strength: either completely *unstrengthened* (i.e., as yet unused) or *strong* (strengthened to near maximum). Such modules together probably encompass a majority of cortex.

However, other cognitive functions (e.g., language) do require each knowledge link to have a strength that is directly related by some fixed function to $p(\psi|\lambda)$. The theory's hypothesis as to how these weightings arise is now sketched.

Although the mechanisms of synaptic modification are not yet well understood (particularly those connected with medium-term and long-term memory), research has established that "Hebbian" synaptic strengthening does occur (Cowan et al. 2001). This presumably can yield a transponder neuron to target symbol neuron synapse strength directly related to the joint probability $p(\psi\lambda)$ (i.e., roughly, the probability of the two involved symbols being co-active). In addition, studies of post-synaptic neurotransmitter depolarization transduction response (i.e., within the neuron receiving the synaptic neurotransmitter output, separate from the transmitting synapse itself) by Marder and her colleagues (Marder and Prinz 2002, 2003) and by Turrigiano and her colleagues (Desai et al. 2002; Turrigiano and Nelson 2000, 2004; Turrigiano et al. 1998) suggests that the post-synaptic apparatus of an excitatory cortical synapse (e.g., one landing on a target symbol neuron) is independently modifiable in efficacy, in multiplicative series with this Hebbian $p(\psi\lambda)$ efficacy. This "post-synaptic signalling efficacy" is expressed as a neurotransmitter *receptivity* proportional to a direct function of the reciprocal of that target neuron's average firing rate; which is essentially $p(\lambda)$. The net result is implementation by this *Marder–Turrigiano–Hebb* learning process (as I call it) of an overall link strength directly related to $p(\psi\lambda)/p(\lambda)$, which, by Bayes' law, is $p(\psi|\lambda)$. Presumably, somehow, this overall graded link strength is implemented in the final long-term memory. Thus, it is plausible that biological learning processes at the neuron level can accumulate the knowledge needed for confabulation. See Chap. 3 for more discussion of knowledge link implementation.

8.5 Implementation of Confabulation

Since only a small subset of the neurons representing target module symbol λ are excited by a knowledge link from source module symbol ψ, how can confabulation be implemented? This section, which presents the theory's hypothesized implementation of confabulation, answers this question and shows that these "sparse" knowledge links are an essential element of cortical design. Counterintuitively, if these links were "fully connected," cortex could not function.

Figure 8.6 schematically illustrates how confabulation is implemented in a thalamocortical (answer) module. The four boxes on the left are four cortical modules, each having exactly one assumed fact symbol active (symbols α, β, γ, and δ respectively). Each of these active symbols is represented by the full complement of the neurons which represent it, which are all active (illustrated as a complete row of filled circles within that assumed fact symbol's module, depicted in the figure in colors green, red, blue, and brown for α, β, γ, and δ respectively). As will be seen below, this is how the symbol(s) which are the conclusions of a confabulation operation are biologically expressed (namely, all of their representing neurons are active and all other symbol-representing neurons are inactive).

In Fig. 8.6 the neurons representing each symbol of a module are shown as separated into their own rows. Of course, in the actual tissue, the neurons of

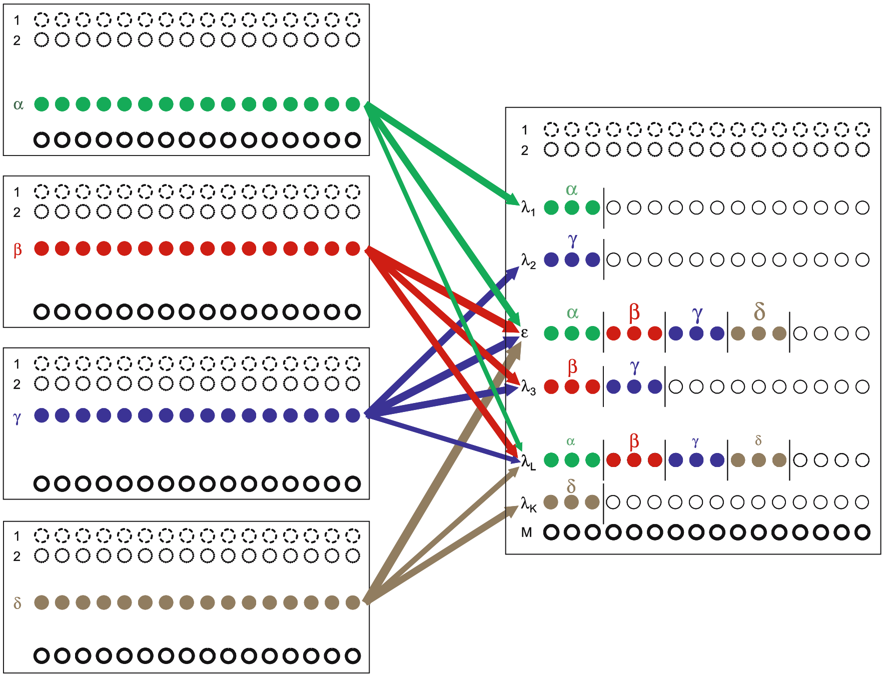


Fig. 8.6. The implementation of confabulation in human cerebral cortex. See text for explanation

each symbol are scattered randomly within the relevant layers of the cortical portion of the module implementing the module. But for clarity, in Fig. 8.6 each symbol's neurons are shown collected together into one row. The fact that the same neuron appears in multiple rows (each symbol-representing neuron typically participates in representing many different symbols) is ignored here, as this small pairwise *overlap* between symbol representations causes no significant interference between symbols.

The answer module for the elementary confabulation we are going to carry out (based upon assumed facts α, β, γ, and δ, just as described in Chaps. 3 and 4) is shown as the box on the right in Fig. 8.6. Each assumed fact symbol has knowledge links to multiple symbols of the answer module, as illustrated by the colored arrows proceeding from each source module to the answer module. The width of each such knowledge link arrow corresponds to the link *strength*, i.e., the value of its $p(\psi|\lambda)$ probability. Each assumed fact symbol in this example (other possibilities exist, but will be ignored here) is assumed to be the sole conclusion of a previous confabulation on its module. Thus symbols α, β, γ, and δ are all active (maximally transmissive).

The symbols of the answer module which receive one or more links from the assumed facts are denoted by ε, λ_1, λ_2, λ_3, and so forth and, for clarity, are grouped in Fig. 8.6. As discussed in the previous section, the actual percentage of neurons of each target symbol which receive synaptic inputs from the assumed fact's transponder neurons is approximately the same for all symbols (this is a function of the roughly uniform – at least for each individual answer module – binomial statistics of the locally random cortico-cortical axons implementing each knowledge link). And, as mentioned earlier, this percentage is low (from 1% to 25%, depending on where the module is located in cortex).

As shown in Fig. 8.6, symbol λ_1 receives only one link (it is a medium-strength link from assumed fact symbol α). In accordance with Fig. 8.5, only a fraction of the neurons of the answer module which represent symbol λ_1 are actually being excited by this input link. These are shown as green circles with α above them (again, for clarity, the target symbol neurons which happen to receive input excitation from a particular assumed fact, which are actually randomly located, are grouped together on the left in each row, and labeled above with the symbol of that assumed fact). Note that, in the case of this group of green neurons of symbol λ_1 receiving input from assumed fact symbol α, that a medium-sized font α is shown above the group, reflecting the fact that the knowledge link delivering this assumed fact excitation has only medium strength $p(\lambda_1|\alpha)$. Similarly, the neurons representing symbol λ_2 are also receiving only one medium-strength link; namely, from assumed fact symbol γ.

Only two of the answer module symbols shown in Fig. 8.6, namely ε and λ_L are receiving links from all four assumed facts. However, note that the links impinging on the neurons of symbol ε are stronger than those impinging on symbol λ_L. Now this discussion of the biological implementation of confabulation will pause momentarily for a discussion of synapses.

Despite over a century of study, synapse function is still only poorly understood. What is now clear is that synapses have dynamic behavior, both in terms of their responses to incoming action potentials and in terms of modifications to their transmission efficacy (over a wide range of time scales). For example, some synapses seem to have transmission efficacy which "droops" or "fades" on successive action potentials in a rapid sequence (such are sometimes termed *depressing* synapses – this has nothing to do with the clinical condition of depression). Other synapses (termed *facilitating*) increase their efficacies over such a sequence; and yet others exhibit no change. However, it has been learned that even these categorizations are too simplistic and do not convey a true picture of what is going on. That clear picture awaits a day when the actual modulations used for information transmission, and the "zoo" of functionally distinct neurons and synapses, are better understood. Perhaps this theory can speed the advent of that day by providing a comprehensive vision of overall cortical function, which can serve as a framework for formulating scientific questions.

Even though little is known about synapses, it is clear that many synapses are weak (unstrengthened), quite likely unreliable, and marginally capable of signaling (confabulation theory claims that over 99% of synapses must be in this category; see Sect. 8.7 below). This is why it takes a pool of highly excited or active neurons representing a symbol (such neurons possess the ultimate in neural signaling power) to excite transponder neurons (each of which receives many inputs from the pool). No lesser neural collection is capable of doing this through unstrengthened synapses (which is why cortical synfire chains have only two stages). However, it is also known that some synapses (this theory claims that these represent fewer than 1% of the total of cortical excitatory synapses, see Sect. 8.7) are much stronger. These stronger synapses (which the theory claims are the seat of storage of all cortical knowledge) are physically larger than unstrengthened synapses and are often chained together into multiple-synapse groups that operate together (see Fig. 8.7). One estimate (Henry Markram, personal communication) is that such a strengthened synapse group can be perhaps 60 times stronger than the common unstrengthened synapse (in terms of the total depolarizing effect of the multi-synapse on the target cell at which they squirt glutamate neurotransmitter). These strong synapses are probably also much more reliable. Figure 8.7 illustrates these two hypothesized types of cortical excitatory synapses used for cognitive knowledge storage.

