DIVISION OF LABOR IN A COMPUTATIONAL MODEL OF VISUAL WORD RECOGNITION

by

Michael Wayne Harm

A Dissertation Presented to the

FACULTY OF THE GRADUATE SCHOOL

UNIVERSITY OF SOUTHERN CALIFORNIA

In Partial Fulfillment of the

Requirements for the Degree

DOCTOR OF PHILOSOPHY

(Computer Science)

August 1998

Dedication

This thesis is dedicated to my parents.

Acknowledgements

I would like to thank my fellow grad students at USC for many helpful conversations, thought provoking and engaging arguments, advice and help throughout the duration of my Ph.D. In particular I would like to thank Joe Allen and Robert Thornton for innumerable after hour discussions, and for their willingness to tell me when I was totally wrong, even when it meant enduring a storm of defensive invective from me. Laura Gonnerman, Joe Devlin and Marc Joanisse also contributed immensely to the development of my thinking on many issues. Thanks, guys.

I further would like to acknowledge the faculty at USC. Dr. Michael Arbib's brisk humor, keen ability to spot flaws in one's reasoning and cut to the core of an issue provided an invigorating environment for developing ideas. Dr. Maryellen MacDonald's fresh perspective and unflinching rigor guided my thinking to different domains while keeping me honest in my primary area of research.

Mostly I would like to thank my advisor Dr. Mark S. Seidenberg for showing me what it means to be a serious scientist. Ever since I was very young I knew I wanted to be a research scientist; I'm grateful to have had the opportunity to learn from one of the best. Thanks, Mark.

Contents

List Of Tables

List Of Figures

Typographic Conventions

The following typographic conventions are used throughout. The spelled form of a word is listed in small caps, i.e. CAT. The phonological form is listed between slashes, i.e. /kæt/. Semantic features are represented in angle brackets, as in <*bad*>; the presence of a feature is denoted <+*bad*> and the absence as <-*bad*>. A semantic concept, comprising a bundle of features, is represented between square brackets, as in [dog]. Hence the word spelled CAT sounds like /kæt/, has the property <+*meows*> and corresponds to the concept [cat].

Abstract

How do we compute the meanings of written words? For decades, the basic mechanisms underlying visual word recognition have remained controversial. The intuitions of educators and policy makers, and the existing empirical evidence have resulted in contradictory conclusions, particularly about the role of the sound structure of language (phonology) in word recognition. To explore the relative contributions of phonological and direct information in word recognition, a large scale connectionist model of visual word recognition was created containing orthographic, semantic and phonological representations. The behavior of the model is analyzed and explained in terms of redundant representations, the development of dynamic attractors in representational space, the time course of activation and processing within such networks, and demands of the reading task itself. The different patterns of results that have been obtained in previous behavioral studies are explained by appeal to stimulus composition and properties of a common experimental paradigm. A unified explanation of a wide range of empirical phenomena is presented.

Chapter 1

Introduction

[phonics-based word lists are] skeleton-shaped, bloodless, ghostly apparitions, and hence it is no wonder that children look and feel so deathlike, when compelled to face them.
— Horace Mann, 1841

. . . the [non-phonic] word method consists essentially of treating children as if they were dogs . . . It's the most inhuman, mean, stupid way of foisting something on a child's mind.
— Rudolph Flesch, 1955

Reading by "phonics" is demonstrably impossible. Ask any computer.
— Frank Smith, 1973

We propose that the relatively invariant correspondence between orthographic representations and phonologic representations explains why word identification appears to be mediated by phonology.
— Van Orden and colleagues, 1990

[phonics] "skills" are helpful in reading only when they help make texts more comprehensible. They should not be the core of a language arts or reading program.
— Stephen Krashen, 1996

How do we read? Literacy has been a part of human culture for millennia, and yet the fundamental mechanisms governing the acquisition, use and breakdown of this skill are still subjects of intense (and often emotional) controversy (see Adams, 1990, for historical review).

Logically, the meaning of a word can in principle be computed directly from the spelling of the word, or by translating the word into its sound form, and then accessing its meaning from that (Figure 1.1). However, the extent to which phonological representation actually enters into this process is generally very contentious. As a result, there are currently no universally accepted methods for effectively teaching literacy, remediating poor developing readers, and assisting patients in the recovery of literacy after brain damage.

Opinions within the field of reading research span the range from those who assign no useful role to phonological processing in the computation of meaning, to those who feel that phonological encoding is the dominant, primary and initial method of lexical access. Given the absence of an explicit understanding of the computational principles that underlie the reading process, intuitions and empirical evidence are widely varied, and often contradictory.

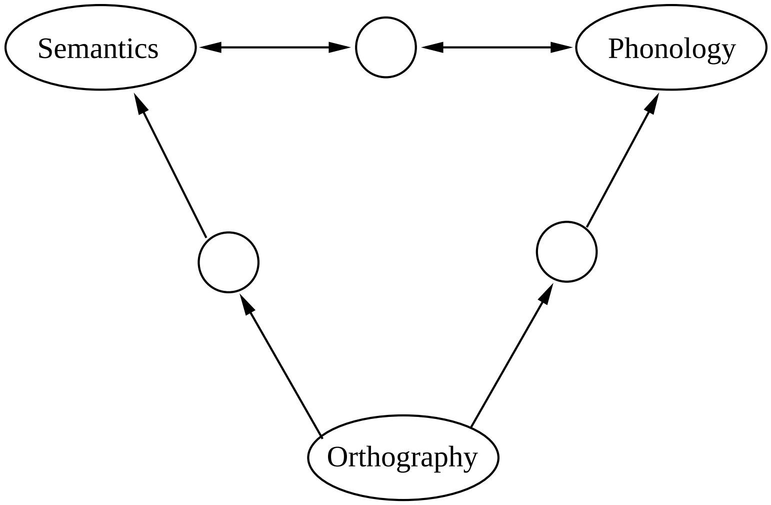


Figure 1.1: The Triangle model of word recognition. There are two paths from orthography to semantics: orth→sem and orth→phon→sem. There are, reciprocally, two pathways to phonology: orth→phon and orth→sem→phon.

I will first review evidence and arguments for varying degrees of phonological implication in the access of meaning. Existing models of word recognition will be introduced and critiqued. Then I will introduce a new, novel theory of word recognition based on connectionist principles of knowledge representation, acquisition and processing. An explicit computational model embodying these principles will be introduced and analyzed, and its behavior will be linked to widely divergent empirical findings and theoretical claims, providing a uniform explanation of the phenomena. The division of labor between phonological and non-phonological processes in the model's computation of meaning will be analyzed with respect to the model's underlying computational principles. The model will further provide insights into a less hotly debated, but still interesting reciprocal question: having asked what is the role of phonological and direct mapping in the computation of meaning, what is the role of semantic and direct mapping in the computation of a word's pronunciation?

1.1 Intuitions and Evidence

Theoretical arguments for the primacy of phonology in word recognition are widely varied and derive from different domains.

Children have a large auditory vocabulary when learning to read; mastery of spelling to sound correspondences allows them to tap into this vocabulary. Jorm and Share (1983) argue that children who can sound out words (either overtly or covertly) can then match them to words known from speech, providing a kind of self-teaching mechanism. In this view, reading is parasitic on existing linguistic knowledge of the associations of the sound pattern of words with their meanings, and intuitively, a reading instruction program should capitalize on this existing base of knowledge.

Van Orden and colleagues (Van Orden, 1987; Van Orden, Johnston, & Hale, 1988; Van Orden et al., 1990) present a different argument. They argue that in English, orth→sem is largely uncorrelated, while orth→phon has broad regularities. As such, reading by orth→sem is much more difficult than by orth→phon. Hence, simply from an ease of mapping viewpoint, orth→sem is too hard to learn. Van Orden (1991) explicitly links this theorizing with Smolensky's (1986) harmony theory, claiming that a dynamical system mapping orthographic to phonological word forms (and back) would exhibit great self consistency, and this would facilitate further learning of other items and constrain the formation of other, less consistent codes (e.g., semantics).

A similar argument is presented, with great vehemence and invective, by Flesch (1955), in which he calls for a reading instruction program which emphasizes the importance of orth→phon. Flesch argues that when the regularities of orth→phon are overlooked in educational programs, it forces children to "read English as if it were Chinese." He rages that these methods of teaching have "thrown 3,500 years of civilization out the window and have gone back to the Age of Hammurabi" (p. 5).

Other more even tempered researchers echo the view that an alphabetic method of writing is more "modern", noting that over time alphabetic scripts have tended to replace syllabic or pictographic ones.

A common thread here is that in English, orth→phon is more easily learned than orth→sem, and children already have acquired phon→semfor a large set of words, so naturally, orth→phon→sem is the simple, easy way to learn to read.

There is empirical support for this position. A vast empirical literature has linked children's phonological abilities (e.g., performance on tasks such as phoneme identification, manipulation, segment splitting, etc.) with their performance in learning to read (Bradley & Bryant, 1983; Tunmer & Nesdale, 1985; Mann, 1984; Olson, Wise, Conners, Rack, & Fulker, 1989; Shankweiler & Liberman, 1989; see Adams, 1990 for review). In such studies, children are typically asked to perform a phonological task, and their performance is then related to their reading development. Tasks include syllable splitting, where children are asked to split a sound pattern into an onset and rime (e.g., /kət/ to /k/ and /ət/; Share, Jorm, Maclean, & Matthews, 1984), phoneme blending (e.g., "what do /b/, /ə/ and /t/ produce?" Lundberg, Olofsson, & Wall, 1980), and phoneme deletion (e.g., "say SPLIT without the puh;" Mann, 1984; Rosner & Simon, 1971).

Additionally, impairments in phonological processing have been implicated in some forms of developmental dyslexia (Manis, Seidenberg, Doi, McBride-Chang, & Peterson, 1996; Stanovich, Siegel, & Gottardo, 1997; Murphy & Pollatsek, 1994). In the Manis et al. (1996) study, children found to be poor at a phoneme deletion task were also found to be impaired in nonword naming, and to exhibit a developmental trajectory which is aberrant from that of normal children. Harm and Seidenberg (1998) have replicated this finding with a connectionist network which maps orthographic patterns onto phonological ones, suggesting that direct semantic mediation is not necessary to account for this early developmental impairment.

These developmental results find a natural interpretation within a theory which states that word recognition is initially phonological. If children initially learn to read by phonologically encoding a word, then clearly an impairment in phonological processing will be disruptive.

Studies of adult readers provide further support for the phonological coding hypothesis. Van Orden (1987) tested subjects on a semantic decision task. Subjects were asked "is it a food?" and then provided with exemplars (e.g., MEAT), and homophonous foils (e.g., MEET), and nonhomophonous spelling controls (e.g., MOOT). He found that subjects made a high number of false positives on phonological foils relative to orthographic controls. How, he argued, could these results obtain if phonological coding was not used in the course of lexical access? The experiment was replicated by Van Orden et al. (1988) using pseudohomophone stimuli (e.g., "is it clothing?"

SUTE). Van Orden and colleagues have since articulated a theory of word recognition in which the phonology of a word is the initial source of information; homophones are disambiguated by a later, "spelling check" procedure only after initial access has occurred (Van Orden et al., 1990).

Lesch and Pollatsek (1993) and Lukatela and Turvey (1994b, 1994a) extend these findings by manipulating the specific time course of access. The reasoning behind these studies is as follows. If, as Van Orden and colleagues suggest, word recognition is initially phonological, and homophones are disambiguated by a subsequent spelling check, then there should be a point in time in which homophones cannot be distinguished, and a (later) point in time in which they reliably are. Lesch and Pollatsek (1993) and Lukatela and Turvey (1994b) used a semantic priming paradigm to explore this hypothesis. Masked priming was used to halt processing of the input stimuli at varying times. In the crucial conditions, a prime word that is either related to a target word, or homophonous with a word that is related to a target word is presented. The prime word is presented for either a short (50ms) or long (250ms) duration. Then, the prime word is "masked" by the target, which must be named.

Inappropriate prime words (e.g., TOWED for FROG) produced facilitation in the naming of the target word only when the inappropriate item was presented very briefly. Lesch and Pollatsek obtained very similar results with a similar paradigm. These effects are taken by the authors as confirmation of the Van Orden hypothesis: lexical access is initially phonological, where the phonological code activates all candidate items. A subsequent spelling check operation inhibits the inappropriate item. By masking the stimuli at an early stage (50ms), this spelling check is prohibited from taking place; the orthographic information is no longer present. These results will be discussed in greater detail in Chapter 6.

These studies, when taken together, provide clear evidence against a theory of word recognition which allots no role to phonological coding of words. But do they demonstrate the polar opposite, extreme view, that word recognition is initially phonological? There is reason to think that the answer is no.

While education policy critics like Rudolph Flesch argue for the ease of phonological coding, other educational policy pundits argue for the efficiency of non-phonological reading. Smith (1971, 1973b) argues that English is too rampant with homophony to make such a strategy reasonable. Using arguments from signal detection theory, he claims that a two stage decoding process (orth→phon, phon→sem) would be too slow to support automatic, rapid reading, and that truly skilled reading must include direct translation of a word's spelling into its meaning. While the English monosyllables contain over 1,000 homophone pairs, there are very few homographs; therefore orth→sem is less ambiguous than orth→phon→sem. Smith (1971) further argues that the spelling to sound rules of English are in fact extremely complex, and mastery of them is a daunting task. He concludes that orth→sem is the way that reading must logically proceed.

The strongest interpretation of Smith's appeal to intuition cannot be correct; there are far too many studies showing phonological effects in the access to meaning. However, a more intermediate interpretation is certainly plausible. This is that there are two ways to access meaning: phonological and non-phonological. This view assumes that differing factors govern which pathway is dominant in a given situation.

Consistent with this view, Jared and Seidenberg (1991) replicated the Van Orden (1987) results, but then went on to show that the effects obtain only for certain kinds of stimuli. They hypothesized that the effects were in part driven by priming of the target item by the category. For example, it is very possible that Van Orden's categories (e.g., "is it a flower?") primed the actual token. To eliminate this possibility, Jared and Seidenberg used very broad categories (e.g., “is it an object?” or “is it a living thing?”), where the likelihood of priming the actual item was far lower. They then experimentally manipulated the frequency of the exemplar (e.g., ROSE) against the frequency of a homophone foil. In their study, only the low frequency foils of low frequency exemplars revealed significant false positives. These results suggest that high frequency items are *not* coded phonologically, or at least, phonological coding is not necessarily the primary means by which high frequency items are read.

Other studies have obtained similar effects. In a lexical decision study, Waters and Seidenberg (1985) found significant effects of the regularity of a word (a hallmark of phonological coding), but only for low frequency items. High frequency items did not show a regularity effect. This could mean that the high frequency items are processed so rapidly that effects of regularity are washed out, but an interpretation consistent with Jared and Seidenberg (1991) is that the most frequent words are not read phonologically but by direct access to meaning.

The results of Lesch and Pollatsek (1993) and Lukatela and Turvey (1994b) provide evidence against this conclusion, however. Lesch and Pollatsek (1993) did not explicitly control for prime frequency in their study, however a post-hoc analysis of their data revealed no effect of prime frequency. Lukatela and Turvey did control for prime frequency by dividing their stimuli into two sublists: one in which the “appropriate” prime (e.g., BEACH for SAND) was higher in frequency than a homophonous distractor (BEECH), and a complementary list in which the “inappropriate” prime (e.g., PAUSE for CAT) was higher. Contrary to the expectations of the Jared and Seidenberg study, they found no effect of relative prime frequency; at short SOAs, the inappropriate prime would prime the target regardless of whether it was more or less frequent than the appropriate prime. Lukatela and Turvey interpret these results as showing that the Jared and Seidenberg study was incorrect: phonological codes are used in word identification regardless of frequency considerations. This interpretation will be explored in greater detail in Chapter 6.

A study by Strain, Patterson, and Seidenberg (1995) provided evidence for a frequency modulation of the DOL in a totally different domain: word naming. The Seidenberg and McClelland (1989) model mapped spelling forms onto phonology. It learned all items in its training corpus except for the most low frequency exception items. Presuming that these items are most difficult to learn by orth→phon, Strain et al. hypothesized that these items must be named by an alternate method; by semantics (orth→sem→phon). Accordingly, these items, and only these items, should show the effect of a semantic variable: imageability. These were the effects they found; a reliable effect of imageability on naming times only for low frequency exceptions. These results will be explored in detail in Chapter 7.

Clearly, intuitions and empirical evidence have yielded conflicting results. Minimally, it appears that the strong forms of either the all-phonology or no-phonology theories must be incorrect. The lexical decision studies consider frequency effects on the computation of semantics, and find that high frequency items are more reliant on orth→sem. The Strain et al. (1995) study looks at frequency and regularity effects on the computation of phonology, and found that low frequency exceptions are more reliant on orth→sem. Different models have considered this possible division in different ways, and will be reviewed next. Finally, a new theory that accounts for these effects, and accommodates the differing intuitions that fuel policy debates will be presented.

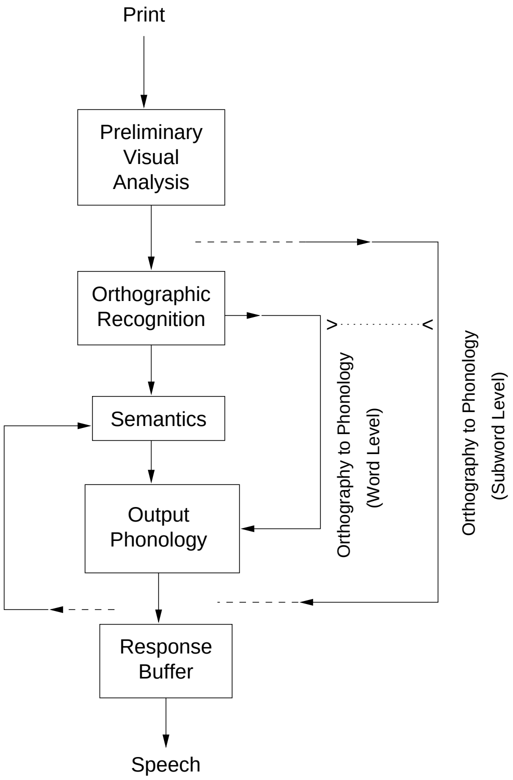

Print
Preliminary Visual Analysis
Orthographic Recognition
Semantics
Output Phonology
Response Buffer
Speech
Orthography to Phonology (Word Level)
Orthography to Phonology (Subword Level)

Figure 1.2: The classical dual route model (the "dust jacket model" from Patterson et al. (1985)).

1.2 Previous Models

1.2.1 The Classical Dual Route Model

The classical dual-route theory posits two distinct mechanisms for generating the pronunciation of a word: lexical, and non-lexical (Coltheart, Davelaar, Jonasson, & Besner, 1977). The non-lexical route is cast as a set of symbolic, frequency insensitive rules which map orthographic units ("graphemes") which correspond to a single sound onto that single sound. These grapheme-phoneme correspondence rules (GPC rules) are the only means by which the model can pronounce nonwords.

A secondary path exists for pronouncing words; the lexical path. Actually, the version presented in Patterson et al. (1985) contains *two* lexical paths, one semantic, and one that is lexical but not semantic (see Figure 1.2). The lexical routes are conceived as a lookup procedure, akin to a content addressable hash table. The orthographic form of a word is analyzed, and the "entry" for that word in the lexicon is accessed.

Different patterns of patient data can easily be accounted for within this framework. Surface dyslexics (Patterson, Marshall, & Coltheart, 1985) are patients who, following brain injury, exhibit a selective impairment in the naming of exception words; words whose spelling to sound correspondences are irregular. Hence, while GAVE would be read correctly by such a patient, HAVE would be read as rhyming with GAVE, SAVE, BRAVE. Such patients are interpreted as having lost access to their lexical store, and can only generate pronunciations by rule.

A complementary impairment, known as phonological dyslexia (Beauvois & Derouesné, 1979; Derouesné & Beauvois, 1979) involves preserved word naming, but grossly impaired nonword reading. These patients are posited to have lost access to their GPC rules as a result of their brain injuries, and can only read words by lexical access. Nonwords, lacking lexical entries, cannot be read.

The non-semantic, lexical pathway is posited to account for patients who have impaired or abolished nonword reading, normal word reading, and impaired semantics. The logic is that they cannot be reading by semantics because their semantics is impaired, and they cannot be reading by rules because their nonword reading is impaired, thus there must be a means by which the pronunciation of a word can be accessed, without recourse to the meaning of the word.

The model also accounts quite easily for the observed frequency by regularity interaction seen in naming studies (Seidenberg, Waters, Barnes, & Tanenhaus, 1984). The standard finding is that low frequency items are named more slowly than high frequency items. However, this is confined to exceptions; regulars do not show this effect. The dual route model accounts for this finding by positing that the GPC rules are frequency insensitive; regulars are read via these rules. The lexical route, in contrast, is frequency sensitive, so frequency effects are seen in exceptions because they can only be read by the lexical route.

The dual route model is really geared as an accounting of naming, not reading for meaning. However, within the framework (Figure 1.2), one can see how access to semantics would be accomplished. There are reciprocal paths from the pronunciation of a word back to semantics; hence, access to the semantic lexicon can be accomplished directly by print, or via phonological encoding. That phonological encoding can take the form of application of the GPC rules which generates a word's pronunciation, or lexical access to a word's pronunciation. This pronunciation then keys access to semantics. Alternatively, semantics can be accessed directly from print. Hence the dual route model admits two means of accessing meaning: via phonological mediation or via direct access.

The dual route model generates pronunciations via rule and via lexical lookup. For items which are not in the lexicon (e.g., nonwords), this procedure fails. However, the GPC rule application always generates a (regular) pronunciation for any stimuli applied to it. The model does not know in advance if an item is an exception or not. Hence, there must be some gating mechanism, or decision process which guides which route is the "winner." The classical dual route model was scant on details of this process (see Seidenberg, 1988).

Paap and Noel (1991) elaborate the notion of a time course of processing into the dual route model. They conceive the lexical and non-lexical routes as being engaged in a "horse race." The non-lexical route produces regular pronunciations for all words it is exposed to, even exceptions (e.g., pronouncing HAVE as /hev/, as in GAVE). For the system to produce the correct response, then lexical route must win the race.

The dual-route model thus explains the frequency by regularity interaction by claiming that regulars and exceptions are handled by different mechanisms. Put less charitably, the dual route model has not *explained* the frequency by regularity interaction, it has *implemented* it. Having explained the frequency by regularity interaction by stating that regulars are read along a frequency insensitive mechanism while exceptions are read along a frequency sensitive mechanism, we are left wondering why the GPC rule system (the "nonlexical horse," in the language of Paap & Noel,

1991) is frequency insensitive. The "explanation" simply begs the question it was purported to answer; this is because it is simply a mechanistic restatement of the phenomena and not an accounting of the phenomena through the application of deeper, independently motivated principles.

The situation is actually worse than this. Having observed the well established finding that the frequency of a word is a strong predictor of its naming time, Coltheart's implemented version of the dual route model (Figure 1.3; see Coltheart, Langdon, & Haller, 1996) pre-sets weights in the lexical access mechanism according to the frequency of the lexical item. The algorithm for processing a word then uses the value of these weights to determine the ramp-up speed for an item. As such, rather than explaining the effect of frequency on naming latency by presenting items to the network according to their frequency distribution, the model instead "explains" frequency effects simply by a brute computation statement of them.

Similar complaints can be raised for the patient data. Observing that some patients are impaired on nonword reading, while others are impaired on exception word reading, Coltheart and colleagues have constructed a model in which nonwords are read by one component and exceptions are read by another. Phonological dyslexics have an impairment here, surface dyslexics have an impairment there. This "explanation" leaves certain facts unaccounted for, however. For example, phonological dyslexics almost universally exhibit phonological impairments in addition to their nonword reading deficits. The dual route theory clearly places the locus of nonword naming impairments in a deficit in the GPC rules; it must treat these co-occurring phonological deficits as an accident of neuroanatomy.

In a similar vein, the dual route theory is at a loss to explain "mixed" cases; patients who are impaired (but not at zero) in nonword reading *and* exception word reading. The dual route theory must stipulate two independent sources of damage to account for these cases, which actually number the vast majority of the patients which have been examined. The literature relating phonological skill to reading acquisition is also problematic for this model; there is no sense in which either GPC rules or the lexical access mechanism are acquired through any process governed by phonological awareness.

Finally, the dual route model provides a poor account of graded effects of regularity. Words can be defined as regular or irregular, but within the set of regulars, there are words that are totally consistent (e.g., CAT) and words that are regular but inconsistent (e.g., GAVE, BRAVE, SAVE; contrast with HAVE). It has been found that inconsistent items are named more slowly than regulars, and more rapidly than exceptions (Taraban & McClelland, 1987). Because the dual route theory cleaves according to a strict notion of regular or irregular, it does not naturally distinguish these intermediary cases. Additional assumptions about the combining of "votes" from the two different mechanisms are necessary to account for these phenomena; the accounting is not at all a natural outcome of the processing framework. Rather elaborate specifications of the relative speeds of the different pathways, and careful tuning of parameters governing the speed of the GPC apparatus are necessary to make it all work right.

In short, the classical dual-route model has an elegant accounting for the broad characteristics of acquired dyslexias and regularity effects in naming. These elegant accounts are purchased at the cost of very inelegant explanations of the more subtle details of these phenomena.

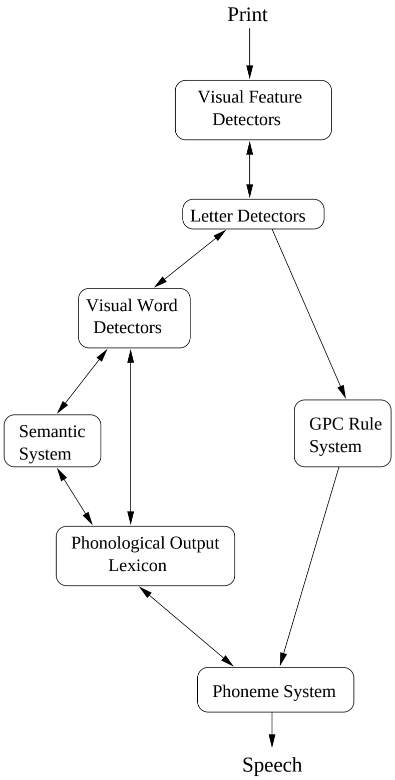


Figure 1.3: The DRC93 model of Coltheart, Curtis, Atkins, and Haller (1993).

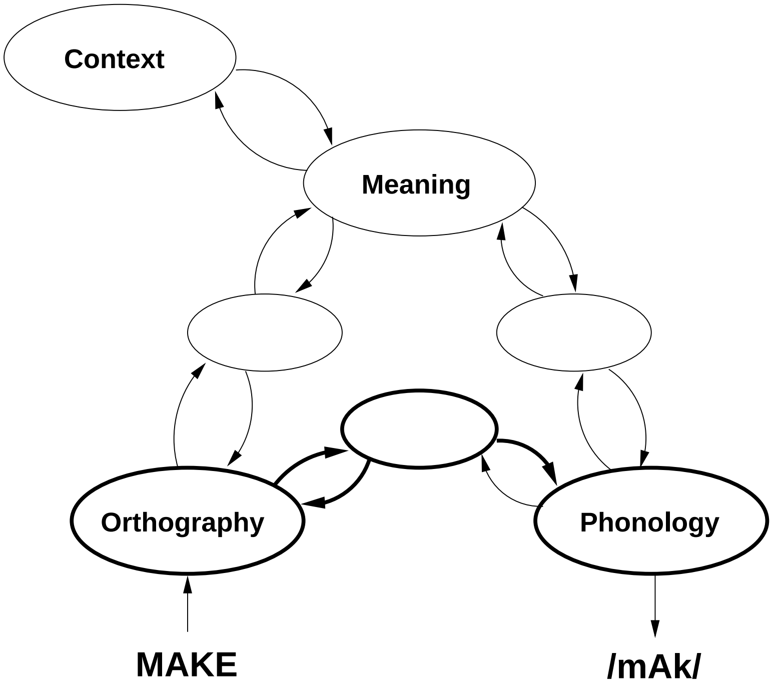


Figure 1.4: The SM89 model of Seidenberg and McClelland (1989). Implemented pathways are shown in bold.

1.2.2 Seidenberg and McClelland 1989

Seidenberg and McClelland's (1989) model (SM89) is shown in Figure 1.4. It represents a theory that is quite different in character from the dual route theory. Here, the notion of "lexical access" is de-emphasized; knowledge of words is not envisioned as a dictionarylike data structure which is accessed, but rather that phonological, orthographic and semantic patterns build up into a stable, coherent configuration. This is what a word is, in this model: a stable mapping between sound, spelling and meaning. Meanings and pronunciations are not accessed from a stored location but are computed on the basis of input, and context.

At first glance, the SM89 model has broad similarities with the classical dual-route model. There are, in essence, two ways to compute the pronunciation of a word; one through the translation of a word into its sound, and one through semantics. Alternatively, there are two ways to compute the meaning of a word; directly from its print, and via translation into its phonological pattern.

These similarities are superficial, however. The dual-route model has strong claims about the nature of the processing mechanisms that underlie these routes. In particular, the dual-route model is really a dual-mechanism model; it incorporates symbolic, frequency insensitive rules, and a lexical, frequency sensitive lookup mechanism. The SM89 model dispenses with these distinct mechanisms, positing instead that all processing is via weighted connections between representations. The SM89 model successfully replicated the frequency by regularity interaction; the computational framework used by SM89 was later used by Plaut, McClelland, Seidenberg, and Patterson (1996) to provide an explicit account of the phenomena in terms of nonlinearities in the network dynamics; the account falls naturally out of the framework which posits that regulars and exceptions are handled by the same computational mechanism subject to the same set of computational principles. Graded effects of regularity are also a natural consequence of this architecture, not requiring a host of ancillary assumptions about the precise timing of differing mechanisms. Further, the SM89 model accounts for frequency effects by presenting words to the model according to their (compressed) frequency distribution; see Plaut et al. (1996) and Harm and Seidenberg (1998) for similar. As such, not only is the character of the model quite different from the classical theory, but the character of the explanations are quite different as well. Explanations are couched in terms of computational principles rather than by architectural fiat.

The cost of such elegant explanations of the subtler aspects of word recognition phenomena has been less-than-obvious explanations of the broader patterns of impairment seen in the patient population. Initial attempts to simulate dyslexia in the SM89 model (Seidenberg & McClelland, 1989; Patterson, Seidenberg, & McClelland, 1989) were not terribly successful. It failed to produce regularization errors, which are the hallmark of fluent surface dyslexics (Patterson et al., 1985). Further, while accounting for mild cases such as patient MP (Bub, Chancelliere, & Kertesz, 1985), the model failed to produce dissociations as stark as that seen in patient KT (McCarthy & Warrington, 1986) (see Coltheart et al., 1993, for discussion). A full accounting of acquired surface dyslexia requires recourse to semantics, which the original SM89 model did not implement. Work in this vein will be discussed in Section 1.2.4. Similarly, the theoretical framework of the triangle model also requires appeal to semantic activation to account for acquired phonological dyslexia. Recently, developmental variants of these dyslexias have been accounted for within this general framework in a model that maps spelling patterns to sound (Harm & Seidenberg, 1998). Accounting for these patterns of impairment within the triangle model framework is an ongoing project and is far from complete.

As stated above, the implemented SM89 model (shown in Figure 1.4 in bold; see also Figure 1.5(a)) did not address issues of semantic activation, focusing instead on the computation of pronunciation directly from print. Computational principles and assumptions that derive from this framework will be detailed in Chapter 2.

1.2.3 Plaut and Shallice 1993

Plaut and Shallice (1993) implemented a model of deep dyslexia. Deep dyslexia is a condition caused by brain injury which is characterized by grossly impaired or abolished nonword reading, with “semantic” or “visual” errors in word reading (Coltheart, Patterson, & Marshall, 1980). An example of a “visual” error would be reading the word SYMPHONY as SYMPATHY. An example of a “semantic” error would be reading the word SYMPHONY as ORCHESTRA. Occasionally patients make both visual and semantic errors, reading SYMPATHY as ORCHESTRA. Other aspects of the performance of deep dyslexics will be considered in greater detail in Chapter 7.

The Plaut and Shallice (1993) sought to build a model which could account for both visual and semantic errors in naming. Noting that these patients typically cannot read nonwords, they left the orth→phon pathway unimplemented, presuming that the patients’ lesions render that pathway unusable. Figure 1.5(b) shows the implemented model. Plaut and Shallice then proceeded to account for a variety of impairment types in deep dyslexia, producing visual errors when lesions were introduced into the model before access to semantics, and semantic errors when lesions were introduced in the semantic attractor.