The theory hypothesizes that synapses which implement knowledge links (as in Fig. 8.5) are always strengthened greatly in comparison with unstrengthened synapses. When the knowledge link requires that a transponder-neuron-to-target-symbol-neuron synapse code the graded probability $p(\psi|\lambda)$ (as opposed to just a binary "unstrengthened" or "strong"), the dynamic range of such a strengthened synapse is probably no more than a factor of, say, 6. In other words, if the weakest strengthened synapse has an "efficacy" 10 times that of an unstrengthened synapse, the strongest possible synapse will have an efficacy of 60. Thus, we must code the smallest meaningful $p(\psi|\lambda)$ value as 10 and the

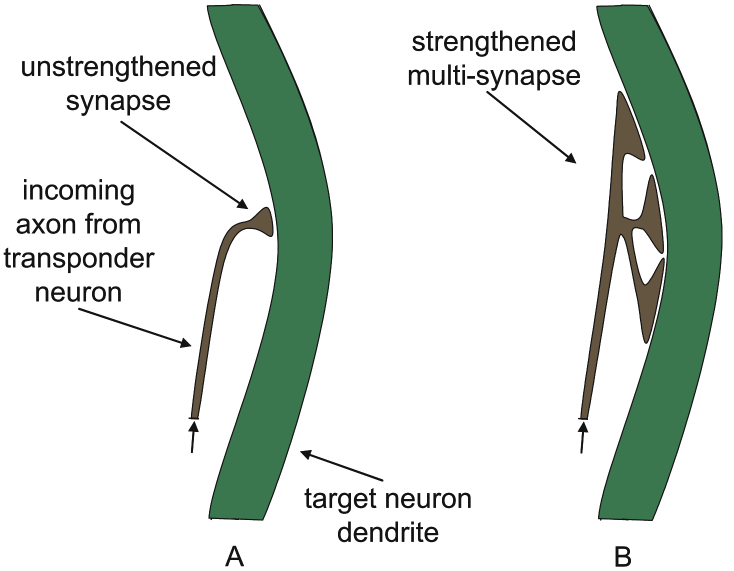

unstrengthened synapse
incoming axon from transponder neuron
strengthened multi-synapse
target neuron dendrite
A
B

Fig. 8.7. Synapse strengthening – the fundamental storage mechanism of cortical knowledge. **A** A weak, unreliable, unstrengthened *vestigial* synapse making a connection from a transponder neuron axon to a target neuron dendrite. The theory hypothesizes that roughly 99% of human cortical synapses with this connectivity are vestigial. **B** the same synapse after learning (i.e., the progression from short-term memory to medium-term memory to long-term memory has been completed). Now, the synapse has blossomed into three parallel synapses, each physically much larger than the original one. This multi-synapse (perhaps what has been recently termed a *ribbon synapse*) is more reliable and has an efficacy ranging from perhaps 30 to 50 times that of the original unstrengthened synapse (learning always yields a great increase in efficacy – the theory posits that there are no such knowledge storage synapses which are only slightly strengthened)[11]

strongest as 60 (remember that $0 < p_0 < p(\psi|\lambda) \leq 1$, where p_0 is the smallest "meaningful" antecedent support probability value).

In our computer confabulation experiments (e.g., those reported in Chaps. 4 and 6), the smallest meaningful $p(\psi|\lambda)$ value (define this to be a new constant p_0) turns out to be about $p_0 = 0.0001$, and the largest $p(\psi|\lambda)$ value seen is almost 1.0. As it turns out, the smaller $p(\psi|\lambda)$ values need the most representational

[11] This is easy to see: Consider the simplified attractor you built and experimented with above. It always converged to a single pure state $\mathbf{x}_k$ (at least when the initial state $\mathbf{u}$ was dominated by $\mathbf{x}_k$), meaning that all of the neurons which represent $\mathbf{x}_k$ are active and all other neurons are inactive. However, each of the neurons of $\mathbf{x}_k$ also belongs to many other stable states $\mathbf{x}_i$, but this does not cause any problems or interference. You may not have seen this aspect of the system at the time you did your experiments – go check! You will find that even though the overlap between each pair of $\mathbf{x}$ field stable states is relatively small, each individual neuron participates in many such stable states. The properties of this kind of attractor network are quite astounding, and they do not even have many of the additional design features that thalamocortical modules possess.

precision, whereas little error is introduced if the larger $p(\psi|\lambda)$ values are more coarsely represented. Clearly, this is a situation that seems ripe for using logarithms! The theory indeed proposes that non-binary strengthened synapses in human cortex have their $p(\psi|\lambda)$ probabilities coded using a logarithmic scale [i.e., $y = \log_b(cx) = a + \log_b(x)$, where $a = \log_b(c)$]. This not only solves the limited synaptic dynamic range problem mentioned above, but it is also a key part of making confabulation work (as we will see below)!

So, given the above estimates and hypothesis, let us determine the base b of the logarithms used for synaptic knowledge coding in the human cerebral cortex, as well as the constant c [actually, we will instead estimate $a = \log_b(c)$]. To fix ideas, say we want $p(\psi|\lambda) = 0.0001$ to be represented by a synaptic strength of 10, and we want $p(\psi|\lambda) = 1.0$ to be represented by a synaptic strength of 60. In other words, we need to find positive constants a and b such that (see Chaps. 3, 4, and 5):

$$a + \log_b(0.0001) = 10 \tag{8.2}$$

and

$$a + \log_b(1.0) = 60. \tag{8.3}$$

Clearly, from Eq. 8.3, $a = 60$ (since the log of 1 is zero for every b). Then Eq. 8.2 yields $b = 1.2023$. Thus, when a highly excited transponder neuron representing source symbol ψ delivers its signal to a neuron of answer module symbol λ, the signal delivered to that neuron will be proportional to $a + \log_b(p(\psi|\lambda))$ (where the constant of proportionality is postulated to be the same for all target neurons of a single module, and where nearby modules typically have very similar proportionality constants).

You might wonder why the signal delivered is not the "product" of the transponder neuron output signal and the synaptic efficacy [as was common in classical "neural network" models such as the Perceptron (Hecht-Nielsen 2004)]. Well, it is! However, exploring this aspect of the theory would quickly take us beyond the scope of this introductory sketch (see the Methods appendix of Chap. 6 for an example). Since transponder neurons coding a single active symbol (assumed fact) on a module, essentially anywhere in cortex, always fire at the about the same signal level (namely, the maximum possible) when they are implementing a link, we can consider this link input signal as constant. Thus, for the purposes of discussing elementary confabulation (the process of reaching conclusions based upon sets of assumed facts), we need not worry about this issue here. Another issue that can be ignored is the influence of the many non-strengthened synapses impinging on target module symbol neurons. This effect can be ignored because the inputs due to this prolific, but unreliable, source is very uniformly distributed across all neurons of all symbols and so it affects them all equally. In other words, this input acts as a low-variance, roughly constant, uniform "background noise." One possibility (that seems worthy of further investigation) is that this "loud" uniform blanket of carefully designed (by evolution) background noise might be an important part of the neuronal attractor network mechanism modules use to carry out confabulation. See Kosko (2006) and Patel and Kosko (2007) for discussions of how such a noise blanket can dramatically improve neuronal calculational precision and accuracy.

The main conclusion of the above argument is that those neurons that represent answer module symbol λ which happen to receive a sufficient number of ψ transponder neuron inputs to allow them to respond will all have about the same response to that input; namely a response proportional to $a + \log_b(p(\psi|\lambda))$.

Recall from the discussion of Fig. 8.6 above that the number of neurons of each answer module symbol which receive sufficient synaptic inputs from the transponder neurons of a source symbol ψ are about the same for each knowledge link and each symbol. You may wonder why only λ neurons having this maximum number of synapses from ψ transponders will respond. It has to do with the events of the confabulation process. As the operate command input rises, these "sufficient" neurons will become active first. In the operation of the module (which is very fast) only those neurons with a sufficient number of inputs from an assumed fact will be able to participate in the dynamical convergence process. Another good question is why the variance in this number of synapses turns out to be small. This is because the binomial statistics of random transponder neuron axons make it such that neurons with unusually large numbers of synapses are extremely unlikely. Otherwise put, binomial (or Poisson) probability distributions have "thin tails." Thus, the set of all λ neurons which have strengthened synapses – the ones which participate in the (strength-weighted) excitation of λ – are those that lie in a narrow range at the high end of the Poisson density right before it plummets.

The binomial statistics of the locally random cortical connections also keep the number of target symbol neurons with near-maximum complements of input synapses very close to being constant for all symbols. Let this number of neurons be K. Then the total excitation of the K neurons which represent answer module symbol λ that are receiving input from ψ symbol transponders (where ψ is one of the assumed facts) is proportional to $K[a + \log_b(p(\psi|\lambda)]$ (again, with a universal constant of proportionality that is the same for all the symbols of one module).

Finally, since the subsets of λ-representing neurons which receive inputs from different links typically do not overlap, the total excitation of the entire set of neurons representing answer module symbol λ (assuming that λ is receiving knowledge link inputs from assumed facts α, β, γ, and δ) is approximately proportional to (again, with a universal constant of proportionality) the *total input excitation* sum I(λ):

$$\begin{aligned} I(\lambda) &\equiv K \cdot [a + \log_b(p(\alpha|\lambda))] + K \cdot [a + \log_b(p(\beta|\lambda))] \\ &\quad + K \cdot [a + \log_b(p(\gamma|\lambda))] + K \cdot [a + \log_b(p(\delta|\lambda))] \\ &= 4K \cdot a + K \cdot \log_b[p(\alpha|\lambda) \cdot p(\beta|\lambda) \cdot p(\gamma|\lambda) \cdot p(\delta|\lambda)]. \end{aligned} \tag{8.4}$$

Recall from the discussion of Sect. 8.2 that when the answer module attractor network is operated (and yields only one winning symbol), all of the neurons representing the winning symbol (which will be the one with the highest total input excitation) are left in the active state and all other symbol neurons are left inactive. By virtue of the above formula, we see that this winning symbol will be the symbol λ with the highest confabulation product $p(\alpha|\lambda) \cdot p(\beta|\lambda) \cdot p(\gamma|\lambda) \cdot p(\delta|\lambda)$ value (e.g., in the specific case of Fig. 8.6 this will be symbol ε). This is the theory's explanation for how thalamocortical modules can carry out confabulation.

Since not all symbols of the answer module of Fig. 8.6 receive knowledge links from all four assumed facts α, β, γ, and δ, what will be the input excitation sums on symbols that receive fewer than four link inputs (total excitation level of the entire ensemble of neurons representing that symbol in the answer module)? For example, consider an answer module symbol θ which receives links only from assumed facts β and δ. The total input excitation sum $I(\theta)$ of the set of neurons which represent θ will be:

$$\begin{aligned} I(\theta) &\equiv K \cdot [a + \log_b(p(\beta|\theta))] + K \cdot [a + \log_b(p(\delta|\theta))] \\ &= 2K \cdot a + K \cdot \log_b[p(\beta|\theta) \cdot p(\delta|\theta)]. \end{aligned} \tag{8.5}$$

Thus, given that each individual term in the first lines of Eqs. 8.4 and 8.5 lies between $K \cdot 10$ and $K \cdot 60$, the value of $I(\theta)$ (Eq. 8.5) could, in extreme cases, be larger than that of $I(\lambda)$ of Eq. 8.4 [although in most cases $I(\theta)$ will be smaller and θ will not be the winning symbol]. In any event, the symbol with the highest I value will win the confabulation.

Note that in cognitive functions which employ binary knowledge (every knowledge link transponder neuron synapse is either unstrengthened or is "strong"), $I(\lambda)$ is roughly proportional to the number of links that symbol λ receives. Thus, in these cortical areas, confabulation devolves into simply choosing the symbol with the most knowledge link inputs. Although it is not discussed in this chapter, this is exactly what such cognitive functions demand.

The seeming problem identified above of having symbols which are missing one or more knowledge links win the confabulation competition is not actually a problem at all. Sometimes (e.g., in early visual processing) this is exactly what we want, and at other times, when we want to absolutely avoid this possibility, we can simply carry out multiple confabulations in succession. Also, some portions of cortex probably have smaller dynamic ranges (e.g., 30–50 instead of 10–60) for strengthened synapses, which also helps solve this potential problem.

As discussed in Sect. 7.1, in mechanizing cognition we explicitly address this issue by appropriately defining a constant called the *bandgap*.

In summary, the theory claims that the above-sketched biological implementation of confabulation meets all information-processing requirements of all aspects of cognition; yet it is blazingly fast and can be accurately and reliably carried out with relatively simple components (neurons and synapses) which operate independently in parallel.

8.6 Action Commands

At the end of a confabulation operation there is often a single symbol active. For example, the triangular red cortical neurons (belonging to layers II, III/IVa) shown in Fig. 8.2 represent one particular symbol of the module which is now active following a confabulation. Of course, in a real human thalamocortical module, such an active symbol would be represented by tens to hundreds (depending on the location of the module in cortex) of "red" neurons, not the few shown in the figure.

A key principle of the theory is that at the moment a single symbol of a module achieves the active state at the end of a confabulation operation, a specific set of neurons in layer V of the cortical portion of that module (or of a nearby module – this possibility will be ignored here) become highly excited. The outputs of these cortical layer V neurons (shown in brown in Fig. 8.2) leave cortex and proceed immediately to sub-cortical *action nuclei* (of which there are many, with many different functions). This is the theory's *conclusion→action principle*. In effect, every time cognition reaches a definitive single conclusion, a behavior is launched. This is what keeps us moving, thinking, and doing, every moment we are awake.

The layer V neurons which become highly excited when a symbol wins a confabulation cause a very specific set of actions to be *executed* (or at least to be considered for execution, depending on the function of the action nucleus that receives the layer V efferents). This is the origin of all non-reflexive and non-autonomic behavior – each successful (one winning symbol) confabulation causes the launch of a set of *associated* action commands. These actions can be part of a movement process, part of a thought process, or both.

During development, the genetically determined program for creating the brain is, barring problems, executed. This program causes the development of axons from neurons in layer V of each cortical module portion which proceed to genetically directed sub-cortical action nuclei (of which there are tens). In other words, genetics can ensure that a module has the layer V neurons it needs to launch those actions which that particular module should be empowered to execute. Thus, each of us has a range of behavioral potentialities which are in this sense pre-determined. This is probably one important mechanism by which various talents and personality traits are transferred from parents to children. This is part of the "nature" portion of the human equation.

Given the behavioral potentialities established by the genetically directed wiring of the axons of the layer V neurons of a module to action nuclei, the big question is how exactly the correct ones of these layer V neurons end up getting "wired" from the population of neurons representing each symbol. Given the exact specificity of effect each layer V neuron produces, there is no room for error in this wiring from each symbol to the action commands it should launch. Since the local geometrical arrangement of the symbol representing neurons and action-command neurons within their respective layers is random, and their local axonal wiring is largely random, this wiring from symbol representing neurons to

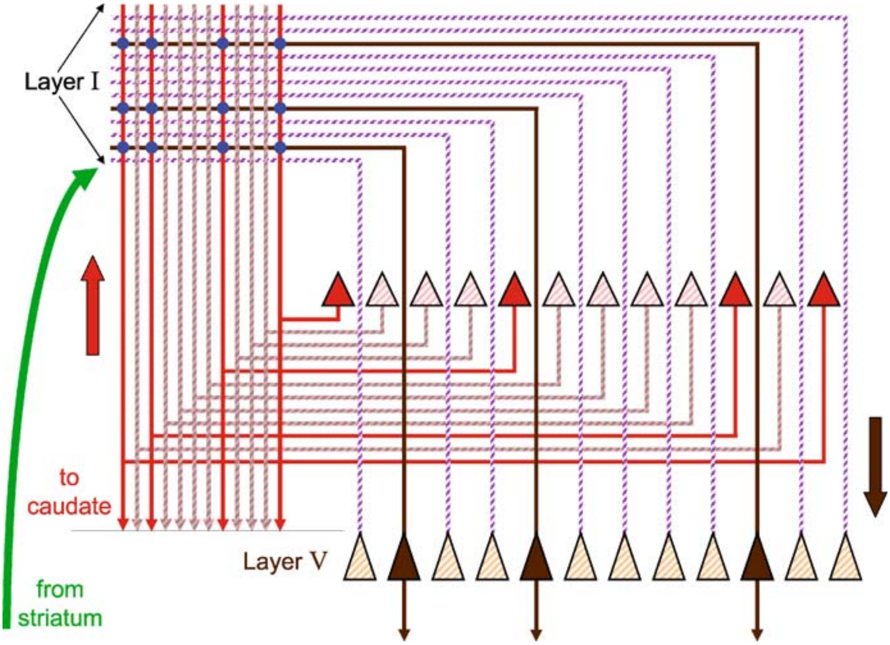


Fig. 8.8. Learning and using the precise associations from symbols to action commands. Keep in mind that the neuron populations involved in these associations, illustrated here as small sets, are, in the brain, extremely large sets (tens of thousands of neurons in every case). See text for explanation of the figure

layer V action-command-generating neurons cannot be genetically determined. These associations must be learned and they must be perfect. Figure 8.8 illustrates the theory's hypothesized mechanism for implementing these precise *symbol to action associations*. This figure will be referred to extensively below.

The learning of symbol to action command associations is almost certainly a totally different learning process from that used in development of module symbol sets or in the establishment of knowledge links. This symbol to action association learning process is hypothesized to take place primarily during childhood, but probably can also occur, when needed, during adulthood. Cognitive module development, cognitive knowledge acquisition, and symbol to action command association learning together make up the most "glamorous" parts of the "nurture" portion of the human equation (there are a number of other, quite different, learning processes that go on in other parts of the brain, e.g., learning that we should use the toilet).

Notice that in Fig. 8.2, every cortical layer of a module is mentioned except layer I (the most superficial). Layers II, III, and IV are primarily involved in symbol representation, precedence principle interactions among feature detector neurons, and the receipt of afferents from thalamus. Layer V is where the action command output neurons reside, and layer VI is where the cortical efferents to thalamus arise. The theory hypothesizes that layer I is where the wiring between the symbol-representing neuron sets and the layer V action command output neurons takes place (and quite possibly some of the wiring for the module's attractor network function as well). It is well known (Paxinos and Mai 2004) that the neurons of layer V (typically these are of the pyramidal category) have *apical* dendrites that ascend to layer I and then branch profusely. Further, neurons of layers II, III, and IV typically send large numbers of axon collaterals to layer I (and also frequently have apical dendrites too – but these will not be discussed here). Further, the *basal ganglia* [BG – a complicated set of brain nuclei known to be involved in multiple types of action learning (Paxinos and Mai 2004)], and specifically the BG sub-structure known as the *striatum*, sends signals in great profusion to layer I of cortex via the thalamus (see Fig. 8.8). This radiation is principally concentrated in frontal cortex (where most behaviors seem to originate), but other cortical areas also receive some of these inputs.

Given the random nature of cortical wiring, the only way to establish correct symbol to action associations is via experimentation. This experimentation is carried out (starting with the simplest actions and then constructing an *action hierarchy*). At the beginning of development of each module, the first item on the agenda is development of the module's symbols (which was discussed in Sect. 8.3). As this module development process begins to produce stable symbols, the problem of associating these to actions is addressed.

At first, action-command neurons are randomly triggered when a particular single symbol is being *expressed* by the module (i.e., that symbol was the lone outcome of a confabulation operation by the module). As this occurs, the BG monitor the activity of this module (via efferents from layer III and layer V – see Fig. 8.8). When a randomly activated action command happens to cause an action that the BG judge to be particularly "good" (meaning that a reduction in a drive or goal level was observed – which the BG know about because of their massive input from the limbic system), that action is then associated with the currently expressed symbol via the mechanism of Fig. 8.8.

(Note: Reductions in drive and goal states are almost never immediate following an action. They are usually delayed by seconds or minutes, sometimes by hours. One of the hypothesized functions of the BG (Miyamoto et al. 2004) is that they develop a large number of predictive models, called *critics* (Barto et al. 1983), that learn [via delayed *reinforcement learning* methods (Sutton and Barto 1998)] to accurately predict the eventual goal-or-drive-state-reduction "value" or "worth" of an action at the time the action is suggested or executed. It is by using such critic models that the BG are hypothesized by confabulation theory to immediately assess the worth of action commands produced by layer V outputs.)

When an action command that is randomly launched is indeed judged worthy of association from the currently expressed symbol of a module, a special signal (the green arrow in Fig. 8.8) is sent (via thalamus) from the striatum of the BG to cortical layer I of the module. This green signal causes the synapses (blue circles) connecting axon collaterals of the neurons representing the currently expressed symbol (these neurons are shown in red in Fig. 8.8 and reside in layer II, III, or IVa) with the apical dendrites of the now-validated action-command neurons of layer V (shown in brown in Fig. 8.8) to be incrementally strengthened. Essentially every neuron representing the expressed symbol gets its direct synaptic connections with the action-command neurons incrementally strengthened.

Notice how different the situation of Fig. 8.8 is from that of knowledge links. In a knowledge link, the source symbol must first amplify its signal by briefly recruiting thousands of transponder neurons to retransmit it. Even then, when the knowledge link signals arrive at the target module, only a relatively small fraction of each target symbol's neurons receive a sufficient number of inputs to complete the link. In Fig. 8.8, we presume that almost all of the expressed symbol's representing neurons synapse directly with the apical dendrites of each layer V action-command neuron. The reason this is a sensible hypothesis is that layer I is well known to be fed extensively with axons from the nearby neurons below it (i.e., neurons of the module that represent symbols), and to be profusely supplied with dense apical dendrites from layer V neurons.