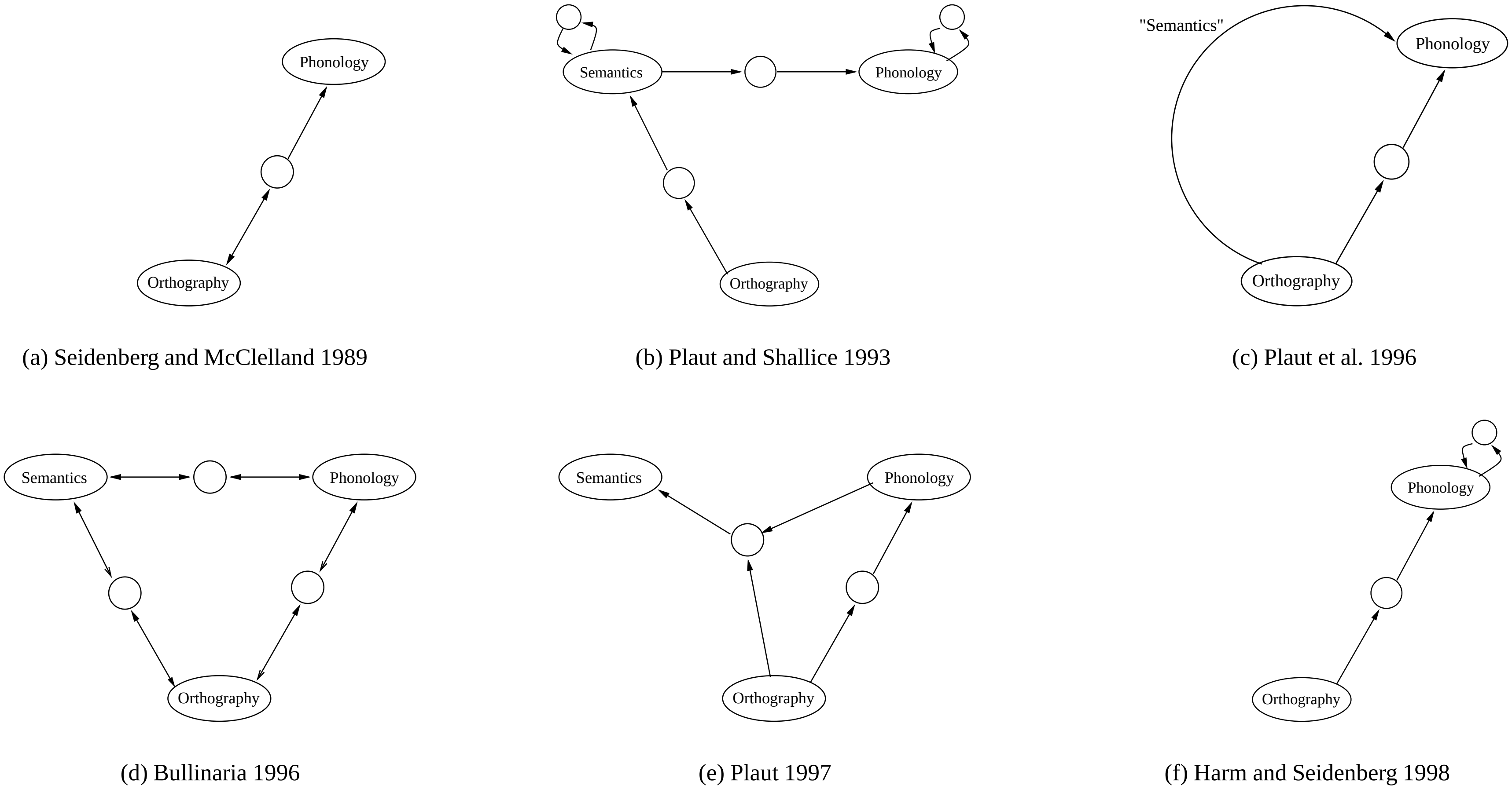
Attractor basins were used in the semantics system; such attractor dynamics were crucial to accounting for the differing forms of impairments (see Plaut, 1991, for more details).

The model did not really examine any kind of division of labor in the computation of phonology, because the orth→phon pathway was not implemented. The orth→sem→phon pathway was trained in the model to criterion, implicitly assuming that virtually all words can be read and pronounced aloud by this pathway, and no attempt was made to model any interactions of the learning of orth→sem→phon with the orth→phon pathway prior to the lesion.

1.2.4 Plaut et al. 1996: Naming

Plaut et al. (1996) (hereafter PMSP) implemented a series of models which extended the framework established by Seidenberg and McClelland (1989). The basic model was a feedforward model mapping orthographic forms onto phonology; its intent was to address the various criticisms of the SM89 model, such as its poor performance on nonword naming, and the account of acquired surface dyslexia.

A different phonological and orthographic coding scheme was used in this model, which provided for much better nonword performance. The original coding scheme for words in the SM89 model “dispersed” the regularities of the training set; items which were close neighbors in spelling and sound (e.g., CAT and HAT) were not sufficiently close to support adequate generalization to novel forms (e.g., GAT). The new scheme proposed by PMSP condensed these regularities by breaking words into clusters of units for the onset, vowel and coda. This scheme incorporated into the representation (and scoring method) knowledge of the legal sound and spelling patterns of English, and delimiting the representational space to reflect that knowledge. Phonotactic and orthographic rules (such as sonorancy rules: if /b/ and /l/ both occur in the onset, the ordering must



(a) Seidenberg and McClelland 1989

(b) Plaut and Shallice 1993

(c) Plaut et al. 1996

(d) Bullinaria 1996

(e) Plaut 1997

(f) Harm and Seidenberg 1998

Figure 1.5: Implemented connectionist models of visual word recognition.

be /bl/ and not /lb/) are imposed by fiat; if the model output a /b/ and /l/ in the onset then the output was interpreted as /bl/.

The model achieved nearly human levels of performance on standard nonword lists. To simulate differential patterns of surface dyslexia, however, required an interaction of semantics; reading via a strictly orth→phon route was not sufficient (see Chapter 7 for more details).

Plaut et al. simulated a developmental contribution of orth→sem→phon to reading by stipulating that during the development of the orth→phon route, the "semantic" route would contribute correct activation to phonological units as a function of the frequency of the word and the number of words the model had been exposed to (see Figure 1.5(c)). Equation 1.1 shows the contribution of semantics S based on the frequency of the word f and the number of sweeps through the training set t (each sweep corresponds to about 3000 word presentations. The model was trained for 2000 sweeps, or about 6 million word presentations. As such, over the course of development the orth→sem→phon pathway increases in competence, with high frequency items performing more accurately than low.

$$S = 5\frac{\log f + 2t}{\log f + 2t + 2000} \tag{1.1}$$

Plaut et al. also imposed mild weight decay on the orth→phon route. This impaired the ability of this route to learn all the items. This decay only applied to the orth→phon route in the model, and not the orth→sem→phon route. The result of this training regime is that low frequency exceptions could not be learned by the orth→phon route; they are the most difficult items and are hence most impacted by the effect of weight decay. By imposing damage on the semantic pathway they were able to simulate the effects of fluent surface dyslexia.

The model of course was not a true division of labor simulation; the division of labor was stipulated by fiat, via Equation 1.1. It was not a *finding* that high frequency items are read by orth→sem→phon better than low, but rather this assumption was built into the network's orth→sem→phon contribution. The learning of the orth→phon pathway was dependent on the orth→sem→phon contribution, because learning was error driven, and the more the orth→sem→phon pathway contributed, the less the orth→phon pathway had to. However the reciprocal was not true: the orth→sem→phon pathway learned according to Equation 1.1 regardless of the orth→phon pathway.

The PMSP model suggests that the "triangle" model formulation (Figure 1.4), in which words can be pronounced according to either orth→phon or orth→sem→phon can account for various forms of acquired dyslexia, and that the central tenets of the dual route theory are not necessary to account for patterns of impaired performance seen in surface dyslexia. However, the explanatory power of the model suffers from the ad-hoc way in which the semantic pathway was implemented. Ideally, one would want a unified set of computational principles to apply to all aspects of a model. By implementing the semantic pathway through a mechanism that is qualitatively different from the orth→phon one, it in fact has absorbed one of the tenets of dual route theory: different learning mechanisms for the two pathways to phonology. And, like the dual route model, the effect of frequency on the semantic pathway is not emergent from basic processing mechanism, but rather stipulated by design. The semantic pathway is in effect a *deus ex machina* which forces the correct activity on the phonological units via unspecified mechanisms.

Path	Accuracy
To Semantics	
Intact	89.1%
orth→sem	0.2%
orth→phon→sem	87.3%
To Phonology	
Intact	98.1%
orth→sem→phon	0.0%
orth→phon	97.5%

Table 1.1: Results from Bullinaria (1996).

1.2.5 Bullinaria 1996

There has been only one previous attempt to model the full triangle framework originally presented in SM89. Bullinaria (1996) presented a model (Figure 1.5(d)) which implemented all pathways between the three vertices of the triangle formed by orthographic, semantic and phonological knowledge. A training corpus of about 300 words was used, with random bits used to encode semantic features. The specific aim of this research was to explore the division of labor between the pathways.

Bullinaria trained the model in a stepwise fashion, first training the pre-literate parts of the model (phon→sem and sem→phon), and then using that network in a larger reading model (see also Seidenberg & Harm, 1995, for similar). Bullinaria then used lesion studies to explore the competence of the different pathways in the network.

The main results are shown in Table 1.1. The intact model is quite good at computing the phonological and semantic representations of words. The orth→phon pathway is quite efficient for computing the phonological form of words, and the orth→phon→sem pathway is quite good at computing the semantics of words. However, orth→sem is almost totally unable to compute the semantics for any word, and correspondingly orth→sem→phon is unable to produce any correct pronunciations.

Bullinaria presents this finding as a discovery about the nature of the reading system. The discovery, in a nutshell, is that words are read phonologically, with the direct orth→sem pathway contributing little.

This finding bears careful examination, however. Such a strong conclusion seems at odds with a host of behavioral studies. If the semantic pathway is incapable of supporting pronunciations, then why do deep dyslexic patients show semantic effects in their naming (reading SYMPHONY as ORCHESTRA)? Further, why do normal subjects show effects of a semantic variable, imageability, in their naming latencies for certain words (Strain et al., 1995)? The notion that orth→sem→phon is capable of reading some words is crucial for accounts of fluent and dysfluent surface dyslexia as well (Seidenberg, 1995; Plaut et al., 1996). If it were the case that orth→sem→phon was totally unable to read words on its own, then these findings and accounts of phenomena would all require a serious re-evaluation.

However, examination of Bullinaria's training regime reveals a potential source of this effect. The model was trained using backprop, an error correcting training regime (Rumelhart, Hinton, & Williams, 1986). Weight updates are driven by the error on output units. The model's sem→phon and phon→sem pathways were pretrained, as would be the case with a normal child who has acquired an auditory vocabulary previous to learning to read. In English, orth→phon is a very (though far from completely) regular mapping; far more regular than orth→sem. As such, any associative learning regime would find it far easier to learn. Hence, phon→sem is known to the model prior to reading, and orth→phon is much easier than orth→sem. The model, therefore, would learn orth→phon→sem far earlier than orth→sem. So what would drive the learning of orth→sem? Recall that an error correcting scheme was used. If orth→phon→sem faithfully drove semantic units to their correct activations, there would be no error in the semantic units, and hence nothing to drive the learning in orth→sem. The same applies for the division of labor to phonology: if orth→phon is learned rapidly, then there would be no error left to drive the learning of orth→sem→phon.

It is important that the processing assumptions used in Bullinaria's model be understood. The model was not placed under time pressure; semantic units were correct if either the orth→sem or orth→phon→sem routes ultimately activated them to their correct values. Bullinaria has implemented a theory of word recognition in which learning is error driven but there is no time pressure; the outcome of this theory is that orth→sem cannot learn. This result flies in the face of numerous empirical studies and behavioral results, and derives from processing assumptions that are very difficult to defend.

1.2.6 Plaut 1997: Lexical Decision

Plaut (1997) presented another model to account for lexical decision phenomena (Figure 1.5(e)). This model had only one route to semantics, via a set of hidden units which pool contributions both from orthographic and phonological units. The orthographic units feed into the phonological units as well. The model is a departure from the triangle framework proposed by SM89. Here, there is no pathway to semantics from orthography that is independent of phonological activation, and vice versa. There is, however, a path to phonology from orthography independent of semantic activation, and no pathway from semantics to phonology at all.

Plaut modeled lexical decision on the basis of a measure termed *stress*, which loosely means the extent to which semantic units are activated to extremal values. Formally:

$$S_j \quad = \quad s_j \log_2 s_j + (1 - s_j) \log_2 (1 - s_j) - \log_2 0.5 \qquad (1.2)$$

The stress of unit j, S_j is zero when the unit is totally at rest (output = 0.5), and increases as the activity of the unit approaches either 0.0 or 1.0. The overall stress for a word S is the sum of the stress of each of the j semantic units, $S = \sum_j S_j$. The idea was that actual words will have higher stress values than nonwords, and that measurements of the semantic stress would provide a basis on which lexical decision could be done.

By establishing a cutoff value of stress for which a decision of "word" versus "nonword," Plaut obtained a high degree of accuracy using standard lexical decision stimuli. Additionally, pseudohomophones (nonwords which sound like words, e.g. KAT) created slightly higher stress values than control nonwords, replicating a standard empirical finding that pseudohomophones produce longer latencies and more errors in lexical decision.

Problems with the simulations will be discussed more thoroughly in Chapter 5. Briefly, there are two major problems. Simulation work conducted by myself has suggested that when a semantic attractor (similar to that used in Plaut & Shallice, 1993) is used in semantics, then stress is no longer a usable measure, because the attractor tends to pull partial, non-extremal values to extremal states. As such, while stress is useful in a feedforward network without recurrence as in Plaut (1997), but a network with dynamic attractors in semantics (as in Plaut & Shallice, 1993) does not lead to interpretable results.

A second problem with the Plaut (1997) account of lexical decision is the pseudohomophone effect. As will be detailed in Chapter 5, pseudohomophones ought to provide very high levels of semantic activation. After all, if KAT activates the phonological form /kæt/, which reliably activates the semantics of [cat], then the only hope of disambiguating KAT from CAT is if the orth→sem activation of KAT is sufficiently disruptive, which it typically is not. How, then, does Plaut's model produce reliable differences between KAT and CAT? Plaut's (1997) model did not pretrain any phon→sem knowledge prior to the reading task; as such, every time the model was exposed to the phonological form /kæt/, it was *in the context* of the orthographic form CAT. As such, the common hidden units which map both orth→sem and phon→sem learn to depend on both sources of information. In a model which learned an auditory vocabulary prior to reading, however, this interdependency would probably not exist. If such a dependency did exist, then how would the model be able to account for *hearing*, that is, the mapping of phonological forms onto meaning in the absence of orthography?

These problems appear to apply in general to models which rely on the strength of semantic activity to support lexical decision. Lexical decision no doubt involves the use of multiple cues; sole reliance on the strength of semantic activation may work for a large set of word/nonword pairs but is probably not sufficient in general, and in particular for pseudohomophones. Intuitively, the rejection of pseudohomophones such as orthkat involve a form of spelling check; a sense in which the subject knows a word and is aware that the presented spelling is incorrect. This proposal will be worked out in greater detail in Chapter 5.

1.2.7 Harm and Seidenberg 1998: Naming

The Harm and Seidenberg model (Harm & Seidenberg, 1996, 1998), shown in Figure 1.5(f), added recurrence to the phonological units of an SM89-style model of naming. This created a phonological attractor which was trained prior to literacy training. The phonological attractor is similar in character and operation to the semantic attractor used by Plaut and Shallice (1993). While Plaut and Shallice varied the capacity and integrity of the semantic attractor to account for varying patterns of impairment in deep dyslexia, Harm and Seidenberg (1998) varied the capacity and integrity of the phonological attractor. These variations were related to behavioral phenomena in the development of literacy. Specifically, an impaired phonological attractor resulted in the model learning in a more item-specific manner: overlapping items were represented in a disparate manner, relative to models with normal phonological attractors. The nature of the attractor and its impact on learning were directly and causally related to the performance of the normal and impaired models.

1.3 Summary

The nature of reading poses two logical questions. First, to what extent does phonological information play a role in the computation of meaning? Secondly, to what extent does meaning play a role in the computation of phonology? These questions have been debated for well over a hundred years, with no clear resolution.

The intuitions of researchers and policy makers have not produced consistent answers; clearly, pure inductive logic is not the solution. Empirical studies of normal and brain damaged patients has yielded similarly inconsistent and contradictory results.

The modeling enterprise has not provided much help. The approach embodied in the dual route model is to *stipulate* the division of labor between phonological and non-phonological factors. This approach is really just a computational embodiment of introspection, not a discovery of any computational principles that explain the behavioral phenomena. Connectionist models hold the promise of the discovery of such computational principles, but here again the plate is rather empty. The SM89 model laid out a framework for the exploration of the division of labor within a connectionist framework. But this framework has not, until now, been realized. The history of connectionist models since SM89 has in fact been a series of partial implementations of the framework, and progress has not been additive. For example, Plaut and Shallice (1993) discovered that semantic attractors were crucial to explaining patterns of behavior in impaired populations. But no model of reading since then has utilized semantic attractors. In a similar vein, extensive empirical evidence exists for the role of phonological knowledge in the development of reading, but only one model (Harm & Seidenberg, 1998) has systematically looked at phonological knowledge. This model, in turn, did not have any semantic representations. Plaut's (1997) lexical decision model in fact bears little resemblance to the naming model discussed in the same paper. Bullinaria (1996) implemented all corners of the triangle model, but without attractors in semantics and phonology which are so clearly important, and with a training regime that is patently incorrect. As such, the results of this simulation attempt are largely uninterpretable.

Hence the current state of affairs is such that the SM89 model has provided a framework in which reading is viewed as the conjoined operation of multiple sources of information in an interactive, dynamical system. However, exactly how these information sources combine, and the factors relevant to the division of labor between them has yet to be investigated.

This thesis work represents the first serious attempt to combine the principles from all of these models in a full simulation of the division of labor to semantics and phonology. The general organization is as follows: Chapter 2 outlines the principles which guided the development of an explicit computational model of visual word recognition. Chapter 3 details a novel approach to deriving realistic semantic representations for a large set of words. Such representations are crucial to explaining the model's behavior. Chapter 4 provides details of the model's architecture and training regime. Chapter 5 outlines the division of labor to semantics, and relates the results to empirical studies. Chapter 6 explores the Lukatela and Turvey and Lesch and Pollatsek results and reconciles these results to the model's behavior. Finally, Chapter 7 explores the division of labor to phonology.

Chapter 2

A New Computational Model

2.1 Principles

There are a number of computational and environmental factors which motivate this new model of word recognition. These factors, their motivation, and their effect on the modeling enterprise, are considered here.

2.1.1 Differing Ease of the Mappings

In English, spelling and sound are highly correlated; spelling and meaning are not. Armed with the first consonant letter of a word one has a strong clue how the pronunciation of the word begins. One has no hint as to the meaning of the word. This makes the learning of orth→phon much easier than orth→sem.

There are exceptions to this tendency, of course. Morphological relations are cued by orthography, and in some cases are more reliable than the pronunciation. The regular past tense of a word can phonologically end in /d/, /t/ or /ʌd/, but invariably is spelled with -ED. The regular plural is the same, ending phonologically with /s/, /z/, or /ɛz/, but always being spelled ending with an S or ES. There are other regularities in spelling to meaning, some of which are more regular than spelling to sound (e.g., SIGN-SIGNATURE, or GHOST-GHAST-GHOUL, which have exceptional orth→phon correspondences, but quite overlapping orth→sem mappings).

The dimension along which the difficulty of orth→phon varies is spelling to sound regularity; by definition, a measure of the degree of predictability of the mapping from spelling to sound. Being a phonological factor, this does not relate to orth→sem. What does affect orth→sem are measures of regularity within that mapping. In this regard, while RAN is perfectly regular with respect to its pronunciation, it is an exception from the viewpoint of the orth→sem mapping (it is a past tense item which does not end in -ED). Hence, different words will vary in their difficulty with respect to the differing mappings, by dint of the nature of the mappings themselves.

The differing regularities of the mappings has an impact on the nature of the representations formed. Learning a correlated mapping produces broad attractors in mapping space, while learning totally uncorrelated mappings tends to produce point attractors. Hence, an input that is novel or new will produce a coherent output for a correlated network (e.g., the pronunciation of a nonword NUST), but not for an uncorrelated network (e.g., the semantic network activation for nonwords like NUST). The tendency of orth→sem to produce noisy, uninterpretable outputs has been used by Plaut (1997) in a simulation of the lexical decision task.

Because the correlated nature of orth→phon is assumed to be an important factor in the division of labor within the reading system, any model which is to accommodate this factor must contain enough words for there to actually *be* a correlated mapping. A network learning orth→phon relationships for a small number of words does not benefit from this correlated mapping nearly as much as it ought to. Serious consideration of this factor, therefore, requires any model to have a reasonably large number of items.

Orthographic, phonological and semantic representational schemes that preserve these facts about English will be used. Morphological regularities will be preserved in semantic, orthographic and phonological space, and the semantic representations will be created with an eye to variability in the arbitrariness of the mappings (Chapter 3 will cover this in detail).

2.1.2 Pre-existing Knowledge

Children learn to read armed with a great deal of prior knowledge. They possess a large auditory vocabulary prior to learning to read. The mapping of sound to meaning is as difficult and uncorrelated as that of spelling to meaning, however when learning to read, this knowledge is already in place for a child to make potential use of. Similarly, children begin reading armed with a structured knowledge of the sound patterns of their language, and of the semantic composition of the world. It has been hypothesized that these existing sources of knowledge can be used in the development of the reading skill (Jorm & Share, 1983; Share, 1995). Additionally, the fact that phon→sem is a largely uncorrelated mapping is offset in part by the fact that this knowledge is largely in place before learning to read. The computation of meaning via orth→phon→sem is, in principle, just as difficult as via orth→sem; what makes phon→sem different from orth→sem with respect to the development of reading is the fact that phon→sem is known for a large vocabulary of words prior to reading training.

The model will be trained in a manner that is broadly faithful to this constraint: the phon→sem and sem→phon mappings, as well as semantic and phonological attractors, will be pre-trained prior to their use in reading.

2.1.3 Integrative Dynamics

Representations are computed, not accessed. There is not an instantaneous moment that units spring to life. Rather, unit activity builds over time with a speed proportionate to the strength of the unit's input. A unit's activity builds more rapidly with stronger input, or convergent input from multiple sources. This is an important theoretical point: the notion of "lexical access" has been a dominant concept in the psychology of reading for decades, yet no neurally plausible mechanism for accessing a dictionarylike mental structure, with all its ancillary knowledge, has ever been proposed. Further, it has been found that partial activation of semantic information of a word can occur depending on the context. For example, the fact that pianos are heavy would be activated in a moving context, but not in a recital context (Barclay, Bransford, Franks, McCarrell, & Nitsch, 1974). Similarly, when experiencing tip of the tongue phenomena, subjects sometimes report knowing the initial phoneme of a word but cannot access the remainder of its phonological representation. Such phenomena suggest that an atomic moment of "access" is not appropriate, and that the graded development of relevant features over time is a better metaphor (Balota, 1990; Seidenberg, 1990).

Distributed semantic representations will be used, in which the meaning of a word is viewed as an activation vector in semantic space. A continuous time version of backprop will be used, which has the property of building activity in units over time as a function of the strength of the input to the unit (see Chapter 4 for details). Because activity in a model accumulates over time, an approximation of reaction times (RTs) in empirical studies could be created. Simulations in Chapter 7 will use the time course of processing to model reaction times from empirical studies.

2.1.4 Cost of Computing Intermediate Representations

Mapping from one representation directly onto another one is (generally) more rapid than mapping a representation onto another through an intermediate representation. Specifically, mapping from spelling to meaning via phonology involves computing a reasonably stable intermediate phonological representation. This representation does not necessarily have to be the final, veridical representation to cause semantic activation, but needs to be sufficient to activate a reasonable semantic representation.

By using time varying networks, there is a time course of activation such that activity does not instantaneously propagate through the network but rather integrates up over time. In such networks, in contrast to simple feedforward nets, an intermediate representation has to build up in order for it to influence a final representation. See Chapter 4 for discussion of the need for time varying networks and the inadequacy of feedforward networks for implementing this theoretical constraint.

2.1.5 Attractor Basins and Dynamical Systems

Plaut and Shallice (1993) made extensive use of attractor networks in semantics to account for varying patterns of performance of patients with deep dyslexia. Harm and Seidenberg (1998) made use of attractor networks in phonology to account for various patterns of behavior observed in developmental phonological dyslexia. These principles have been used profitably in other domains of connectionist psychology (e.g. Devlin, Gonnerman, Andersen, & Seidenberg, 1998; Allen & Seidenberg, In Press; Joanisse & Seidenberg, 1998).

Attractor basins are important because their existence influences the nature of learning accomplished by a system that maps onto them (Harm & Seidenberg, 1998). When an attractor system is available to repair partial or noisy patterns, a system that maps representations onto that attractor space is relieved of pressure to produce a precise, exact output. It makes the goalposts wider.

The use of attractor basins and recurrence in the reading system also adds a time-varying component to processing; the network can change its state in response to its own state, as well as to external input. Such a system forms dynamical systems; systems whose state varies over time in complex ways. This will be discussed in greater detail in Chapter 4.

The quality and nature of the attractor basins are an important factor for the reading system's dynamics. Devlin et al. (1998) proposed that variation in different items' semantic properties (e.g., artifacts such as tools, versus natural kinds such as fruits) are important in accounting for patient data regarding differential performance on these items. These semantic properties influence the ability of the semantic attractor to boost noisy or degraded patterns. In a similar vein, Plaut and Shallice (1991) explored the effect of a word's abstractness or concreteness (e.g., TRUTH versus BRICK) in explaining deep dyslexic patients' patterns of responses. Again, the network performance is interpreted in terms of attractor basins within semantics. Further constraints on the formation of semantic attractors are discussed in Chapter 3.

These data, coupled with the developmental phonological dyslexia simulations, provide strong evidence for the importance of attractor basins as an explanatory tool in the modeling enterprise.

2.1.6 Greed is Good

It is assumed here that the reading system is under constant pressure to perform rapidly; as rapidly as possible. It is a standard assumption in eye-tracking paradigms that the limiting factor on reading speed is not the physical process of an eye saccad, but rather the cognitive process underlying the comprehension of text. Indeed, if the reverse would be true, eye tracking data would be totally uninteresting; it would be a measure of oculomotor speed independent of the text before the subject. As such, the model should not only be driven by the need to be asymptotically accurate, but to recognize a word as rapidly as possible using all available information. This tenet, combined with principle 2.1.3, results in a system that is *greedy*; it demands activation from all available sources to the maximum degree possible.

This constraint is operationalized by penalizing the network not only for producing incorrect responses, but for being slow; error is injected into the network early in processing to encourage the quick ramp up of activity (contrast with the discussion of Bullinaria, 1996 on page 14).

2.1.7 Error Driven Learning

This is perhaps the most tendentious of the assumptions for this framework. Here we assume that learning is in part at least error driven. Operationally, what this means is that the intact system is under pressure to produce a correct output by whatever means necessary. If one component of the system (e.g., orth→phon→sem) fails or is slow for a given item, there is implicit pressure on the redundant computations (here, orth→sem) to make up the difference.

There are, of course, learning mechanisms that are strictly correlative, and not driven by error; the classic example being Hebb (1949). In such a system, learning of an item by one component (again, for example, orth→sem) would be totally independent of the success or failure of orth→phon→sem for that item. However, it will be shown in subsequent sections that sensitivity to the success or failure of other parts of the system is important, and plays a crucial role in accounting for more subtle patterns of performance.

Error driven learning methods such as backpropagation of error (Rumelhart et al., 1986) have been criticized for lacking biological plausibility. These criticisms will be addressed in Chapter 4.

2.2 The Focus of the Research

Much of the debate within the field of education has been characterized by the question: orth→phon→sem or orth→sem? Do we teach children to read using phonics-based methods, or do we emphasize the direct translation of print to meaning?

Within the dual-route model, their central research question has been: which words go by which route? What are the speeds of the two routes? Which route wins, and under what circumstances? How do we characterize reading impairments in terms of damage to the two routes?

Here, the emphasis is different. The research question is: what are the properties of the system, the words and the task which mitigate the relative contributions of the two paths? What abstract computational principles drive the development of the two routes, their use in skilled reading, and their behavior in the face of damage?

Chapter 3

Semantic Representations

3.1 The Problem With Meaning

Stan: Cartman doesn't know a rainforest from a poptart!
Cartman: Yes I do! Poptarts are frosted!
– South Park

Semantics is hard. While the modeling of phonology poses interesting and difficult challenges (e.g., the "slot problem"), there is broad agreement on the general composition of phonetic features. If one chooses to represent a phonological form according to binary features, you can consult a textbook on phonology to find matrices detailing the feature values for the target language's phonemes. The set of features used to encode a phoneme varies according to theory, but is generally on the order of 10-30 features; always finite and easily enumerable. The set of actual phonemes in an inventory additionally varies according to theory, but is typically on the order of 20-50.

Contrast this with semantics. While phonemes can be described with, say, 22 binary phonetic features, how many *semantic primitives* are there? And while the number of phonemes is limited and bounded, how many *concepts* are there? Its easy to state that /p/ and /b/ are similar yet distinct phonemes, and state the ways in which they are distinct. A well-studied articulatory apparatus must realize the differences between these phonemes. But what of the concepts [rock] and [stone]? They are clearly not identical, but how are they different? And while /p/ and /b/ differ in the same way as /f/ and /v/ (voice onset time), do [rock] and [stone] differ in the same way as [jem] and [jewel]? Conceptual structure is not constrained by a physical articulatory mechanism, as phonology is, so direct study is not possible, and the problem space is far less bounded.

In this section I will first outline some desiderata for a semantic representation system for use in connectionist modeling. I will then review methods used in previous work. I will show that while all previous methods have advantages, they also are inadequate for this research. I will then present a novel approach to generating large corpora of recognizable semantic representations that satisfy the design constraints.

3.2 Design Constraints

There are a number of desirable properties for a semantic system to have.

1. Scalability of effort. Ideally, the effort necessary to generate semantic representations for a large set of words should not be prohibitive.

2. Scalability of size. The number of units to encode n words should scale no faster than $O(n)$.[1] This is to keep the size of the simulations manageable.

3. Neural Implementation. For the purposes of connectionist modeling, the semantic representation ought to map in a simple way onto neuron-like units whose state is simply a level of activation.

4. Distributed Representation. A distributed representation should be used, allowing varying degrees of overlap between the representation of different concepts.

5. Transparency of Feature Meanings. Ideally, the representation for a word would be composed of units whose meaning is interpretable. If people's common sense intuition is that the meaning of [dog] is composed of properties such as <*barks*>, <*has-fur*>, <*animal*> and such, then the features representing [dog] ought to share some reasonable, identifiable overlap with these features.

6. Transparency of relationship with other words. Words that people intuitively feel are related to each other should have representations that overlap; those that are distant should have less overlap. For example, [dog] and [cat] should overlap in representation, more with [puppy] and less with [truck] or [truth].

7. Systematicity of morphological relationships. One (new and controversial) approach to the study of morphology relationships is that the system of language that is called morphology is really not an autonomous module of the language apparatus, but instead is emergent from strong overlaps in meaning coupled with corresponding, quasi-regular overlaps in sound (Gonnerman, Devlin, Andersen, & Seidenberg, 1995). The idea, characterized grossly, is that there is a common phonological relationship between CAT and CATS, and that the phonological difference between these words is very similar to the difference between PUFF and PUFFS. Additionally, the semantic difference between [cat] and [cats] overlaps or shares recognizable properties with the semantic difference between [puff] and [puffs]. Hence, what can be described as a system of rules that transforms a stem into a plural through the addition of an /s/ phoneme (e.g. Pinker, 1991), can instead by characterized as generalization within a quasi-regular domain, sharing many computational principles as reading and other productive tasks.

8. Inheritance. Semantic relations are clearly not a strict hierarchy; many properties cut across instances (e.g., both a school bus and a banana have the property <*yellow*>). Nonetheless, there is a sense that some semantic relations are inherited. A beagle, for instance, is a type of dog, sharing properties that can be assigned to dogs. A beagle can be thought of as a dog, with all the normal properties of a dog, but with additional defining features.

This does not need to incorporate strict, object oriented programming notions of inheritance, with all the ancillary assumptions about information encapsulation that can go with it. Further, the idea is that there is not a strict distinction between categories and types. Here, the idea is simply that a set of features define the concept of [animalhood]. Those features are, by definition, common to animals.[2] The concept [cat] contains the features of [animal], plus some additional

[1] The notation $O(f(n))$ means that asymptotically, the function scales no faster than a constant multiple of $f(n)$. Hence $O(n)$ means the function scales linearly with n.

[2] In WordNet, the concept [animal] is actually a very high level concept; reptiles are kinds of animals, as are mammals, fish and insects. The concept [animal] has has-part relations containing features such as <*head*>, <*limb*>,

features, which may or may not overlap with other concepts. The concept [calico_cat] contains features of [cat], plus additional features. In a sense, [animal] is a category; an abstraction derived from its members. But then so is [cat]. Ideally, some features would be probabilistic: cats typically have claws, but a cat that has been de-clawed is still a cat. The use of probabilistic features would naturally lead to *fuzzy* categories, where membership in a category is a more stochastic decision process than a strict aspect of the architecture.