The synapses from symbol-representing neurons to action-command neurons are hypothesized to be quite different from those used in knowledge links. In particular, these synapses can slowly and gradually get stronger (if repeatedly strengthened over many trials over time), and can slowly and gradually get weaker (if not strengthened very often, or not at all, over time). This is why "skill knowledge" decays so fast (in comparison with cognitive knowledge, which lasts for very long periods of time, even if not used). A major benefit of this dynamic synapse characteristic is that occasionally erroneous strengthening of synapses (e.g., when a random action-command set includes some irrelevant commands along with some effective ones) will, in general, not cause problems (as long as the vast majority of strengthenings are warranted). This is very different from cognition, where correction of erroneous knowledge is often impossible (and then the only solution is to specifically learn to not use the erroneous knowledge).

The universal truism that "practice makes perfect" is thus exactly correct when it comes to behavior. And for a difficult skill (e.g., landing a jet fighter on an aircraft carrier at night) to be usable, that practice must have been recent. The associations from symbols to action-command sets are constantly being reshaped during life. If we live in a highly stable information environment we might not notice much change in our behavioral repertoire over many years. If we are subjected to a frequently and radically changing information environment, our behavior patterns are constantly changing. In some respects, people who undergo such changes are being constantly "behaviorally re-made." The workings of the neuronal network of Fig. 8.8 are now briefly discussed.

Clearly, the size of the set of specific layer V action-command neurons which need to be triggered by the expression of a particular symbol is arbitrary. One symbol's association might involve activating a set of five specific layer V neurons, another might involve activating 79, and yet another might activate no layer V neurons. Keep in mind that each individual neuron in the population of tens to hundreds of neurons which together represent one particular symbol in a module also participates in many other such representations for other symbols. So, this association must be between the population representing a symbol and a specific set of layer V neurons.

This requirement suggests a unidirectional Willshaw-type associative network structure wherein the "retrieval keys" all have almost exactly the same number of neurons (which is exactly what the symbol representation neuron sets are like); but where the "output" neurons activated by each key have an arbitrary number of neurons. This is exactly what a Willshaw structure can do – the retrieval keys ("stable states") $\mathbf{x}_k$ must be random and must each have almost the same number of neurons, but there can be as many or as few "output neurons" in the associated $\mathbf{y}_k$ as desired, with no restriction, and the individual neurons making up each $\mathbf{x}_k$ population can appear in many other such populations. These are fundamental mathematical properties of the basic Willshaw design and are likely to apply to a wide class of similar systems. [Note: If you don't see this, consider again the computer experiments you performed in Sect. 8.3 above. You will see that it does not matter how many $\mathbf{y}_k$ neurons there are for each $\mathbf{x}_k$, as long as we are not implementing the second, $\mathbf{y}$ field to $\mathbf{x}$ field, part of the cycle (this is not well known, because, for analytical simplicity, the original Willshaw model used the same number of neurons in both the $\mathbf{x}_k$ and $\mathbf{y}_k$ vectors). Further, as long as the $\mathbf{x}_k$ keys are random and have almost exactly the same numbers of active neurons, the reliability of the $\mathbf{y}_k$ neuron responses is extremely high.]

However, as mentioned earlier, unlike the situation in knowledge links (where only a few of the target symbol neurons receive connections from the transponder neurons of a source symbol), in this case, almost all of the neurons of each active symbol must connect to each of the desired action-command neurons. Partial connectivity will not work here, since there is no feedback to implement a "convergence" process. However, the enormous local connectivity within layer I is hypothesized to make achieving a sufficient level of this connectivity no problem.

By incorporating inhibitory neurons into its intrinsic design, such a one-way Willshaw network (with inhibition added) will only respond with a $\mathbf{y}_k$ when its input is a newly active single symbol (multiple symbols will fail to yield any association output because they induce excessive inhibition, which shuts down all of the layer V neurons). This is hypothesized to be why action commands are only issued when a confabulation produces a single winning symbol. Also, when considering what action to take for a given $\mathbf{x}_k$ input, only those layer V neurons having a sufficient input excitation will respond (much like in confabulation competitions). In other words, even near the beginning of learning, when behavioral symbol to action associations are all weak, the layer V response will be based upon this "competitive" criterion, not a fixed threshold.

That a vast majority of cortex would be involved in issuing thought action commands, as opposed to movement action commands, makes sense because there are many more modules (and knowledge bases) than muscles. [It is not discussed here, but each knowledge base may also need to receive an "enable" command in order to function – if this is true, this function probably involves the large “higher order” (Sherman and Guillery 2001) portion of the thalamus that is not included in the thalamocortical modules.] So it probably requires a much larger portion of cortex to producing such thought process control action commands (muscle action commands come mostly from layer V of modules located within the relatively small *primary motor area* of cortex).

Most action commands represent “low-level housekeeping functions” that are executed reflexively whenever a single symbol (often one of a large set of symbols that will elicit the same action-command set) is expressed on a module. For example, if a confabulation in a module that is recalling a stored action sequence (such modules are typically located in frontal cortex) ends in the expression of a single action symbol, then that module must be immediately erased and prepared for generating the next sequence symbol. This is an action command that is issued along with the expression of the current action sequence symbol. Overriding such reflexive thought progressions is possible, but generally involves shutting off tonic cortical arousal (one of multiple adjuncts to the module operation command input) in a general cortical area via action commands issued to brainstem thought nuclei. The result is a momentary freezing of the halted function as a new thought process stream is inaugurated. This is what happens when we see that we are about to step on dog poop. It takes a only fraction of a second for us to recover from the suspension of the ongoing action and activate an alternative. Further, since muscle tone and rhythmic actions such as walking are nominally controlled by other brain nuclei (not cortex and thalamus), all the cortex typically needs to do (once the prior action sequence has been suspended) in such instances is issue a momentary set of *corrective alteration* action commands which are instantly executed as a momentary perturbation to the ongoing (sub-cortically automated) process, which then typically resumes.

It is important to note that the details of how sequences of “action” symbols – each representing (via its symbol to action command association) a particular specific set of action commands that will be launched every time that symbol is the sole conclusion of a confabulation – are learned, stored, and recalled are the same as with all other cognitive knowledge. However, unlike many other types of knowledge (e.g., event knowledge or factual knowledge), only the action symbol replay knowledge is rehearsed and solidified at night. The action symbol to action command associations can only be learned and refined via awake rehearsal. This accounts for the fact that anyone learning a new skill will frequently find themselves (either through vague memories of dreams upon waking, or via reports from their sleep partner) carrying out “silent practice” of those skills in their sleep. These do not involve launching the involved actions (a function that is normally suppressed during sleep), but simply running through the involved action symbol sequences. Such activities can help solidify the symbol sequences and this often yields improved skill performance the next day.

Quite a bit of experience has been gained with learning and recalling action symbol sequences in my year-long UCSD ECE-270 graduate course. For example, a checker-playing system that learns by expert-guided rehearsal has been demonstrated. However, issues surrounding the replay of action sequence hierarchies are complicated and not within the scope of an introductory book (e.g., provisions for automatic real-time, moment-by-moment modification of an ongoing lower-level action sequence replay in response to the exact current state of the world, with no modification at the higher level – a process called *instantiation* – must be introduced). Therefore, action symbol sequence learning and recall are not discussed in this book.

In summary, confabulation theory proposes that the unidirectional symbol pair links used in confabulation are the only knowledge learned and stored in cortex that is used in cognition. However, as described in this section, there is a second kind of knowledge learned and stored in cortex: the associations between each symbol and the action commands that its expression as a confabulation conclusion should launch. This knowledge is not really part of cognition. It is the mapping from *decisive cognitive conclusions* (single active symbols resulting from confabulations) to *behaviors*. Thus, the ultimate end product of cognition is the origination of action commands, some of which are unconditionally *executed* immediately and others, termed *suggested actions*, must be approved (*vetted*) by the basal ganglia before they can be executed.

8.7 Discussion

The theory's hypothesized cortical implementation of knowledge links has some important universal properties. First, the locally random wiring of the cortical axons can be established during development and then frozen; essentially for life (although there may be a very slow *replenishment* of some types of neurons throughout life that helps keep the brain functional as neurons slowly die; but this has not been established – the vast majority of neurons probably live a very long time, perhaps for the full life span of the individual). Knowledge links, by means of a parallel, two-step synfire chain communication process through the random cortical signaling network, can be immediately formed, as appropriate, between almost any two symbols in any two modules that genetics have provided connection possibilities for. A link can be temporarily established instantly (via the short-term memory mechanism) and then, if it is warranted, the link can be progressively transformed into permanent knowledge during the subsequent few sleep periods.

The price of this ability to instantly learn almost anything without need for rewiring (to carry out such wiring by growing new axons would take days and would require the involved axons to have unbelievable navigation abilities) is probably a vast over-wiring of cortex. A prediction of the theory is that only roughly 1% of cortical synapses are actually used to store knowledge (i.e., have been strengthened). The rest are there to provide the capacity for *instant arbitrary learning*. Thus, the old saw that "we only use 10% of our brain" is probably wrong on the high side; 99% of unstrengthened synapses are hypothesized to simply be sitting around waiting to be needed. This may seem wasteful, but unstrengthened cortical knowledge synapses and axon collaterals are small, and humans have about 10^{14}–10^{15} of them (Mountcastle 1998; Nicholls et al. 2001; Nolte 1999; Steward 2000). Clearly, the survival value of instant arbitrary learning vastly outweighs whatever inefficiency is incurred. This hypothesis helps explain one of the most puzzling findings of neuroscience: the vast majority of synapses that have ever been individually evaluated [e.g., by manipulating them, and monitoring their effects on the target cell, using multiple patch clamps (Cowan et al. 2001)] have turned out to be very unreliable and only marginally functional. This is exactly what you would expect to find if 99% of synapses are in a state of minimal existence, awaiting the possible moment that they will be needed.

Humans live for roughly 3×10^9 seconds. So, for example, if we acquire an average of one item of knowledge during every second of life (86,400 knowledge items per day), and if an average of 300 transponder neuron synapses are used to implement each knowledge item, far less than 1% of all synapses will ever be used (of course, not all cortical synapses are available for knowledge storage, but most probably are, so this conclusion is still probably correct). Thus, the theory proposes that the potential amount of cognitive knowledge that can be stored is huge.

In my laboratory's computer implementations of confabulation, a startling fact (which is consistent with the above numbers) has emerged: a staggeringly large number of knowledge items is needed to do even simple cognitive functions. The theory postulates that the average human must possess billions of items of knowledge. This has many startling and profound implications and, assuming that the theory gains acceptance, many philosophical and educational views of humans (and other animals) will likely be completely altered. For example, the theory implies that children (and adults too!) probably accumulate tens of thousands, or more, new individual items of knowledge every day. Thus, the process of reconsidering each day's short- and medium-term memories and converting selected ones into a more permanent form is a huge job. It is no wonder that we must sleep a third of the time.