9. Variation in correlations between features. There is extensive neuropsychological evidence that the density of a concept's semantic attractor is an important factor in accounting for normal and disordered word recognition. Patients with deep dyslexia tend to be better able to name concrete words such as BRICK than abstract words like TRUTH. Word class effects are also observed: nouns are read more easily than verbs, which are read more easily than adjectives, which in turn are read better than closed class words (Coltheart et al., 1980). Similar effects have been found in populations of acquired phonological dyslexics. Friedman (1996) analyzed the results of 11 phonological dyslexic patients. Of the ten patients that were tested for noun versus closed class word reading, six exhibited a deficit in closed class words relative to nouns. Of those six, four also exhibited impaired verb reading relative to noun reading, and of those four, two also showed an advantage of concrete nouns over abstract nouns. Friedman proposed that there is a predictable succession of impairment characteristics: the difference between nouns and closed class words is greatest, so most subjects show that effect. The difference between nouns and verbs is smaller, and the difference between concrete and abstract nouns smaller still. Patterson, Suzuki, and Wydell (1996) argue that it is variability in the richness of the semantic representations of the different types of words that accounts for these effects (see also Plaut & Shallice, 1993, 1991).

Providing further evidence for the importance of intercorrelated features, patients with Alzheimer's disease have exhibited differential performance in naming artifacts versus natural kinds. Differences in the semantic properties of these classes of items is part of one account of these effects (Gonnerman, Andersen, Devlin, Kempler, & Seidenberg, 1997); natural kinds have more densely intercorrelated features, which can support concepts in the face of low levels of damage.

Normal subjects exhibit evidence for an influence of semantic factors as well. In a careful study, Strain et al. (1995) found effects of imageability, a semantic factor, on the naming of low frequency exception words. Words with high imageability were named more rapidly than words with lower imageability. This effect is interpreted as showing the influence of the attractor depth on the formation of a stable semantic pattern: items are named faster because the item's semantic pattern is formed more rapidly. McRae, de Sa, and Seidenberg (1997) found that for natural kinds (but not artifacts), the intercorrelatedness of semantic features, rather than featural overlap, was a reliable predictor of subjects' latencies in an online semantic categorization task. McRae et al. (1997) simulated their empirical results with an attractor network which is sensitive to the degree of correlational overlap between features.

The common thread of these different phenomena is the strength of the semantic attractor for a word. The imageability effects are standardly not interpreted as reflecting imageability per se, but rather the strength of the semantic attractor. Concreteness effects and word class effects are

<*face*> and so forth. The subclass [mammal] further includes features such as <*hair*>. The current project does not utilize these has-part relations, instead treating nodes such as [animal] as placeholders for the relevant features, but future work could utilize this information.

interpreted as evidence of the richness and strength of the semantic representation of different words; closed class words have weaker semantics than nouns, abstract nouns are weaker than concrete nouns (Friedman, 1996; Patterson et al., 1996; Plaut & Shallice, 1991, 1993).

The explicit claim here and elsewhere is that the intercorrelatedness of semantic features contributes to the richness and strength of the semantic attractors. This richness affects the speed of responses in normals, and the resilience to damage in brain damaged patients. Such findings can and should guide the methods employed for developing of semantic representations.

3.3 Previous Methods

Connectionist models have utilized various methods to encode semantics. In this section I provide a description and critique of each of them.

3.3.1 Random Features

Random features are the most obvious way to encode semantic relationships. In this scheme, features are assigned values of on or off for each word based on some random algorithm. Plaut and Shallice (1993) utilized a system of quasi-random feature assignments. It was not totally random, as "pools" were created corresponding to different categories, though the actual meanings of features were opaque and randomly assigned. This scheme has the advantage of being trivially scalable. Additionally, one can algorithmically vary properties of the semantic representations along dimensions such as feature intercorrelatedness and number of features per item. Hierarchical relationships could also be enforced algorithmically. The drawback is that the correspondences between feature values and actual meaning are totally arbitrary; the semantic features for [dog] are no more likely to overlap with [puppy] than with [tire-iron] or [truth]. Hence, it would be impossible to match the results of existing priming studies with the same items. Additionally, subtle correlations that exist in the world are missed; for example, closed class words tend to be short in length and high in frequency. Such properties must be explicitly coded in. Additionally, there are phonological cues to a word's status which one would want to code in: a voiced initial TH provides a cue to open versus closed class status (compare THIN, THICK, THIEF with THUS, THE, THEM, THERE). There are voicing differences for noun/verb distinctions, e.g., the noun versus verb usages of words like HOUSE, USE, ABUSE or WREATH/WREATHE, TEETH/TEETHE. People are aware on some level of these probabilistic regularities and are able to make use of them (see Kelly, 1992, for review). Further regularities exist that are broken by random pattern matching. For example, systematic morphological regularities would not exist unless one coded items as <*plural*>, etc. by hand. Lastly, the construction of stimuli for controlled experiments (e.g., does TOWED prime FROG) would be very difficult when the semantic overlap between items is totally arbitrary.

3.3.2 Feature Norms

Another method used in the literature (e.g. Devlin et al., 1998) is feature norms. Subjects are provided with a list of items and asked to write down things they know about the items. The most common features of the provided items are then treated as binary features and used as the representation for the item.

In a sense, this method is the polar opposite of the random features method. It has good face validity; the semantic features clearly correspond to the items in question, and items that intuitively ought to overlap generally tend to. Because the representation is nominally faithful to actual semantic representations, subtle regularities such as effects of word length and frequency on semantics are preserved.

One problem with this method is that it is very difficult to generate semantic representations for a large set of words. Collecting the feature norms from subjects is labor intensive both for researchers and for the subjects, and places a practical limit on the number of items which can usefully be coded in a reasonable amount of time.

An additional problem is that the features one gets tend to not include hierarchical relationships. People certainly know that one thing that [builder] and [baker] have in common is that they are both <*causers*> (in contrast with, for example, [thinker]). When a subject is asked to define a concept such as [cat], the implicit assumption is that definitions are sought which contrast [cat] with other items in that same category. Hence, features tend to focus on in-category distinctions. People do not tend to provide as a semantic feature for [cat] the fact that it is a [living thing], or for that matter, a [thing]. Such distinctions could be elicited through varying contrasts (e.g., rather than asking a subject what a [cat] is, ask them how [cat] differs from [truth], or [brick], or [swim], or [dog]). But such contrastive pairings would make the scalability of the method even worse.

Lastly, morphological relationships and entailment relationships may not be detected. Asking a subject to define [bake] and another subject to define [baker] may not yield any overlapping features; the first may refer to heating food while the second may refer to perceptual features such as being in a kitchen and wearing an apron. One could not rely upon obtaining the notion that [baker] entails the concept [bake], plus a <*causer*> relation.

3.3.3 Hand Coded Features

This method is very similar to the feature norm method. Rather than asking subjects to write down features of items, the experimenter simply does so. This method shares many of the benefits of the feature norm method, while also allowing the experimenter to explicitly add in hierarchical and morphological relationships and enforce a consistency among them.

The problem is that the method scales terribly. Hand coding the semantics for 5,000 monosyllables is prohibitively time consuming.

3.3.4 Co-occurrence Statistics

This final method that has been used (Lund, Burgess, & Atchley, 1995) is an interesting departure from previous methods. Here, a very large corpora (on the order of 400+ million words) is scanned. The co-occurrence statistics are computed for words. This produces a large (70000^2 cell) matrix of distances. The rows and columns of this matrix are individual words, and the cells are the co-occurrences between those words in the corpus. Hence, the co-occurrence of CAR and ROAD would be high; the co-occurrence between CAR and CLAM would be quite low. Mathematical techniques are used to reduce the dimensionality of this large matrix into a smaller set of units (say, 200) which preserve the general topology of the matrix. The representation of words is then coded from these units. It has been found that these units preserve intuitive semantic relationships including syntactic distinctions and semantic category distinctions (Lund et al., 1995).

One problem with this method is that none of the features have a particular meaning. One cannot point to unit 132, for instance, and say "oh, that means concept [xyz]." The units are a compressed 200 unit representation of a weighted 10 word window over 70000^2 cells of a matrix; it would be very surprising to be able to infer clear semantic features such as <*has-fur*> or <*pounds-nails*> out of an individual dimension. The data reduction techniques would spread such regularities across multiple features, making them very difficult if not impossible to interpret.

Further, it is unclear that morphological regularities would be enforced. While it is granted that BAKE and BAKER would overlap, and BUILD and BUILDER would overlap, what is questionable is that the features that distinguish BAKE from BAKER would overlap in a systematic way with those that also distinguish BUILD from BUILDER. If one wishes to characterize morphology as systematic overlaps in meaning cued by systematic phonological overlap in sound, one would clearly want this property. You cannot have systematic overlaps in the mapping of sound to meaning if you do not have systematic overlaps in meaning to begin with. Put differently, if we wish to characterize the morphological transformation of BUILD to BUILDER as a semantic/phonological regularity, then the translation of BAKE to BAKER ought to exhibit very similar regularities. Since the phonological transform is the same, the semantic transform therefore needs to be at least very similar. Otherwise, what is there to generalize from, if one wanted the system to be able to generalize GLORPER from GLORP?

The problem is, why would such similarities in overlap obtain in the co-occurrence matrix? Exactly which lexical items would differentiate BAKE from BAKER and also distinguish BUILD from BUILDER and TAKE from TAKER, RAKE from RAKER and so forth? It is not the surrounding lexical environment that provides this overlap, but the deeper meaning of these words. But the *meaning* of words is not used as input to the matrices, that is considered the *output* of the system!

The importance of quasi systematic overlap in deeper meaning is illustrated by Levin (1993). She enumerates a large list of verbs and their syntactic alternations, and concludes that the underlying meaning of the verbs is fundamentally what drives most of their syntactic behaviors. For example, consider the verbs CUT and BREAK (see Levin, 1993, p. 6). Both verbs can be used transitively (*Margaret cut the bread*, *Paul broke the window*), in diathesis alternations (*The bread cuts easily*, *The window breaks easily*), but only CUT can be used in a conative construction (contrast *Margaret cut at the bread*, versus *Paul broke at the window*). The explanation is semantic: CUT does not imply the action was completed, while BREAK does. Items that imply completion cannot be used in the conative construction, which inherently implies partial success. This generalization applies for a large class of CUT verbs (CUT, HACK, SAW, SLASH . . .) and BREAK verbs (BREAK, CRACK, RIP, SHATTER . . .). The co-occurrence statistics method would correctly categorize these verbs into differing pools on the basis of their different lexical environments (only the CUT-verbs appear before an AT, so their matrix entries would differ from the BREAK verbs). So we would in fact have different classes. But this would get the causal relationship identified by Levin backward: it is not that these words mean different things because some of them have the word AT following them; it is rather the case that some of them imply completion, and *that* is what governs whether they can appear in the conative construction (e.g., with an AT following them).[3]

[3] This is not to say that such local lexical constraints have no role in acquisition; on the contrary it is quite likely that the environment which words appear in is an important source of information as to their lexical properties. The

When a person learns a new word, the semantics of that word cue the syntactic environments it can appear in. If a child is shown a picture of a person, say, snapping a stick in half, and is told that the person is FLOOBING the stick, the child would know that FLOOB is semantically in the category of BREAK verbs, and as such one could not FLOOB at a stick. But if we take co-occurrence statistics to *be* meaning, then how could a novel item be provided with a semantic cue that the operation involves completion? Again, one would have to inform the system that FLOOB was a BREAK verb by appeal to its lexical environment, but that is exactly what semantics itself is supposed to inform! Again, the operation would work, but in the reverse direction from what it should.

Perhaps if the algorithm were recursive, i.e. rough computations of meaning made by co-occurrence statistics of lexical items, then further refinements generated based on the co-occurrence of those derived rough meanings, then the method would hold greater promise for such deeper relationships such as [causer], [plural], [implies-completion] and such.

The point here is not that the method is worthless. On the contrary, the co-occurrence statistics method generates plausible semantic representations for broad classes of uninflected items: ROAD and STREET are very similar in representational space, presumably due to the common occurrence of CAR, BUS and such in their environment; CAT and DOG overlap, again probably due to the co-occurrence of words like PET, FUR, CLAWS, VETERINARIAN and FLEAS. Inflected items like CAT and CATS would be quite similar as well. The morphological problem, in a nutshell, is that it seems very unlikely that a simple 10 word window of lexical items would make the feature(s) that distinguish CAT from CATS overlap at all with those that distinguish, for example, LOAF from LOAVES. Lexical relationships should be learned by the system, it should not be the target state to which phonological forms are trained to.

3.4 A Different Method

This project demanded a new method for generating semantic representations for a large set of words.

The method will be outlined in detail below. Here is a general overview. A large set of monosyllables was collected from the CELEX corpus (Baayen, Piepenbrock, & van Rijn, 1993). The Francis and Kuçera (1982) corpus was used to identify the most frequent sense of a word (e.g., verb, noun or adjective). The WordNet project (Miller, 1990) was then used to extract meanings for the nouns and verbs in the corpus. Inflected past tense and plural items were detected by WordNet and appropriate semantic features were algorithmically added. Adjectives and adverbs, which were far fewer in number than nouns and verbs, were coded by hand, along with closed class items. The result was a corpus of over 5,500 words, each coded with semantic features from a pool of about 1,600 features.

problem is that the co-occurrence statistics do not treat lexical statistics as a *cue* to some ultimately derived meaning of words, but rather as *meaning* itself. Bluntly, it is not that the presence of AT *cues* the system that CUT and BREAK have different meanings, but rather that this method implicitly claims that the lexical environmental difference *is* the difference in meaning.

3.4.1 Generating Part of Speech

The CELEX electronic corpus was used to identify a set of 8,500 monosyllables. These monosyllables were used as initial input into the system.

These items were matched against the Brown corpusFrancis and Kuçera (1982). This corpus contains about 55,000 English word forms with a breakdown of usage frequencies. For example, it shows that HIT is used as a noun 26 times, and as a verb 126 times. This corpora was processed to produce a listing of the single most frequent sense of each orthographic form in the corpus (e.g., HIT is most often used as a verb, not a noun). Words from the initial list were matched against this corpus to determine their most common part of speech (noun, verb, adjective, adverb, closed class). This resulted in a list of 5,600 words. The remainder were not coded in the Brown corpus.

The 3,000 items which were unknown to the Brown corpus were tested against the WordNet corpora to determine if they were more often a noun, verb, morphological variation on a noun or verb (e.g., AWLS) or other. Items that were nouns or verbs were marked as such and added to the list of items whose status was determined by the Brown corpus. This resulted in a total list of 6,500 words coded for their most frequent part of speech.

Words that were coded as being morphological variants on a root word were removed at this point. This resulted in 3,600 words remaining. At this point the list of words was split into four lists. One contained nouns, one verbs, one adjectives, and one other, consisting of adverbs and closed class words.

The 250 adjectives were hand coded according to a taxonomy of features inspired by Frawley (1992). Modifiers were broken down into a hierarchy of mental properties, physical properties, quality/value properties and spatial properties. These were each broken down further: for example mental properties could be social properties properties (e.g. *friendly, mean*), mental (e.g., *smart, dumb*) or dispositional (e.g., *careful, careless*). Physical properties were similarly broken down into those describing temperature, size, weight, or color. As before, <*bad*> or <*negative*> features were used to describe negation (such as the difference between [smart] and [dumb], [bright] and [dark]).

The 127 closed class items were also hand coded, using grammatical and pragmatic features such as <*determiner*>, <*definite*>, <*singular/plural*>, <*disjunction*>, <*article*>, <*interrogative*>, <*male/female*>. Table 3.1 shows a sample of typical items and their semantic coding.

The remaining 3,200 uninflected items (2,200 nouns and 1,000 verbs) comprise the bulk of the uninflected monosyllables. The method for deriving features for them is now described.

3.4.2 WordNet Data Structures

WordNet's noun database consists of about 34,000 word forms organized into a set of approximately 29,000 noun SYNSETS, or synonym sets. Each synset consists of a set of words that are taken, at WordNet's level of granularity, to be synonyms (e.g., [cur] and [mutt] are synonyms in WordNet). Here, the notion of word refers to a particular sense of a word; different senses have different entries in different synsets. There is an index of lexical forms, where each lexical form contains a sequence of pointers to synsets containing different senses of the word. The lexical form DOG has pointers to synset entries meaning a canine, an ugly person, a scoundrel, and the thing in the fireplace that you put logs on. These entries are sorted by order of frequency, so the first synset for a given lexical entry is the most common one (generally; there are odd cases where this doesn't

Item	Features
A	*<singular, indefinite, determiner>*
THE	*<singular, definite, determiner>*
HOW	*<interjection, puzzlement>*
WOW	*<interjection, astonishment>*
YES	*<interjection, agreement>*
YEAH	*<interjection, agreement, slang>*
AYE	*<interjection, agreement, archaic>*
DANG	*<interjection, negative>*
THEIR	*<possessive, pronoun, 3rd_person>*
OUR	*<possessive, pronoun, reflexive>*
YOU	*<nominative, pronoun, 2nd_person>*
YOUR	*<possessive, pronoun, 2nd_person>*
THEE	*<nominative, pronoun, 2nd_person, archaic>*
THINE	*<possessive, pronoun, 2nd_person, archaic>*
HE	*<definite, article, male>*
SHE	*<definite, article, female>*
MORE	*<post, determiner, relative-quantity, relation>*
LESS	*<post, determiner, relative-quantity, relation, negative>*

Table 3.1: Examples of codes for closed class items.

work. For example, WordNet codes the most frequent sense of [chess] as a gramineous herb, and that [blimp] as a Civil War colonel).

Each synset contains a set of pointers encoding different relations. The IS-A relation is a single pointer which indicates a single synset which is superordinate to the synset. The [dog] synonyms set points to the [canine] synonym set, indicating that a [dog] IS-A [canine]. The synonym set for [wolf] also has an IS-A pointer to [canine], making [dog] and [wolf] siblings in the IS-A tree. The IS-A relations for all words proceed up the semantic tree until a top level node is reached (e.g., [entity], [abstraction], [action]).

The organization for verbs in WordNet is identical, except that verbs contain additional information regarding the syntactic frames they can appear in. Such information is not used in this project but could be used in other work.

Verbs in WordNet are not as detailed as the nouns. While the representation for nouns can be quite rich, most verbs are quite simple. For example, the verbs [heat] and [warm] are both children of the [change] verb set, but so is [improve]. The notion that [heat] and [warm] have to do with changing a physical property, particularly temperature, is absent. Nonetheless, the representations are still rich enough to be usable.[4]

[4] Adjectives and adverbs in WordNet are totally impoverished, and so these items were hand coded instead.

3.4.3 Generating Features from WordNet

The IS-A relations for a word are used to generate semantic features as follows. For each lexical item, the most common sense (the first one listed in WordNet) is used. A text representation of the synset corresponding to that word is printed. Then, the IS-A relations are followed up the tree. As each synset node is visited, the text representation of that synset is printed. This repeats until the top of the tree is reached.

3.4.4 Compaction

There is tremendous redundancy in the representations that WordNet produces. Many features serve no role in distinguishing words. This is particularly true of smaller word sets. For example, if you have a set of words whereby every creature that is a <*chordate*> is also a <*vertebrate*>, then the feature <*chordate*> is redundant and irrelevant. If the training set contained items that were <*+chordate*> but <*-vertebrate*> (like a lobster), then the feature <*chordate*> would serve a useful role.

To economize the number of features in the final representations, features that serve no distinguishing role are detected algorithmically and deleted. This pruning operation is depicted in Figure 3.1, where the node <*chordate*> is deleted from an example semantic tree.

An additional form of pruning was then performed on the network. The leaves of the tree were removed. Formally, all nodes which belong to one and only one lexical item were deleted. For a representation of n words, this tends to remove approximately $0.8 \times n$ nodes. This removes item-specific, or localist nodes. It also introduces a large amount of ambiguity into the representation. Before this pruning operation, the representation of [pup] consisted of the representation of [dog], plus an additional <*pup*> feature. Since the <*pup*> feature is only used for the concept [pup], it would be deleted. This would leave [dog] and [pup] ambiguous; they would have the same representation.

For example, in WordNet [mud] is a kind of [dirt]. Similarly, [marsh] is a kind of [land]. The single leaf pruning procedure would leave [mud] and [dirt] as synonyms. The sets of synonyms were detected algorithmically and printed out. Features were added which disambiguated synonyms. In this example, a feature <*wet*> was introduced, which disambiguates [mud] from [dirt], and also distinguishes [marsh] from [land] (and [wade] from [walk], and [scald] from [burn]). Register features such as <*formal*>, <*archaic*> and <*slang*> were also used. The words YOU and THEE were distinguished by marking THEE as an <*archaic*> form; similarly YOUR and THOU. The concepts MOM and MA were distinguished by marking MA as <*informal*>.

The basic reason for the condensing operations is that item specific nodes can be deleted, which cuts the number of features needed almost in half, and then ambiguous items can be disambiguated either by hand or algorithmically, by re-using a set of features. These pruning operations were not theoretically driven; it is quite reasonable that the human semantic system actually has tremendous redundancy in it. These operations were done solely to keep the size of the representation manageable. Prior to pruning, each non-synonymous word in WordNet contains a node for itself as well as all its hypernyms, hence n words required approximately $2.0n$ nodes. For the present simulation, this was too computationally expensive. The pruning operation reduces the number of

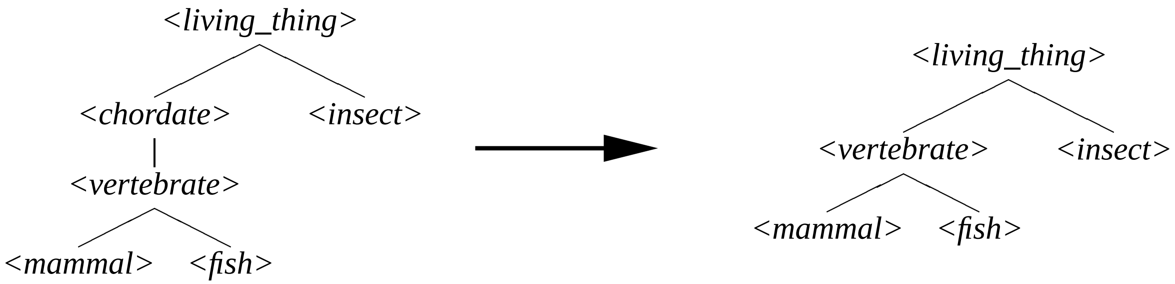


Figure 3.1: The process of removing redundant semantic features

semantic nodes for n words to approximately n, while preserving the positive properties of the representation (distributed features, with the correct variation in the degree of intercorrelations among features).

WordNet's IS-A hierarchy does not distinguish directionality in its features. For example, the nouns [wealth] and [poverty] share the feature of being financial conditions, but the notion that they are in fact *opposites* is not encoded. The same goes for nouns like [heat] and [cold], [truth] and [falsehood]. A <*negative*> feature was introduced to amend this; hence [cold] is the negation of [heat], [poverty] is the negation of [wealth], and so on.

WordNet does include antonym pointers, but these are rather sparse. For example, [heat] does not have any antonyms listed, nor does [quiet]. If the WordNet listing of antonyms was more complete, the process of assigning <*negative*> features could be accomplished algorithmically. They weren't, however, so the process had to be done by hand.

The IS-A hierarchy also does not distinguish degrees of magnitude. Intuitively, the relationship between adjectives such as [warm] and [hot] relates to the relationship between [cool] and [cold]. The features <*low_intensity*> and <*high_intensity*> were introduced. The <*high_intensity*> feature distinguishes items such as [shout] from [shriek], [warm] from [hot], and [jibe] from [barb]. Coupled with the <*negative*> feature, the intensity features form a simple two dimensional matrix along with items such as [warm], [hot], [cool] and [cold] are intuitively categorized.

A majority of the synonym sets (about 1600) were disambiguated this way. The remaining 500 consisted of items whose meaning was not transparently different (e.g., GROAN and MOAN). These items were disambiguated algorithmically. A set of 30 features (labeled <*R1*> through <*R30*>) were used. For each set of n synonymous words, n features were chosen at random from this pool. These features can be thought of as corresponding to overlapping distinguishing features whose meaning is opaque. Through this method, the 500 ambiguous sets can be disambiguated with just 30 features, without sacrificing underlying similarity structure. Again, it is not a theoretical claim of this work that a small set of common, semantically empty features disambiguate words. Rather, a small set of features is re-used to keep the size of the featural representation small. If additional unique features were used to disambiguate all 500 remaining ambiguous synonym sets, an additional 500 units would be necessary, which is too costly.

3.4.5 Morphological Relations

A list of plurals, past tenses, and 3rd person singular word forms was assembled from other studies (Hoeffner, 1996), from the CELEX electronic corpus (Baayen et al., 1993) and from the Penn Treebank Project (Marcus, Santorini, & Marcinkiewicz, 1993). These lists were in the form of an association list, mapping an uninflected word form onto the appropriate morphological form (e.g. the plural list contained an entry CAT → CATS). For the plurals, items in the set which were coded as nouns by the Brown corpus (and hence had a noun representation from wordnet) were matched against the plural list. The resultant plurals were assigned the semantics of the uninflected item, plus a <*plural*> feature. Hence [cat] and [cats] differ only by <*plural*>, [dog] and [dogs] likewise. The past tense items and third person singulars were created similarly; the list was combed for items coded as verbs (and hence having verb semantics from WordNet). Verbs found in the past tense list were given a <*past_tense*> feature; verbs found in the third person singular list were given a <*3rd_person_singular*> feature. These features were appended to the core features for the uninflected verb. This operation created 1,800 plurals, 730 past tenses and 730 third person singulars. There were far more plurals than past tenses because nouns outnumbered verbs in the corpus. Although [dog] can be used as a noun or a verb, it is most commonly a noun; hence the representation contains [dogs] but not [dogged].

3.4.6 Complexity Analysis

For the purposes of this analysis I will assume that the IS-A hierarchy in WordNet is a perfect k-ary tree, that is, a symmetric tree where each non-leaf node has k children. The hierarchy in fact isn't, but this assumption makes the analysis tractable and does not deviate far from the actual form of the tree.

The first step, involving matching the n items in the original word list against the Brown corpus is $O(n)$ (assuming both lists are sorted) and accomplishable by machine, so the operation is essentially cost-free.

The cost of computing the features for n items involves n traversals up the tree. A k-ary tree with n leaves has height $O(\log_k n)$, so the total computational cost is $O(n \log_k n)$. This operation is fully automated. Each item has $O(\log_k n)$ features, so $O(n \log_k n)$ space is required to store the features. The removal of leaves is $O(n)$, and the detection and removal of redundant features is $O(n \log_k n)$, again accomplished by machine.

Of greater concern, from a practical standpoint, is the human cost. The disambiguation of items is currently done by hand. When the leaves of the tree are removed, this removes $O(n)$ features, but introduces $O(n/k)$ ambiguous sets of $O(k)$ concepts each. These $O(n/k)$ sets must be disambiguated by hand. This is the most labor intensive part of the operation.

It should also be noted that it is not the case that one needs to create representations by hand for the $O(n/k)$ sets; the similarity structure is already present. One need only introduce sufficient features to disambiguate these items from each other.

It should also be noted that the entire operation could be foregone, by algorithmically using a random k features to disambiguate each set. The quality of the representations would suffer (see above discussion of negation and magnitude, for example), however it would make the entire task of generating noun and verb semantics fully automated. In the current study, both methods were used; hand coding to disambiguate a large proportion of the sets, and algorithmically generated features for the remainder.

The closed class items, and adjectives must be coded by hand. This task is time consuming and difficult, but the items are few in number. Additionally, as the size of the training corpus increases, this task does not increase with it. A large proportion of the CVC words are closed class items; a far smaller proportion of CCCVCCC words are, and a dramatically smaller proportion of the polysyllabic words are. In short, as one uses a larger and larger corpus, the size of this task scales very slowly.

The addition of morphological relationships is $O(n)$, provided a list of morphological translations is available (that is, one that keys [cats] as the plural of [cat]). Such a list for larger sets of words could be generated algorithmically from the CELEX corpus (Baayen et al., 1993), which codes words by lemma in addition to word form, and includes a morphological relation database.

3.4.7 Imageability and the Representation

A property of pure binary trees is that at each level, there are more nodes than in all the levels above it (one more, to be exact). Hence the leaves of the tree account for half of the mass of the tree. While WordNet is not a pure binary tree, this property is approximately true of its IS-A hierarchy. By deleting the leaves, and disambiguating the remaining ambiguous items with a finite and reusable set of features, the number of nodes in the tree is reduced by half.

However, the introduction of reused nodes (e.g., <*bad*>, <*wet*>, <*archaic*>) means that the semantic representation is no longer a proper tree. Other dimensions such as register, magnitude and goodness cut across the IS-A hierarchy.

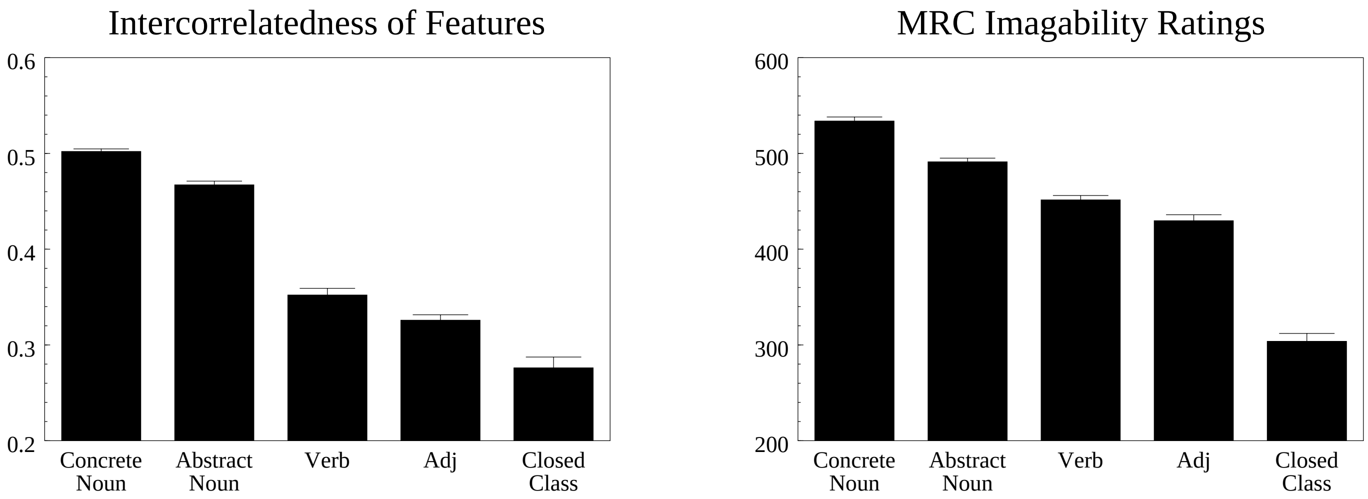
The representation deviates from a pure binary tree in other important ways, however. A strict tree, with each concept encoded as a leaf and containing features following the height of the tree would have the same number of features for each concept. Additionally, the degree of intercorrelatedness of the features would be the same for each concept (ignoring variation due to the frequency of the concepts).

As shown earlier, the intercorrelatedness of semantic features is important for accounting for a wide range of behavioral phenomena. To verify that the derived semantic representations have the desired properties, the intercorrelatedness of features was measured for various word classes. For each pair of semantic features (1600^2 pairs), the Pearson correlation over all words was computed. For each word having k features, the degree of intercorrelatedness of the semantic features for that word was computed as the mean Pearson r for the $k(k-1)$ pairs of features (auto correlations are of course 1.0 and so are not included).

This degree of intercorrelatedness is plotted for different word classes in Figure 3.2(a). The syntactic word classes are derived from the Brown corpus; the distinction between abstract and concrete nouns is based on the presence of the <*object*> semantic feature.

The MRC psycholinguistic database (Coltheart, 1981) contains ratings of the imageability of words. Figure 3.2(b) shows the imageability ratings for various word classes. The word classes are defined by the derived semantic representation described in this chapter; syntactic category from the Brown corpus and the division of concrete and abstract nouns from the <*object*> feature from WordNet. A qualitative match is obtained between the imageability ratings and the intercorrelatedness measures (Figure 3.2(a)). Additionally, both ratings match the empirical results obtained from patient data: concrete words $>$ abstract nouns $>$ verbs $>$ adjectives $>$ closed class words. The degree of intercorrelatedness of individual items correlated with the MRC imageability ratings weakly but reliably: $r = 0.27$, $p < 0.001$. A better match is obtained when the frequency of a word and its intercorrelatedness are used jointly to predict the MRC imageability norm: $r = 0.38$, $p < 0.001$. This may be due to the MRC imageability norms being partially confounded with the frequency of a word: subjects' judgments of the imageability of low frequency words (e.g., BARGE) could be more error prone and noisy than that for high frequency words (e.g., BED).