To appreciate the vast storage capacity of your cerebral cortex, imagine for a moment that you are being asked a long series of detailed questions about the kitchen in your home. Describe all of the spoons and where they are kept; then the forks, then the drinking glasses, and so on. Describe how you select and employ each item; where and when you obtained it; and some memorable occasions when it was used. Obviously, such a process could go on for tens of hours and still turn up lots of new kitchen information. Now consider that you could probably answer such detailed questions for thousands of mental arenas. Humans are phenomenally smart.

Another cortical property which the theory's hypothesized design of cortex imparts is an insensitivity to occasional random neuronal death. If a few of the transponder neurons which represent a particular symbol randomly die, the remaining knowledge links from this symbol continue to function. Newly created *replenishment neurons* (which may slowly arise throughout life) which turn out to have the appropriate connectivity (once they have spread out and connected up and reached maturation), can be incorporated into such a weakened link to replace lost neurons, assuming the link is used from time to time.

If a link is not used for a long time, then as the transponder neurons of its source symbol slowly get *redeployed* (see below) or die, the axons to the target symbol neurons of the link will not be replenished and the link will become gradually weaker (other links having the same source symbol, which are used, will not suffer this fate because they will be replenished). Eventually, the unused link will become so weak that it cannot function by itself. Sometimes, when a link has become weak, but is not completely gone, it can be used if accompanied by additional assumed fact inputs to the same target symbol – a faded-memory recall trick known popularly as *mnemonics*. This is the theory's explanation for why we forget long-disused knowledge.

Another aspect of the hardware failure tolerance of cortex is the primary representation of each symbol within its own module. With tens or hundreds of neurons representing each symbol, the module's symbols too have some redundancy and failure tolerance.

When new inputs to a cortical module arise which do not fit any of the existing symbols well, and continue to appear repeatedly, new symbols can be formed, even in adulthood. Depending on how close to capacity the involved module is, these new symbols may or may not displace existing symbols. This *module rebuilding* process is often used to add new symbols to modules when we learn a subject in more depth (e.g., when we take Calculus III after having already taken Calculus I and Calculus II). Total rebuilding of a module typically only occurs in the event of trauma (e.g., stroke), where the entire information input environment to the module has dramatically changed. Total rebuilding takes weeks and requires lots of practice with the new symbols. This is why recovery of function after a stroke takes so long and why intensive and immediate physical and mental therapy based upon practice and use is so important. Aspects of childhood development are being recapitulated on an abbreviated schedule.

Modules also slowly incorporate replenishment neurons into existing symbol representations that are used. As with forgetting of knowledge, long-disused symbols eventually have their sets of representing neurons redeployed (see below) or eroded beyond functionality. A person who spoke French as a child, but who has not used French at all for 40 years, will likely have many of their French word representation symbols eroded beyond recovery.

The only instance of deliberate fast knowledge erasure in human cortex is *redeployment*, where a source symbol in a module, which used to be linked to a particular set of target symbols in other modules, suddenly has an entirely new ensemble of links to new target symbols arise for it, and these new links persist (and the old ones are disused). For example, when we move to a new home, it may be necessary to learn that the alarm clock is now on the left side of the bed, not the right. What happens in this instance is that the sets of transponder neurons representing the involved source symbol have a finite limit to the number of highly strengthened synapses that they can have at any time (this probably has to do with a total individual cellular limit on synthesis of certain consumable biochemicals – the critical ones of which are produced only in the neuron's soma and dendrites, where the ribosomes reside). [NOTE: The ultimate limit to knowledge storage capacity is probably not total synapses; it is the number of strengthened (knowledge link) synapses that each transponder neuron can support at one time. There are probably people (e.g., perhaps the author) who have spent their entire lives studying and who reached this capacity limit long ago for many symbols.] As the transponder neuron synapses implementing the many new links are learned and strengthened, many of the old, now unused, links must be immediately sacrificed (their synapses shrivel to the unstrengthened vestigial state). Within a few weeks, we instinctively reach left. The old knowledge has been effectively erased. The synapses of many of the old knowledge links have shriveled (but not all of them; some remnant knowledge links often remain – you can experience this, for instance, by revisiting one of your old haunts and trying to carry out formerly familiar patterns, like skipping down stairs at a childhood residence). Fragments of your former knowledge will still be there. Please be careful in conducting these experiments.

Redeployment is a critical cognitive capability that allows us to adapt to environmental change quickly. It is also hypothesized to be the only mechanism of deliberate forgetting in cognition.

Finally, it is important to note that any global theory of human cerebral cortex and thalamus is bound to be vastly oversimplified. For example, it is well known (Paxinos and Mai 2004) that different areas of cortex have some layers dramatically attenuated (e.g., layer IV in certain areas of frontal cortex). Other areas have layers that are dramatically elaborated (e.g., in primary visual cortex, layer IV becomes tripartite). These local modifications almost certainly must have significant meaning for the nuances of function. However, the theory proposes that these are all relatively small variations of the same overall grand theme.

The central notion of the theory – that cognition, that greatest engine of animal ennoblement, is universally mechanized by one information-processing operation (confabulation) employing a single form of knowledge (antecedent support), with each singular conclusion reached launching an associated set of action commands – seems to me to now be secure. The concreteness and specificity of this theory guarantee that it is falsifiable.

References

Abeles M (1991) Corticonics. Cambridge Univ. Press, Cambridge

Ackley DH, Hinton GE, Sejnowski TJ (1985) A learning algorithm for Boltzmann machines. Cognitive Science 9:147–169

Amari S-I (1989) Characteristics of sparsely encoded associative memory. Neural Networks 2:451–457

Amari S-I (1974) A method of statistical neurodynamics. Biological Cybernetics 14:201–215

Amari S-I (1974) A mathematical theory of nerve nets. In: Kotani M (ed) Advances in biophysics, vol 6. Tokyo University Press, pp 75–120

Amit D (1989) Modeling brain function: The world of attractor networks. Cambridge University Press, Cambridge

Anderson JA, Silverstein JW, Ritz SA, Jones RS (1977) Distinctive features, categorical perception, and probability learning: some applications of a neural model. Psychological Review 84:413–451

Anderson JA (1972) A simple neural network generating an interactive memory. Mathematical Biosciences 14:197–220

Anderson JA (1968) A memory storage model utilizing spatial correlation functions. Kybernetik 5:113–119

Barto AG, Sutton RS, Anderson CW (1983) Neuronlike adaptive elements that can solve difficult learning problems. IEEE Transactions on Systems, Man and Cybernetics SMC-13:834–846

Bender EA (1996) Mathematical methods in artificial intelligence. IEEE Computer Society Press, Los Alamitos, CA

Brown JW, Bullock D, Grossberg S (2004) How laminar frontal cortex and basal ganglia circuits interact to control planned and reactive saccades. Neural Networks 17:471–510

Caid WR, Hecht-Nielsen R (inventors) (9 January 2001) Representation and retrieval of images using context vectors derived from image information elements. US Patent 6,173,275

Caid WR, Hecht-Nielsen R (inventors) (6 July 2004) Representation and retrieval of images using context vectors derived from image information elements. US Patent 6,760,714

Canfield DE, Poulton SW, Narbonne GM (2007) Late-Neoproterozoic deep-ocean oxygenation and the rise of animal life. Science 315:92–95

Carpenter GA, Grossberg S (eds.) (1991) Pattern recognition by self-organizing neural networks. MIT Press, Cambridge

Casagrande VA, Guillery RW, Sherman SM (eds) (2005) Cortical function: A view from the thalamus. Elsevier Science, Amsterdam

Chomsky N (1980) Rules and representations. Columbia University Press, New York

Cohen MA, Grossberg S (1983) Absolute stability of global pattern formation and parallel memory storage by competitive neural networks. IEEE Transactions on Systems, Man and Cybernetics SMC-13:815–826

Cowan WM, Südhof TC, Stevens CF (eds.) (2001) Synapses. Johns Hopkins University Press, Baltimore

Crick F (1984) Function of the thalamic reticular complex: The searchlight hypothesis. Proceedings of the National Academy of Sciences 81: 4586–4950

Daugman JG (1988a) Complete discrete 2-D Gabor transforms by neural networks for image analysis and compression. IEEE Transactions on Acoustics, Speech and Signal Processing 36:1169–1179

Daugman JG (1988b) Relaxation neural network for non-orthogonal image transforms. In: Proceedings International Conference on Neural Networks, vol I. IEEE Press, Piscataway, NJ, pp 547–560

Daugman JG (1987) Image analysis and compact coding by oriented 2-D Gabor primitives. In: Barrett E, Pearson JJ (eds) Image understanding and the man-machine interface (SPIE Proceedings Vol. 758). SPIE Bellingham, WA, pp 19–30

Daugman JG (1985) Uncertainty relation for resolution in space, spatial frequency, and orientation optimized by two-dimensional visual cortical filters. Journal Optical Society of America, Part A 2:1160–1169

Daugman JG, Kammen DM (1987) Image statistics, gases, and visual neural primitives. In: Proceedings International Conference on Neural Networks, vol IV. IEEE Press, Piscataway, NJ, pp 163–175

Desai NS, Cudmore RH, Nelson SB, Turrigiano GG (2002) Critical periods for experience-dependent synaptic scaling in visual cortex. Nature Neuroscience 8:783–789

Duda RO, Hart PE, Stork DG (2000) Pattern classification and scene analysis, 2nd edn. Wiley-Interscience, New York

Einstein A (1961) Relativity: The special and the general theory. Three Rivers Press, New York

Fike DA, Grotzinger JP, Pratt LM, Summons RE (2006) Oxidation of the Ediacaran ocean. Nature 444:744–747

Freeman WJ, Holmes MD (2005) Metastability, instability, and state transition in neocortex. Neural Networks 18:497–504

Freiwald WA, Kreiter AK, Singer W (2001) Synchronization and assembly formation in the visual cortex. Progress in Brain Research 130:111–140

Fries P, Schröder J-H, Roelfsema PR, Singer W, Engel AK (2002) Oscillatory neuronal synchronization in primary visual cortex as a correlate of stimulus selection. Journal of Neuroscience 22:3739–3754

Fukuda T, Kosaka T, Singer W, Galuske RAW (2006) Gap junctions among dendrites of cortical GABAergic neurons establish a dense and widespread intercolumnar network. Journal of Neuroscience 26:3434–3443

Fukushima K (2005) Restoring partly occluded patterns: A neural network model. Neural Networks 18:33–43

Fukushima K, Miyake S, Ito T (1983) Neocognitron: A neural network model for a mechanism of visual pattern recognition. IEEE Transactions on Systems, Man, and Cybernetics SMC-13: 826–834

Fukushima K (1975) Cognitron: A self-organizing multilayered neural network. Biological Cybernetics 20:121–136

Gabor D (1969) Associative holographic memories. IBM Journal of Research and Development March:156–159

Gray CM, König P, Engel AK, Singer W (1989) Oscillatory responses in cat visual cortex exhibit inter-columnar synchronization which reflects global stimulus properties. Nature 338:334–337

Gregory R (2004) The blind leading the sighted. Nature 430:836

Grossberg S (1997) Cortical dynamics of three-dimensional figure-ground perception of two-dimensional patterns. Psychological Review 104:618–658

Grossberg S, Mingolla E, Ross WD (1997) Visual brain and visual perception: How does the cortex do perceptual grouping? Trends in Neurosciences 20:106–111

Grossberg S (ed) (1987) The adaptive brain, vols I and II. Elsevier, Amsterdam

Grossberg S (1976) Adaptive pattern classification and universal recoding. Biological Cybernetics 23:121–134

Hahnloser RHR, Seung HS, Slotine J-J (2003) Permitted and forbidden sets in symmetric threshold-linear networks. Neural Computation 15:621–638

Haines K, Hecht-Nielsen R (1988) A BAM with increased information storage capacity. In: Proceedings 1988 IEEE International Conference on Neural Networks, vol I. IEEE Press, Piscataway NJ, pp 181–190

Hebb D (1949) The organization of behavior. Wiley, New York

Hecht-Nielsen R (2006a) The mathematics of thought. In: Yen GY, Fogel DB (eds) Computational intelligence: Principles and practice. IEEE Computational Intelligence Society, Piscataway, NJ, pp 1–16 [Chap. 3 of this book]

Hecht-Nielsen R (2006b) The mechanism of thought. In: Proceedings of the World Congress on Computational Intelligence. Vancouver, BC, Canada. July 16–21, IEEE Press, Piscataway, NJ [Chap. 5 of this book]

Hecht-Nielsen R (2006c) The mechanization of cognition. In: Bar-Cohen Y (ed) Biomimetics. CRC Press, Boca Raton, FL, pp 57–128 [Chaps. 7 and 8 of this book]

Hecht-Nielsen R (2005) Cogent confabulation. Neural Networks 18:111–115 [Chap. 4 of this book]

Hecht-Nielsen R (2004a) A theory of cerebral cortex. Institute for Neural Computation, University of California, San Diego, Technical Report #0404

Hecht-Nielsen R (2004b) Perceptrons. Institute for Neural Computation, University of California, San Diego, Technical Report #0403

Hecht-Nielsen R, McKenna T (eds) (2003) Computational models for neuroscience. Springer, London

Hecht-Nielsen R, Zhou Y-T (1995) VARTAC: A foveal active vision ATR system. Neural Networks 8:1309–1321

Hecht-Nielsen R (1989) Neurocomputing. Addison-Wesley, Reading, MA

Herculano-Houzel S, Munk MHJ, Neuenschwander S, Singer W (1999) Precisely synchronized oscillatory firing patterns require electroencephalographic activation. Journal of Neuroscience 19:3992–4010

Hopfield JJ (1984) Neurons with graded response have collective computational properties like those of two-state neurons. Proceedings of the National Academy of Sciences 81:3088–3092

Hopfield JJ (1982) Neural networks and physical systems with emergent collective computational abilities. Proceedings of the National Academy of Sciences 79:2554–2558

Hyvärinen A, Karhunen J, Oja E (2001) Independent component analysis. Wiley Interscience, New York

Izhikevich EM (2007) Dynamical systems in neuroscience. MIT Press, Cambridge

Izhikevich EM (2006) Polychronization: Computation with spikes. Neural Computation 18:245–282

Karten HJ (1991) Homology and evolutionary origins of the "Neocortex". Brain, Behavior and Evolution 38:264–272

Kerr RA (2006) A shot of oxygen to unleash the evolution of animals. Science 314:1529

Kohonen T (1995) Self-organizing maps, 3rd ed. Springer, Berlin, Heidelberg, New York

Kohonen T (1984) Self-organization and associative memory. Springer, Berlin, Heidelberg, New York

Kohonen T (1972) Correlation matrix memories. IEEE Transactions on Computers C-21:353–359

Korb KB, Nicholson AE (2003) Bayesian artificial intelligence. CRC Press, Boca Raton, FL

Kosko B (2006) Noise. Viking/Penguin, New York

Kosko B (1988) Bidirectional associative memories. IEEE Transactions on Systems, Man and Cybernetics SMC-18:49–60.

Kryukov VI, Borisyuk GN, Borisyuk RM, Kirillov AB, Kovalenko EI (1990) The metastable and unstable states in the brain. In: Dobrushin R, Kryukov VI, Toom A (eds) Stochastic cellular systems: Ergodicity, memory, morphogenesis. Manchester University Press, UK

Mack A, Rock I (2004) Inattentional blindness. MIT Press, Cambridge

Mai JK, Assuher J, Paxinos G (2004) Atlas of the Human Brain, 2nd Edition. Elsevier, San Diego

Makeig S, et al. (2002) Dynamic brain sources of visual evoked responses. Science 295:690–694

Marder E, Prinz AA (2003) Current compensation in neuronal homeostasis. Neuron 37:2–4

Marder E, Prinz AA (2002) Modeling stability in neuron and network function: The role of activity in homeostasis. BioEssays 24:1145–1154

Markram H (2003) Elementary principles of nonlinear synaptic transmission. In: Hecht-Nielsen R, McKenna T (eds) (2003) Computational models for neuroscience. Springer, London, pp 125–169

Miller GA (1956) The magical number seven, plus or minus two: Some limits on our capacity for processing information. Psychological Review 63:81–97

Miyamoto H, Morimoto J, Doya K, Kawato M (2004) Reinforcement learning with via-point representation. Neural Networks 17:299–305

Mountcastle VB (1988) Perceptual neuroscience: The cerebral cortex. Harvard University Press, Cambridge, MA

Nicholls JG, Martin AR, Wallace BG, Fuchs PA (2001) From neuron to brain, 4th ed. Sinauer, Sunderland, MA

Nilsson NJ (1998) Artificial intelligence: A new synthesis. Morgan Kaufmann Publishers, San Francisco

Nilsson NJ (1965) Learning machines. McGraw-Hill, New York

Nolte J (1999) The human brain, 4th ed. Mosby, St. Louis, MO

Oertel D, Fay RR, Popper AN (2002) Integrative functions in the mammalian auditory pathway. Springer, New York

Palm G (1980) On associative memory. Biological Cybernetics 36:19–31

Patel A, Kosko B (2007) Levy noise benefits in neural signal detection. In: Proceedings of ICASSP-07 Conference, Honolulu, HI, April 15–20, IEEE Press, Piscataway NJ

Paxinos G, Mai JK (eds) (2004) The Human Nervous System, 2nd edn. Academic Press, San Diego

Pearl J (2000) Causality. Cambridge University Press, Cambridge

Pepperberg IM (1999) The Alex studies: Cognitive and communicative abilities of grey parrots. Harvard University Press, Cambridge MA

Sagi B, Nemat-Nasser SC, Kerr R, Hayek R, Downing C, Hecht-Nielsen R (2001) A biologically motivated solution to the cocktail party problem. Neural Computation 13:1575–1602

Sejnowski TJ, Destexhe A (2000) Why do we sleep? Brain Research 886:208–223

Sherman SM, Guillery RW (2006) Exploring the thalamus and its role in cortical function, 2nd ed. MIT Press, Cambridge

Shibata T, Tabata T, Schaal S, Kawato M (2005) A model of smooth pursuit in primates based on learning the target dynamics. Neural Networks 18:213–224

Sommer FT, Palm G (1999) Improved bidirectional retrieval of sparse patterns stored by Hebbian learning. Neural Networks 12:281–297

Steinbuch K (1965) Automat und Mensch, 3rd ed. Springer, Heidelberg

Steinbuch K (1963) Automat und Mensch, 2nd ed. Springer, Heidelberg

Steinbuch K (1961a) Automat und Mensch. Springer, Heidelberg

Steinbuch K (1961b) Die Lernmatrix. Kybernetik 1:36–45

Steinbuch K, Piske UAW (1963) Learning matrices and their applications. IEEE Transactions on Electronic Computers December:846–862

Steinbuch K, Widrow B (1965) A critical comparison of two kinds of adaptive classification networks. IEEE Transactions on Electronic Computers October:737–740

Steward O (2000) Functional neuroscience. Springer, New York

Sutton JP, Strangman G (2003) The behaving human neocortex as a network of networks. In: Hecht-Nielsen R, McKenna T (eds) (2003) Computational models for neuroscience. Springer, London, pp 205–219

Sutton JP, Anderson JA (1995) Computational and neurobiological features of a network of networks. In: Bower JM (ed) Neurobiology of computation. Kluwer Academic, Boston, pp 317–322

Sutton RS, Barto AG (1998) Reinforcement learning: An introduction. MIT Press, Cambridge

Tsypkin YZ (1973) Foundations of the theory of learning systems. Academic Press, New York

Turrigiano GG, Nelson SB (2004) Homeostatic plasticity in the developing nervous system. Nature Neuroscience Reviews 5:97–107

Turrigiano GG, Nelson SB (2000) Hebb and homeostasis in neuronal plasticity. Current Opinion in Neurobiology 10:358–364

Turrigiano GG, Leslie KR, Desai NS, Rutherford LC, Nelson SB (1998) Activity-dependent scaling of quantal amplitude in neocortical pyramidal neurons. Nature 391:892–896

von der Malsburg C (1990) Considerations for a visual architecture. In: Eckmiller R (ed) Advanced neural computers. Elsevier North Holland, Amsterdam, pp 303–312

von der Malsburg C (1981) The correlation theory of brain function. Internal Report 81-2, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany

Weir AAS, Chappell J, Kacelnik A (2002) Shaping of hooks in New Caledonian crows. Science 297:981

Widrow B, Hartenstein R, Hecht-Nielsen R (2005) Eulogy for Karl Steinbuch: 1917–2005. IEEE Computational Intelligence Society Newsletter, IEEE Press, Piscataway, NJ Fall:5

Willshaw DJ, Buneman OP, Longuet-Higgins HC (1969) Non-holographic associative memory. Nature 222:960–962

Xie X, Hahnloser R, Seung HS (2001) Learning winner-take-all competition between groups of neurons in lateral inhibitory networks. In: Proceedings of NIPS 2001 – Neural Information Processing Systems: Natural and Synthetic. MIT Press, Cambridge, pp 350–356

Zador PL (1963) Development and evaluation of procedures for quantizing multi-variate distributions. PhD Dissertation, Stanford University, Palo Alto, CA