This section has shown that a measure of the semantic features' intercorrelatedness has many desirable properties. This measure is a static property of the representation, relating to which features are assigned to which concept. It does not show how learnable such intercorrelatedness is, nor what effect this property has on the behavior of an attractor network trained on this representation. Such dynamic properties of an attractor network trained on this representation will be presented in a later section detailing the training regime and performance of the semantic attractor network.



(a) Intercorrelation of Semantic Features

(b) Imageability from MRC Database

Figure 3.2: Semantic Correlations and Imageability.

3.4.8 Objections to Feature Based Semantics

Philosophers have raised a number of objections to traditional feature-based semantic representations. One complaint is that semantic representations are actually dynamic and context dependent. It is very difficult to identify the necessary and sufficient features for a concept. Such phenomena seem to demand a probabilistic and context dependent notion of features.

It should be emphasized, however, that the current approach does not preclude such optional or probabilistic features. Phonological attractor networks have been shown to be able to complete partial, degraded or noisy patterns into a veridical configuration (Harm & Seidenberg, 1998). Semantic attractors can similarly overcome or repair small amounts of damage; the degree of repair being dependent on the intercorrelations of the features (Devlin et al., 1998). A concept can thus be "recognized" even if it does not have a full set of typically-present features, much in the same way that the auditory presentation of a degraded /p/ can still be recognized as a /p/ (Liberman, Harris, Hoffman, & Griffith, 1957), or in the same way that a word with noise-masked phonemes can be recognized (Warren, 1970). On this view, concepts are points in high dimensional space, and are moved around by network dynamics. The "identification" of a concept, then, is measured not as a match of a minimal set of defining features, but rather a graded measure of proximity in multi dimensional semantic space (e.g. Churchland, 1989). The use of probabilistic, variable and context dependent semantic features is beyond the scope of the current research but is easily admissible within the general framework.

Further criticisms of classical notions of semantic features in an IS-A hierarchy derive from reaction time studies. Collins and Quillian (1969) presented a model of semantic memory based on an IS-A hierarchy. Crucially, inference was conceived as traversing the hierarchy from the word node (the leaf of the tree) on up. This model is problematic, however, because the notion of traversing the tree implies that concepts higher up the tree ought to be slower to verify. However, Rips, Shoben, and Smith (1973) found that concepts lower in the tree are often verified slower (i.e., that a dog is a mammal is slower to verify than a dog being an animal). This has led to amendments of the basic network hierarchy idea, with the additional proposal that higher frequency concepts are located more proximally to word nodes (Collins & Loftus, 1975). In short, the conclusion is that semantic memory cannot be organized as a strict IS-A hierarchy.

It is crucial to understand that the semantic model being presented here is an IS-A hierarchy *only* in the method that generates the features. Fundamentally, the representation encoded here is a set of units with activities that change over time. Processing in the actual network does not involve traversal of the graph; that is instead the offline method for deriving the features. The units are instead activated in a standard parallel, interactive connectionist framework. Hence, earlier criticisms of such semantic hierarchies as Collins and Quillian (1969) simply do not apply here.

Fodor (1998, 1994; Fodor & Lepore 1992) presents arguments against *holistic* semantic theories; that is, theories in which the content of a concept is dependent or defined based on the content of other concepts. Such an approach is explicit in the semantic representation presented here, where features are selected in part simply to distinguish one concept from another. Fodor's argument goes like this: we want to discover psychological laws; these laws ought to be similar to the physical laws that a physicist seeks to discover. The physicist does not want one law for, say, the behavior of gold in New Jersey and a different one for gold in Florida. Gold is gold, if we want the laws governing it to be invariant. Similarly, then, he argues that if we wish to find psychological laws governing [cat] concepts, then patently different people must have the same base [cat] concepts, and a rigorous notion of content equality must be in force. Fodor and Lepore (1992) darkly intone:

> The consequences of option 1 [buying into holism] would seem to be horrendous: either behaviorism or materialistic eliminativisim in the philosophy of mind and massive unemployment in cognitive science . . . (p. 187)

The requirement for content *identity*, rather than *similarity*, therefore, leads Fodor and Lepore to reject holistic approaches to meaning: if my concept of [cat] is dependent on all my other thoughts and experiences with cats, then patently my concept of [cat] cannot be identical to yours or anyone else's.

It seems, however, that Fodor's complaint may arise simply from his notion of the goal of the cognitive enterprise. It is a *conjecture*, not an *axiom*, that the goal of cognitive science is to, for example, discover psychological laws governing [cat]-thoughts. It is not really the goal of a physicist to enumerate a set of laws for gold, but rather to identify the forces and processes and constituents of matter that give rise to phenomena in the world; gold and lead differ in certain crucial ways (one dissolves in mercury, one does not) and are similar in other crucial ways (both are heavy relative to, say, tin, or soft relative to steel). One could conceive of my [cat] thoughts and your [cat] thoughts as being more like lead and gold; the goal of the cognitive scientist, then, is to discover the properties and mechanisms that make our concepts behave similarly *and differently*. The weighted, distributed, dynamic representations under consideration here work well within this framework: [dog] and [brick] share the common node <*entity*> and as such could be the subject of a sentence, but only [dog] contains the feature <*living_thing*> and hence only a [dog] can think, live or breathe.

In a different vein, Fodor raises other objections to feature based approaches to meaning, particularly within the realm of connectionist models. Fodor and Pylyshyn (1988) argue that a distinction must be made between the *role* and *filler* of an expression; the sentence JOHN KISSED MARY contains JOHN as the subject and MARY as the object. Any cognitive system that can entertain JOHN as a subject and MARY as an object must also, *by definition*, be able to entertain MARY as a subject and JOHN as an object, yet if representations were simply laid out as the concatenation of features then this would not be the case. Similarly, Fodor and McLaughlin (1990) argue that the content of the concept [cup of coffee] must contain the concept [coffee] and [cup], but not simply be the concatenation of the features for the two concepts. The role relations must be preserved, in a way that does not change that of the constituent parts. That is, coffee is coffee, whether its in a cup or not. They argue that adding features to clarify the relation of coffee and cup (e.g., that the [coffee] is in the [cup] and not the other way around) would change the core representation of [coffee]. Its very similar to the gold example; Fodor claims we can't have coffee in a cup being different from coffee in, say, a mug, or a different cup, or a pot. Fodor claims therefore that the standard connectionist method of representing concepts (either as concatenations of features or tensor products, as in Smolensky, 1990) is not reasonable and cannot account for important cognitive phenomena.

Connectionist networks do have a means of separating roles from fillers, however. The temporal binding of variables (Shastri & Ajjanagadde, 1993; Hummel & Biederman, 1992; Henderson, 1994) offers one such mechanism. The extent to which the brain actually utilizes temporal binding is controversial, but there is some empirical support for it (Shastri & Ajjanagadde, 1993). Fodor and Pylyshyn (1988) argue that systems with such properties actually are "mere implementations" of the classical approach to cognitive science. Whether they are or are not is not, however, the important question; what is important is whether such mechanisms actually are used by the brain or not. In any event, such temporal binding systems work within the general connectionist framework and seem to be capable of providing for content-preserving role and filler relations.

A further objection to feature based semantic theories is that they are very poor at relating relations between concepts, or at forming rapid inferences about concepts. Certainly, a feature <*can-crush-walnuts-by-stepping-on-them*> would not plausibly be part of anyone's core concept of [hippopotamus], but people can perform such inferences quite rapidly. Murphy and Medin (1985) present an experiment in which subjects are asked to divide sets of symptoms from two unknown diseases. The subjects' decisions tend to follow higher order, causal relationships between items rather than lower level defining features. Further, when subjects were asked to define novel compound words (e.g., OCEAN DRIVE), they tended to provide relational definitions rather than a simple concatenation of features for the two concepts. Murphy and Medin (1985) and Keil (1989) conclude that people's concepts are in fact best describable as *theories* of the relationship between concepts and general world knowledge.

Note that most of these examples of inference resemble Fodor's example of the cup of coffee; both involve relations between concepts that preserve a role and filler status rather than a simple union of features. However, Shastri and Ajjanagadde (1993) presents examples of how a temporal binding network could perform novel inferences based on a form of world knowledge. Such networks provide a neurally plausible version of production rule systems commonly found in AI research. For example, the concept [hippo] has core features such as <*heavy*> and <*large*>. A temporal binding network such as that proposed by Shastri and Ajjanagadde would link the <*heavy*> feature with <*can-crush-fragile-things*>; the link between these features would have been derived from non-hippo-specific world knowledge. Upon learning that a novel item has the feature <*heavy*>, the system could form the correct inference <*can-crush-fragile-things*> without being explicitly trained on that semantic relationship.

Therefore, while a bland list of static features would fall prey to Murphy and Medin's criticism, it is not at all clear that a set of probabilistic features, embedded within a larger connectionist system (containing attractor networks, and inference engines as envisioned by Shastri and Ajjanagadde) could not account for the relevant phenomena. In fact, it is not clear that such a system would not in fact resemble, at some level of description, a neurally plausible implementation of their theory. Under this hypothesis, the featural knowledge used in immediate, automatic word recognition is not different *in kind* from that used in higher order inference. Such higher order inference machinery is permissible within this approach, however its absence from this work is justifiable because such inferences seem largely unimportant to the task of automatic, online word recognition, which is the focus of the current study (see also McRae et al., 1997; Devlin, 1998).

Finally, a last criticism of standard feature theories is that they fail to capture the notion that a concept has an "essence" to its meaning. Putnam (1989) provides examples of "twin cases" to illustrate the point. Suppose on some other planet (a "twin earth") there was a clear colorless liquid that behaved exactly as water, except instead of being water it was material XYZ. If it behaved exactly as water but we knew it was not H_2O but rather XYZ we would not have the same concept of it as we do of water; the *reference* to "water" that we use would refer to *our* water, not the XYZ on our twin earth. This is because there is more to the concept of water than a simple set of perceptual and functional features. Keil (1989) provides an similar example. School children were shown a picture of a raccoon, and told that the raccoon's fur was painted black with a white stripe down its back, and odor glands were added to its body. Younger school children generally said it was a skunk, but the older school children almost unanimously said that the creature was still a raccoon, despite its perceptual and functional features being far more skunklike than racoonlike. The idea is that there is an "essence" to raccoon-ness beyond the set of perceptual and functional features that define the creature, and that as children develop concepts, they rely less on characteristic features and more on essential properties.

This is clearly a criticism of feature theories where features are limited to being perceptual and functional features. But the idea that the "essence" of a concept goes beyond the concepts mere perceptual and functional features can be accommodated within this framework. The WordNet features utilized here do actually have features for concepts such as [canine]; such features can be thought of capturing the essential nature (DNA, or whatever) of a canine; similar for other higher level concepts such as [metal] or [gas].

Considerations of holism, essentiality, compositional structure and role-filler relations are therefore not compelling reasons to reject a connectionist approach to featural semantics. Certainly, for the purposes here, such representations are sufficient. The primary focus of this research is automatic, online word recognition, not higher order judgments of category membership or concept similarity. We of course do not want to commit to a theory of meaning that rules out the possibility of future extensions into higher order cognitive tasks; there seems no reason to believe that the current approach has done so.

While there seems to be no *principled* limitations to the general approach undertaken here, there are many limitations to the current instantiation of semantic meaning described here. These are based on limitations of current technology, methodology, design decisions based on the need for simplicity. They will be described next.

3.4.9 Limitations of the approach

There are many limitations of this approach. The most serious is the reliance on WordNet's IS-A hierarchy for the bulk of the semantic content. There are many things about concepts that are not encoded in a strict IS-A hierarchy; properties that cut across categories. For example, both birds and airplanes have wings and fly. The senses of [wing] and [fly] are slightly different, but should share some overlap. In WordNet's IS-A hierarchy there is no such overlap. I should stress that this is not a limitation of the representation scheme, but of the method for deriving the features. A concept of <*wing*> that is shared by both [bird] and [airplane] could be easily encoded; it is not present simply due to WordNet's limitations.

Similarly, associatedness is not encoded; the representation of [doctor] and [hospital] do not overlap any more than [doctor] and [fireplug]. The concept [doctor] is in the <*living_thing*> hierarchy and [hospital] is in the inanimate object hierarchy; in WordNet such things share no overlap except for the top level <*entity*> feature. The ability to infer such relations would involve presenting meaningful strings or concepts to the network, and the addition of a means by which relations between items could be absorbed. Allen and Seidenberg (In Press) present a model which forms such inferences between the semantic structure of items in a sequence via exposure to strings of natural text. In the current implementation, however, no such inferential mechanism or sequential training regime is used.

An additional problem is that WordNet does not have any overlapping structure between the noun and verb databases; there is no overlap between the noun sense of [hit] and the verb sense of [hit].

There is evidence from neuroimaging of a physical dissociation in the brain between functional and perceptual features (Perani et al., 1995). WordNet makes no distinction between functional and perceptual features; in general, functional features do not even exist. So while the fact that a hammer is a tool, an object, and an artifact are all encoded, the fact that it is used to pound nails is not represented at all. Such a dissociation between perceptual and functional features is of theoretical importance for accounting for dissociations in naming performance of Alzheimer's patients (Gonnerman et al., 1997).

The method of creating plurals and past tenses, while working well for the vast majority of items, is not appropriate for some items. While it is reasonable that the semantics of most plurals (e.g., [cats]) have identical semantics to the singular form except for the notion of pluralness, there are exceptions. Consider the sentence *Ahab sailed many waters*. Here the meaning of [waters] is not simply that there were several instances of [water], and Ahab sailed them; [waters] is not related to [water] in the transparent way that [rocks] is related to [rock]. Many forms are even more opaque. To "put on airs" is not to put on one air, and another air, and another. And while the word Alps in the phrase "Swiss Alps" is in fact a plural, it would be quite odd to hear a speaker of English say "look at that Alp over there." The formation of plurals and past tenses is a quasi-regular domain, with exceptions not only in phonology but in meaning. The quasi-regularity in the phonological domain is captured here but the semantic irregulars are not.

In a similar vein, words are limited to only one sense. The word DOG is trained to the canine sense, not to the ugly person sense, the scoundrel sense, or the sense of the thing in the fireplace that holds the logs. Solving these problems in a thorough way, however, would involve taking seriously the representation of context. This in of itself would be a huge task, one well beyond the scope of the current project.

The total reliance on IS-A relations is a limitation as well. This could, in fact, be extended to other relations. WordNet codes HAS-PART relations as well as IS-A relations. Each node in the tree contains pointers to HAS-PART objects; this tree can be traversed by following the IS-A tree. Figure 3.3 shows a subset of the WordNet relations for some simple words that illustrate. Here, a [dog] is a [canine], which HAS-PART <*paws*>, just as [feline] does. Canines and felines are [mammals], which have HAS-PART <*hair*>. A [mammal] is an [animal], which HAS-PART <*head*>, <*limb*>, and so on. The HAS-PART objects also define a connected graph: the concept [head] has HAS-PART relations <*eyes*>, <*nose*>, <*ear*>. By making a directed walk of the entire tree from a given node, a full set of IS-A and HAS-PART constituents can be generated for that node.

The current implementation does not utilize these relations. Generating the HAS-PART relations would involve a tremendous increase in the number of features. The node for [car] alone has over 30 HAS-PART entries (e.g., <*throttle*>, <*mirror*>, <*glove-box*>, etc.) As more rapid hardware and more efficient simulation technology becomes available, such relations can be easily added. Such additions would increase the richness of the representation in interesting ways: [bird] and [airplane] would overlap in having a part <*wings*>.