Index

A

a posteriori probability 77, 97
a priori probability 77, 97, 109
Abeles synfire chain 102, 210
Abeles, Moshe 6, 87, 89, 90, 102, 120, 210, 229
action
  behavioral program 196
  deconflicting 196
  dithering 196
  hierarchy 221
  potentials 214
  sequences 148
  symbol 196
action commands 12, 73, 196, 197, 219, 221, 223
  associated 219
  muscle 224
  thought process control 224
action symbol 196
action symbol replay knowledge
  rehearsed and solidified at night 224
action unit 106
actions 14
  candidate 148
  deconflicting all 196
  hierarchy of 196
additive knowledge combination 11, 83
additive knowledge combination law 110, 123
almost logical deductions 147
Amari, S.-I. 104, 118, 201
Amari–Hopfield 'energy function'
  attractor network 201
Amit, D. 97, 104, 118, 201
Anderson et al. 97, 104, 118
Anderson, James 89, 90, 200, 201
answer module 75, 91
antecedent support 75, 102, 109, 110, 128, 143, 167, 201, 215, 228
  probabilities 151
antecedent support probability 75
Aristotelian information environments 147, 156
Aristotelian logic 18, 76, 91, 109
  information environment 91, 92
Aristotle 10
associative memory 89, 200, 229, 233, 234
assumed fact 74
assumed fact symbols 91
assumed fact word string 154
assumed facts 152, 153, 216
attended object 170, 185–188
attended object segmentation 170
attractor network 104, 200, 203, 204
  Amari–Hopfield energy function 201
  biological 205
attractor networks 87, 198
attractor states IX
attribute 8
audio front end 167
axon collaterals 84
axonal links
  are not direct 102

B

background noise 169, 216
  constant uniform blanket of 216
bandgap VIII, 82, 128, 140, 218
Barrington, Luke 117, 125

Barto, A.G. 221
basal ganglia (BG) 17, 106, 139, 148, 156, 221, 225, 229
  critics 221
Bayes
  law 97, 211
  naïve Bayes formula 77
Bayesian
  analysis 77
  approach 77
  pattern classifiers 77
  religion 76, 77
  viewpoint 108
behavioral program 196
behavioral triggering 90
behaviors 13, 106
  microbehaviors 106
Bender, E.A. 97
binary knowledge 218
binding problem 8
binomial statistics 217
biological implementation of cognition 193
biological implementation of confabulation 213, 218
biological learning processes 211
Boltzmann machines 158
Boolean AND 91
Borisyuk, G.N. 232
Borisyuk, M. 232
brain
  axon collaterals 84
  basal ganglia 139
  cerebellum 80, 139
  cerebral cortex 193, 195
  cerebral cortex and thalamus 139
  descriptive power of a 15
  entorhinal cortex 83
  higher-order thalamus 83
  hippocampus 83
  human cortex 81, 209
  neurons 81
  nucleus 193
  striatum 221
  superior colliculus 178, 182
  synapses 81
  thalamocortex 193
  thalamus 193, 197
  thought control nuclei 83
  transponder neurons 84
  universal language of 80
Brain State in a Box (BSB) 89, 201
Brown et al. 107
Busby, David 125

C

C software 202
candidate actions 148
canned thought processes 158
capitalized words 120
Carpenter, Gail 205, 206
Casagrande et al. 88, 100, 104, 118
causal pair of symbols 143
cerebellum 80, 139
cerebral cortex 193, 195
  storage capacity of 226
cerebral cortex and thalamus 139
checker-playing system 224
childhood 81, 208
children
  inherited talents and personality traits 219
Chomsky, Noam 115, 124, 167, 230
cocktail party 174, 233
cogency 75
cogency maximization 17, 76, 77, 83, 92, 109, 156
cogency maximization hypothesis 109
cogent 94
cogent confabulation 91
cognition 5, 99, 139, 193,
  See maximization of cogency
  alien nature of 166
  biological implementation of 193
  human sound 176
  mathematical process of 107
  mechanizers of 144
  pure 142
  true nature of 165

universal information processing operation of 204
visual 177
cognitive knowledge 5
cognitive knowledge acquisition 220
cognitive module 189
cognitive module development 220
cognitive world state 149
Cohen and Grossberg 104, 118
column sum 152
complex feature detector 183
computability questions 79
computer simulations 201
conclusion 12
conclusion → action principle 12, 107, 148, 156, 196, 219
conclusions 195
concrete numerical parameter values 99
confabulation 10, 73, 92, 99, 195
-based cognitive systems 146
-based vision system 146
computer implementations of 226
conclusion 105
entrepreneurs 144
variants 140
wave of 185
confabulation architecture 18, 89, 110, 112, 117–120, 125, 143, 150, 155, 159, 160, 170, 178
elements 153
confabulation conclusion
active 142
highly excited 142
confabulation consensus 12, 113, 114, 120, 122, 127
confabulation neuroscience 6, 99, 193
confabulation processes
ensembles of properly phased 165
confabulation product 75
Confabulation Theory VII, 5, 99, 102
Element 1 100
Element 2 102
Element 3 104
Element 4 106
implications 15
monkey-see/monkey-do principle of 115
of animal cognition 193
origins of 88
connection matrix 200
consensus building 140, 158
consensus conclusions 122
consensus of conclusions 78, 158
constraint satisfaction 158
contemplation 208
context 153
context sentence 160, 161
grouping 119
continuation sentence
grouping 119
control of movement 201
control of thought 201
convergence speed 88
co-occurrence counts 113, 121, 143, 151, 152, 188
ordered symbol pair 151
corpus colossum 10
cortical columns 6
cortical knowledge 210, 211, 214, 215, 225
cortical patch 6, 8, 79, 80, 84, 99, 100, 102, 105, 107, 197, 199
Cowan et al. 82, 226
Crick, Francis 81, 89
crosstalk 102

D

data flow viewpoint IX
Daugman, John 183
deconflicting all actions 196
defuzzification 78
development 219
distal modules 184
dithering 196
drive or goal level 221
duck test 92
Duda et al. 76, 77, 108
DVD video presentation 17, 139, 142, 191
DVDs VII, 19

E

$E = mc^3$ 77
education 139, 143, 144, 146
Einstein, Albert 81, 89
English grammar
  and composition 165
English text 93
entorhinal cortex 83
evolution 196
Excel spreadsheet 83
exceptional case 93
excited neurons 201, 207, 210
excited symbols 80
exhaustive knowledge 92, 142, 147
expectation 80, 91, 141, 172, 175, 184, 188
expresses one symbol 8
eyeball image 178
eyeball vector 182

F

Fair Isaac Corporation X, 90, 98, 110, 117, 125
fair use 144
Fan, Adrian T. 98
feature detector neurons 205–208, 220
  complex 205
  simple 205
feature representations 208
feature vector 171, 179
  Gabor logons 179
  image 179
feature vectors 168
field interactions 207
Fiofan, Monk 89, 232
firing threshold 201
fixation point 177, 178, 186, 190
  candidate 181
fixation points
  oversampling 182
Freeman and Holmes 106
freezing 95, 142, 208, 224
Freiwald et al. 89
Fries et al. 89, 106
Fukuda et al. 88
Fukushima, K. 104
Fundamental Theorem of Cognition 75
fundamental theorem of vertebrate cognition 94

G

Gabor 200
Gabor logon 179
  local image features 180
gap junctions 88
gaze controller 177, 178, 181–185, 188–190
genetics 219
global theory of human cerebral cortex and thalamus 228
Gödel paradoxes 79
graded blends of multiple symbols 80
grammar 96, 115, 124, 165–167
Gray et al. 89
Grossberg, Stephen 205, 206

H

Hahnloser et al. 87, 102, 118
Haines and Hecht-Nielsen
  102, 104, 118
  network 87
Haines, Karen 87, 89, 201
hearing 208
Hebb, Donald 6, 89, 102, 103, 121, 210, 231
Hebbian
  efficacy 211
  learning 82
  synaptic modification 9
  synaptic strengthening 211
Hebbian learning 200, 233
Hecht-Nielsen, Zeus IX, 90
Herculano-Houzel et al. 106
hierarchical architecture
  precedence principle 155
hierarchical confabulation architecture 155
hierarchical ring architectures 150

hierarchies 80, 104
hierarchy 196
higher-order thalamus 83
highest-abstraction-level symbols 166
hippocampus 83
holistic symbols 186
honing 173
Hopfield, J.J. 104, 118, 201
human cognitive functions
  action 194
  hearing 194
  language 194
  origination 194
  seeing 194
  somatosensation 194
human cortex 81, 114, 120, 122, 198, 209, 216
  fast knowledge erasure in 227
  gray and white matter of 209
  primary motor area of 224
  vast over-wiring of 225
human educator 189, 190
human language
  children learning 165, 176, 226
  universality of grammar and syntactic structure in all 167
human neuroanatomy and neurophysiology 194
human supervision 144

I

information processing 218
information processing effects 140, 204
information theory community 208
information transmission 214
inhibition 87, 88, 201, 203, 205, 207, 223
inhibitory elements 88
inhibitory interneurons 88
input excitation 75, 140, 202
  maximization 128
instantaneous learning 82
instantiation 225
intellectual property restrictions 144
intelligent machines 144
interference 102
interoperable 80
item of knowledge 9, 73, 75, 92, 104, 196, 198, 226
items of knowledge 95
Izhikevich, E.M. 88, 116, 124

K

Karten, H.J. 116, 118
Kirillov, A.B. 232
k-means 205
knowledge
  additive knowledge combination law 110
knowledge acquisition process 95
knowledge base enablement 83, 148
knowledge bases 5, 103, 120, 172, 182, 197, 198, 209
  causal 151
  primary visual layer 185
knowledge link 7, 73, 102, 103, 120
  antecedent support probability weightings of 109
  axonal 110
  dense connections of 200
  probabilities 152
  source symbol of 103
  sparse 212
  sparse connections of 200
  target symbol of 103
knowledge storage capacity
  limit to 228
Kohonen, T. 200, 205, 207, 208
Korb and Nicholson 108
Kosko, Bart 104, 118, 201, 216
Kovalenko, E.I. 232
Kryukov, V.I. *See* Fiofan, Monk

L

LabVIEW 202
language cognition 149
language generation 150

language hierarchies 147, 154
- phrase level 154
- phrase module 154
- word level 154
- word module 154

language hub cognitive architecture 16
language is the hub of human cognition 16
layer I 220
layer II, III, or IVa 222
layer V 220, 222, 224
layer V neurons 223
layer VI 220
layers II, III and IVa 197
layers II, III, and IV 220
learning 113
- instant arbitrary 225
- phases 121
- taking place on demand 104

learnmatrix 200
Lenneberg, E. 115, 124
likelihood 76
limb load 81
limbic system 139, 221
link strength 213
links per knowledge base 113
local inhibition 88
locked symbols 78
logarithmic scale 173, 216
logarithms 92, 216
long-range context 159
long-term memory 196, 210
long-term potentiation 82
lowpass filtering 167

M

Mai, J.K. 233
Makeig et al. 106
mammalian auditory neurophysiology 171
mammalian primary visual cortex 183
Marder and Prinz 82
Marder, E. 82, 211
Marder–Turrigiano 82
- saturable efficacy factor 82
- variable post-synaptic receptivity learning 82

Marder–Turrigiano–Hebb learning process 211
Mark, Kate X, 98, 117, 125
Markram, Henry 214
massively parallel application of relevant knowledge 81
MATLAB 202
maximization of cogency 91, 107
meaning content 153
- critic subsystem 191

meaning fuzzification 78
meaningful symbol co-occurrence 103
Means, Robert W. X, 82, 98, 117, 125, 154, 159
mechanism of thought 117
Mechanism of Thought video presentation 19
mechanization of cognition 192
mechanizing sound cognition 176
medium-term memory 196, 210
memory 143
- and mnemonics 227
- and sleep 226
- long-term 196, 210
- medium-term 196, 210
- short-term 196, 210

mentation 193
microbehaviors 12, 106
Miller, G.A. 115, 124
mnemonics 227
Model 5 88
Model 6 88
Model 7 88
Model A 83–87
module
- command input 204
- cortical patch 6, 8, 79, 80, 100, 102, 105, 107, 197
- cortical patches 84
- freezing 142
- input vector 182

- source 198
- target 198
- vision 189

module tightenings 114, 122

modules 14, 197
- as muscles of thought 105, 106
- groupings of 119
- next-word acoustic 173, 175
- primary sound 171, 172
- primary visual 182
- sound phrase 172

monkey-see/monkey-do 115
- skill acquisition principle 124

monkey-see/monkey-do principle 165, 191

monocular (cycloptic) vision 177

Mountcastle, V.B. 226

movements 197
- ongoing 197

moving 208

multiconfabulation 12, 15, 78, 99, 110, 114, 117, 122, 140, 153, 158, 194, 198

multilayer perceptron 178

multi-sentence language units 159

multi-word conceptual units 154

multi-word phrases 112

muscles of thought 105, 106, 124

mutual consultation 115, 123

N

Nemat-Nasser, Syrus C. X, 98, 117, 125

Neocognitron 104

network of networks 100, 200, 233, 234

network of networks hypothesis 100

neural field 200

neurocomputing 87

neuron populations 6

neuronal attractor network 11, 105, 118

neuronal calculational precision and accuracy 217

neuronal implementation of knowledge links 83
- Model A 83

neuronal information processing 'hardware' of thought 7

neurons 81
- active 195
- excited 195, 210
- inactive 195
- layers II, III and IVa 197
- massively parallel mutual interactions of 195
- target 216
- transponder 209

neurotransmitter receptivity 211

Newton, Isaac 81

next word 97

next-word acoustic module 170–176

Nicholls et al. 226

Nilsson, N.J. 97, 205

noise blanket 217

Nolte, J. 226

non-exceptional assumed facts 93

non-holographic associative memory 200

normalization 129

NRT *See* thalamic reticular nucleus

null conclusion 129

null state 78

null symbol 140, 195

Nyquist rate 168

O

object of the mental universe 8

object segmentation 185

Office of Naval Research (ONR) X, 90

ordered symbol pair co-occurrence counts 151

output neurons 223

oversampling 182

P

Palm, G. 201

parallel synapses 215

parsing 154

patch clamps 82, 104, 226

Patel, Ashok 217

pattern measurements 77

Paxinos and Mai 100, 102, 171, 221

Paxinos, G. 233
Pearl, J. 77, 97, 108
perceptron 178, 181, 182, 190, 216
phrase completion 94, 149, 152
phrase level 154
phrase module 154
Poisson probability distributions 217
  binomial 217
  thin tails 217
polysynchrony 116, 124
pose insensitive 186
pose insensitivity 181, 189
position-dependent 150
post-synaptic signaling efficacy 211
postural goal 14
PowerPoint VIII
precedence principle 155, 156, 183, 186, 206, 207, 220
Presentation Notes VIII, 19, 78
primary auditory 81
primary layer modules 188
primary layer symbols 187
primary sound modules 170, 172, 173, 175, 184
primary visual cortex 80, 182, 206, 228, 230
primary visual layer 183, 184
primary visual layer knowledge bases 185
primary visual modules 182
Prinz, A.A. 211
professional self-study VII
programmed computing 148
proper English 93, 94, 113, 121, 123, 143, 146, 150, 155, 161
punctuations 112
pure cognition 142
purple box 110, 112, 118, 120, 159, 160

Q

quasiorthogonalization 171
quasiorthogonalization of sources 171
quasiorthogonalizing features 171

R

raven 15
rehearsal learning 14
reinforcement learning methods 221
relevant knowledge 78
replenishment neurons 226
re-representation 85
retrieval keys 223

S

Sagi et al. 169, 170, 174
saturable efficacy factor 82
secondary layer modules 188
secondary module symbols 187
Sejnowski and Destexhe 83, 103
self-organization process 205
semantically replaceable elements (SREs) 147, 166
sentence continuation 149, 153
  thought process 121, 122, 127
  with contex 125
  with context 110, 117, 119, 161
  without context 110, 117, 119, 125, 161
sentence generation 191
sentence meaning content summary module 112, 120, 159, 160
sentence starter 117, 160
Seung networks 87
Seung, H.S. 87
Shannon worship 208
Sherman and Guillery 88, 100, 104, 118, 224
Shibata et al. 107
short-course VII
short-term (or working) memory 74, 196, 210
signal processing 167
  techniques 167
Singer, Wolf 89, 90
skill knowledge 13, 73, 222
skill learning 13
sleep 226

smallest meaningful antecedent
  support probability value 215
Smith, Andrew X, 117, 125
Solari, Soren 197
somatosensation 194, 208
somatosensory cortex 197
Sommer and Palm 87, 97, 104, 118
Sommer, F.T. 201
sound feature bank 168
sound phrase modules 170, 172–175
sound sample vector 168
sound symbol excitation vector 169
source module 10, 198
source symbol 75
sparse knowledge link 212
speech recognizer
  for colloquial speech 176
spike timing 89
stable states 201–204, 208, 215, 223
  bipartite 202
stable terms of reference 8, 79
standard language constructions 154
Steinbuch learnmatrix 200
Steward, O. 226
striatum 221
stroke 227
subcortical action nuclei 219
suggested actions 225
superior colliculus 178, 182
surrounding objects 185
Sutton and Strangman 100
Sutton, J.P. 221
swirling 12, 78, 112, 113, 120, 121, 160
syapses
  synfire chain terminal 104
symbol
  active 148
  convergence property of 204
  co-occurrence counts 128
  excitation levels 128
  excitations 129
  input excitation 128
  source 102
  target 102
  tightening 129
  to action command associations 220
symbol pair counts 95
symbolic processing 102
symbol-representing neuron
  population 197
symbol-representing neurons 8
symbols 8, 73, 79, 101
  associations to action command
    sets 222
  causal pair of 143
  complex feature detector 183
  excited 169
  highest-abstraction-level 166
  holistic 186
  locked 78
  merging 188
  stable terms of reference 9
  terms of reference 101
synapses 81, 104, 213, 215, 228
  cortical excitatory 214
  efficacy of 216
  function of 214
  knowledge coding of 216
  limited dynamic range of 216
  non-binary strengthened 216
  parallel 215
  patch clamps 226
  strengthening of 103, 215
  vestigial 215
synfire chain terminal synapses 104
synfire chains 77, 87, 89, 102, 104, 210, 225
  cortical 214
syntax 115, 124, 166

T

target module 10, 198, 211
target neurons 216
target symbol 75
tertiary layer module symbols 186
tertiary layer modules 188–190

text corpus 143
text generation 166
text training corpus 191
thalamic reticular nucleus
  88, 197, 199
thalamic zone of all module 6
thalamic zones 197
thalamocortex 193
thalamocortical loop 88
thalamocortical module 5, 73, 100, 106,
  118, 194, 212
  development process of 205
  physiology of 197
thalamus 193
  higher-order portion of 224
thinking 5, 208
  is a phylogenetic outgrowth of
    movement 14
thought command input to
  a module 14
thought command signal 11, 73,
  104, 105
thought control command 12
thought control nuclei 83
thought control signal input 105
thought controller 15
thought process
  execution 196
thought processes 14
threshold control 201
time-warp insensitive 174
total input excitation 11
total input excitation sum 217, 218
training 139, 143, 187
training and education 143
training process 151, 152, 182
transduction 167
  of a soundstream 167
transit time 81
transponder neurons 85, 103, 209,
  211, 228
trout 15
Tsypkin, Y.Z. 205
Turrigiano and Nelson 82
Turrigiano et al. 82
Turrigiano, G.G. 82, 211
two-stage synfire chain 83, 120

U

UCSD ECE-270 83, 87, 88, 224
unidirectional weighted links
  between pairs of symbols 209
unidirectional Willshaw-type
  associative network structure 223
unified theory of cognition 191
universal information processing
  operation of cognition 204
universal interoperability of cognitive
  knowledge 83
Universal Language of Cognition 79
university course VII

V

variable post-synaptic receptivity
  learning 82
vector quantization 183, 205, 207, 208
vector quantizer 171, 176, 208
vertebrate cognition 91
vertebrate species 193
  birds 193
  cats 193
  dogs 193
  fish 193
  frogs 193
  parrots 193
  primates 193
  ravens 193
  reptiles 193
vestigial knowledge link final
  synapses 82
vestigial synapse 215
viable conclusions 76, 81
video camera sensor 177
video presentation VII
Video Presentation Viewcells 21
vision 208
  eyeball vector 182
  fixation point 177, 178, 181, 190
  image feature vector 179

- monocular (cycloptic) 177
- pose insensitive 186
- primary visual layer 183
- secondary visual layer 184
- tertiary visual layer 184

visual attributes 177, 190
- color 177
- form 177
- lighting 177
- pose 177

visual environment 183
visual module 189
- secondary 186
- tertiary 186

visual object completion 189
visual object segmentation 148
von der Malsburg 8, 89, 234
Voronoi-like partitioning 208
VQ *See* vector quantization
- codebook 183, 206, 207
- codebook vectors 207

W

wave of confabulations 185
Willshaw et al. 87
Willshaw, David 87, 89, 90, 200, 201, 223
winners-take-all 10, 12, 73, 81, 89, 97, 104, 110, 117, 195, 201
- competition process 99, 105, 118
- implementation of confabulation 97

word level 154
word module 154
words 112
working memory 142, 150, 156

X

Xie et al. 87, 102, 118

Z

Zador, P.L. 171, 205