Having acknowledged these limitations, it should be re-emphasized that while these problems restrict the range of phenomena that can be explored with this representation, they do not seem to be crucial to the central research question of this work. Further, the general approach outlined here

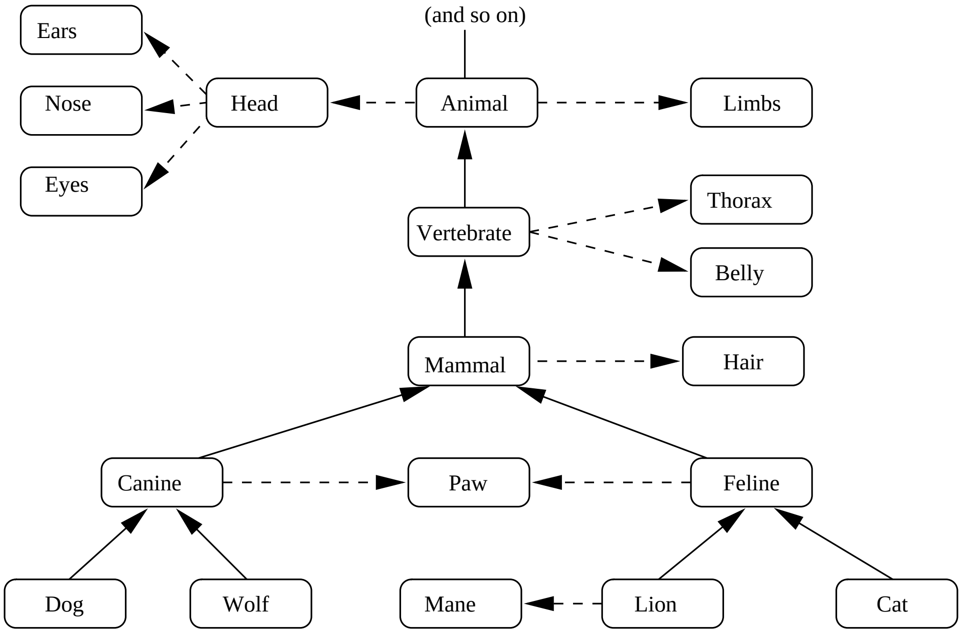


Figure 3.3: A segment of the WordNet tree. IS-A relations are shown with solid lines; HAS-PART relations with dashes. The set of nodes for [cat] could be generated by directionally traversing the tree following the IS-A links and HAS-PART links.

is extensible to accommodate these concerns. Such extensions will be the focus of future work, as the general model is adapted to model higher order phenomena such as lexical and syntactic ambiguity resolution. Furthermore, with respect to the criteria outlined in the beginning of this chapter, this method of deriving semantic representations represents a substantial advance over all previous connectionist models.

Chapter 4

Model Details

4.1 Continuous Time Networks

Since the discovery of the backprop algorithm (Rumelhart et al., 1986), the simplest form of connectionist network has been the feedforward network (Figure 4.1). The network consists of a set of *input* units, a set of *output* units, and a set of *hidden* units mediating between them. On each trial, the j input units u_j are clamped to some desired value. The hidden units compute their values based on the input unit activity and the weights w that map the input units to the hidden units. Each hidden unit h_i for each of the i hidden units computes its output value as $h_i = f(\sum_j w_{ij} u_j)$, where f is a nonlinear squashing function. Similarly, each of the k output units o_k computes its output based on the hidden unit outputs: $o_k = f(\sum_i w_{ki} h_i)$.

Such networks adhere to a neural metaphor to the extent that the processing of each unit is driven by the local propagation of activity along weighted connections, rather than, for example, a central processing executive. However, the metaphor stops there. Such networks are explicitly *stateless*, that is, there are no state transitions in the network; just the final computed state in which activity has propagated through the entire system. There is no time course of activation, no processing dynamics, and no sense in which the current state of the network modifies its subsequent states.

Recurrent networks utilizing backprop through time (hereafter BPTT; Williams & Peng, 1990) address some of these limitations. In such networks, a notion of time is added, such that the output of a unit at time t depends not on the activity of units in a previous layer, as in feedforward networks, but on that of all units at a previous time slice. This kind of network is a generalization of the feedforward network, and allow for recurrent, or cyclic connectivity patterns. The activity of a unit u_i at time t, u_i^t is defined as $u_i^t = f(\sum_j w_{ij} u_j^{t-1})$. A units' activity at time t then is totally determined by the activity of all units connected to it at time $t - 1$. These networks can form dynamical systems, exhibiting either stable fixed points or oscillating behaviors. Further, activity within a group of units can build up over time, with the units influencing each other's states.

However, the temporal dynamics of such networks is still quite simplistic. They operate in a lock step fashion, where the output of unit is the squashed sum of its input regardless of anything else. The output of units, then, tends to "jump;" activity does not ramp up or down gradually but instead can respond instantaneously. Hence, while the network will exhibit global temporal dynamics, each individual unit still has a very simple time course of activation.

Pearlmutter (1989, 1995) formalized a way to train backprop networks with much more subtle time courses of activity. Continuous time networks such as those introduced by Pearlmutter add

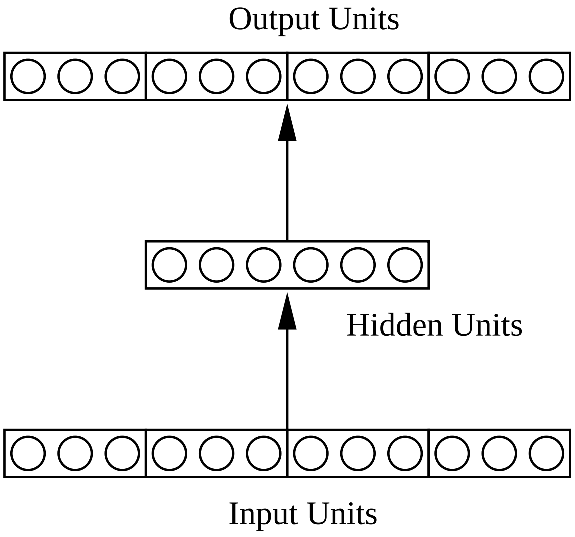


Figure 4.1: A canonical feedforward network architecture.

unit dynamics: a unit's output ramps up gradually as a function of its input, based on a leaky integrator equation:

$$\sigma\frac{\partial o_i}{\partial t} = (y_i - o_i) + b_i \qquad (4.1)$$

$$y_i = f(\sum_j w_{ij} o_j) \qquad (4.2)$$

where y_i is the squashed input to the unit (or, what its output would be in a discrete time network), o_i is the instantaneous output of the unit, and b_i is a resting state of the unit. The parameter σ controls the speed at which a unit ramps up or down. Essentially, the rate of increase of a unit's activity is proportional to the difference between its current activity o_i and what its activity *ought* to be (y_i).

Pearlmutter generalized the backprop equations to allow error gradients to be integrated up over time, the way that activity is integrated up over time. This allows us to train such networks with the full power of the backprop algorithm.

Plaut et al. (1996) introduced a subtle but important change to the Pearlmutter equations. The Pearlmutter (1989) formulation had the *output* of a unit ramping up over time in response to the instantaneous squashed *input* to that unit. Plaut made the *output* of a unit the instantaneous squashed value of the *input* to a unit, and caused the *input* to units to ramp up over time. Formally,

$$o_i = f(y_i) + b_i \qquad (4.3)$$

$$\sigma\frac{\partial y_i}{\partial t} = (x_i - y_i) \qquad (4.4)$$

$$x_i = f(\sum_j w_{ij} o_j) \qquad (4.5)$$

While mathematically similar, there are important theoretical differences between these two processing dynamics. In the Pearlmutter formulation (which has been termed time-averaged-outputs, or TAO), the maximum output of a unit (typically 1) determines the maximum rate of climb of the unit. As such, if one unit receives an input of 10, its asymptotic output is 0.99999, and so it climbs to that value; if a second unit receives an input of 100, its asymptotic output is 0.999999, and it climbs to that value at almost exactly the same rate as the first unit. The error gradient equations reflect this: if a unit is ramping up as rapidly as it can, additional input does not help, and the error gradient is zero. In contrast, with the time averaged input (TAI) networks, if one unit gets an input of 10 and another gets 100, the second unit ramps up much more rapidly than the first. Equation 4.1 cannot evaluate to more than 1.0 (assuming b_i is zero, as is typical), while Equation 4.4 is unbounded, because the summed input to a unit, x_i, is unbounded.

Early pilot simulations utilizing TAO networks failed, because they simply implemented the wrong theory. A crucial design principle of this project is that summed activation causes more rapid rise times of units (Section 2.1.3). It was found early on that if orth→phon→sem was driving semantic units as rapidly as they could be driven (i.e., with a derivative of 1.0) then there was no advantage to additional input from orth→sem; such input would not drive the semantic units any faster. Like Bullinaria's (1996) model, this implemented a *wrong theory*, one in which a network that would *ultimately* compute the correct output received no pressure to compute that output more rapidly, and hence utilize as many other sources of information as possible. For these reasons the TAI networks are used throughout this work.

All of the networks discussed here are error correcting networks using variants of the backprop learning algorithm. The use of error correcting methods is theoretically important (see Section 2.1.7), however backprop has been criticized for lacking biological plausibility. Specifically, the backprop algorithm requires that error signals propagate backwards along connections; activity moves forward and error signals move backward. There is nothing currently known in neurobiology that could support such bidirectional movement of information in neurons.

However, it is important to consider that the backprop algorithm is not meant to be a literal simulation of the exact cellular mechanisms of learning, but rather a means by which the weights are adapted to follow an error gradient in weight space. Other, more biologically plausible mechanisms such as a deterministic Boltzmann machine have been formally shown by Hinton (1989) to follow an error gradient in weight space. Hence, there are algorithms whose behavior is qualitatively identical to backprop. These methods, however, are not used here for reasons of computational economy. Plaut (1991) reported several simulation using backprop, and a replication using a more neurally inspired Hopfield net. He found that the networks exhibited very similar behavior; the primary difference being that the Hopfield net took approximately 40 times longer to train. For the simulations reported here, this time scaling is unacceptable; the time to train the network would run about 2.5 years. Backprop is used because, while it has problems with its biological plausibility, it is far more computationally efficient than more neurally inspired methods.

Further, O'Reilly (1996) has introduced a learning algorithm which is argued to be a biologically plausible form of backprop learning. In this algorithm (termed the *generalized recirculation algorithm*), rather than propagating error backwards, a set of backward connections propagate both output and target signals, and the difference is computed at each neuron. O'Reilly argues that this mechanism is reducible to deterministic Boltzmann machine learning, hence showing that the two learning mechanisms, with the appropriate architectural modifications, are not so different after all.

While the continuous time networks are considerably more sophisticated and interesting than the feedforward networks of yesteryear, they are still appallingly simplistic compared to what is known about actual neurons and the techniques of modeling their activity (cf. Koch & Segev, 1989). However, the research is following a normal progression where the range of phenomena to be modeled is expanding, and with it, the fidelity to actual biological systems is also increasing. Plaut and Shallice (1993) and Harm and Seidenberg (1998) utilized attractor dynamics in BPTT networks to explain patterns of impairment in deep and phonological dyslexia. Such studies revolve around the idea of attractors in state space, and hence would not have been possible with simple feedforward networks. In a similar vein, the current study demands continuous time networks to fully implement the principles outlined in Chapter 2. Further advances in understanding behavioral phenomena, in understanding the neurobiology of learning and processing, and in the power of computer simulation systems will both enable and demand greater biological realism.

4.2 Training Corpus

A subset of the monosyllables was used in these simulations, to reduce simulation time. The training corpus was limited to the monosyllables which have a CVC phonological structure: 1,825 words in total. There were 272 sets of homophones, containing a total of 586 words (some homophone sets contained 3 or more words, such as DO, DEW and DUE). There were 21 homograph pairs, generally consisting of items whose inflected form matches the uninflected form (e.g., SHEEP, FISH).

The semantic representations were derived according to the algorithm described in Chapter 3. The algorithm generated a total of 1,343 semantic features.

The phonological representations were similar to those presented in Harm and Seidenberg (1998) (hereafter HS98). Three phoneme slots were used to encode the CVC words, with vowel centering to minimize the “dispersion” problem (see Plaut et al., 1996). A set of 25 phonological features were used to describe each phoneme; these were derived from feature matrices in Chomsky and Halle’s (1968) work “The Sound Pattern of English” (hereafter SPE). The same features were used to encode vowels and consonants, so each slot contained the same features which are the union of those SPE features relevant for consonants and vowels. All features were binary, taking on values of 0 or 1. The HS98 model used continuous valued features; while having the desirable property of naturally encoding processes that are fundamentally continuous (such as tongue placement), it has the property of making it very difficult for the network to train to intermediate values. Intermediate values are located on the “slippery” or steep part of the activation curve; small perturbations in the input to a unit have large consequences on the output side. Training to binary targets results in more rapid convergence. SPE treats all features in the matrices as binary.

The representation of consonant phonemes is given in Table 4.2; vowels are in Table 4.2.

The frequency of each item was coded using a logarithmic compression of the Wall Street Journal corpus (Marcus et al., 1993) according to the formula

$$p_i \quad = \quad \frac{\log{(f_i + 1)/5}}{\log m/5} \tag{4.6}$$

where f_i is the WSJ frequency of the ith item and m is 30,000 (a reasonable cutoff frequency). Values over 1.0 were set to 1.0; those less than 0.05 were set to 0.05. Unit 1 is added to f_i to

Table 4.1: Consonant Phonological Representation

item	Labial	Dental	Alveolar	Palatal	Velar	Glottal	Stop	Fricative	Affricate	Nasal	Liquid	Glide	Voice	Front	Center	Back	High	Mid	Low	Tense	Retroflex	Round	Pre y	Post y	Post w
/_/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/p/	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/b/	1	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/t/	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/d/	0	0	1	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/k/	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/g/	0	0	0	0	1	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/f/	1	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/v/	1	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/θ/	0	1	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/ð/	0	1	0	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/s/	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/z/	0	0	1	0	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/ʃ/	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/ʒ/	0	0	0	1	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/h/	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/ɕ/	0	0	0	1	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/ʝ/	0	0	0	1	0	0	0	0	1	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/m/	1	0	0	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/n/	0	0	1	0	0	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/ɣ/	0	0	0	0	1	0	0	0	0	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/l/	0	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0
/r/	0	0	0	1	0	0	0	0	0	0	1	0	1	0	0	0	0	0	0	0	1	0	0	0	0
/w/	1	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0
/y/	0	0	0	1	0	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0

Table 4.2: Vowel Phonological Representation

item	Labial	Dental	Alveolar	Palatal	Velar	Glottal	Stop	Fricative	Affricate	Nasal	Liquid	Glide	Voice	Front	Center	Back	High	Mid	Low	Tense	Retroflex	Round	Pre y	Post y	Post w
/⌟/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
/i/	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	1	0	0	0	0	0
/ɪ/	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0
/e/	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	1	0	0	0	0	0
/ɛ/	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	1	0	0	0	0	0	0	0
/a/	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	0
/ə/	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0
/u/	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	1	0	0	0
/ʊ/	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	0	0	1	0	0	0
/o/	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	0	0
/ʌ/	0	0	0	0	0	0	0	0	0	0	0	0	1	0	1	0	0	1	0	0	0	0	0	0	0
/ju/	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	1	0	0	1	0	1	1	0	0
/aj/	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	1	0
/aw/	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	1	0	0	0	0	0	1
/oj/	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	1	0	1	0	1	0	1	0	1	0

prevent taking the $\log$ of zero. The values of f_i and m are divided by 5 to give a distribution of frequencies which is in greater conformance with Zipf's law (Zipf, 1935) (see HS98 for discussion of reasons for log compression of frequencies). The values p_i give the (unnormalized) probability of presentation for item i. The actual probability of presentation of an item is $p_i/\sum_i p_i$.

4.3 The Pre-Literate Model

First, a non-reading model was constructed, in which the phonological and semantic spaces are organized, and the mapping from phonology to semantics and back was trained. This model is intended to represent the state of a pre-reading child who has acquired a reasonable auditory vocabulary, and knows about the phonological structure of his language, and about the semantic organization of his world (e.g., the world contains objects, living things, animals, actions, and states). Further, previous work (Harm & Seidenberg, 1996) suggests that models of word recognition can be sensitive to structure in the phonological space, and that such structure exerts a qualitative impact on the learning the model performs. Similarly, structure in the semantic as well as phonological space has been shown to be important in connectionist models of reading (Plaut & Shallice, 1993, 1991). This work brings together these two threads of research.

4.3.1 Architecture

Figure 4.2 depicts the hearing model used in these simulations. The semantic component consisted of the 1,343 semantic features. These units were all connected to 50 units in the semantic cleanup apparatus. These units projected back onto the semantic features. This architecture, when trained properly, is capable of forming "attractors" in semantic space which repair noisy, partial or degraded patterns and tend to pull the state of the semantic units into consistent patterns (Plaut & Shallice, 1993).

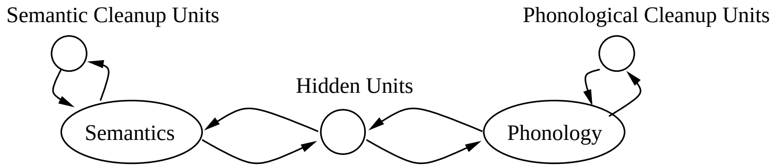


Figure 4.2: The "hearing" model. Semantic and phonological representations are organized, and the mapping from sound to semantics is learned.

The phonological units consisted of a set of 75 phonological units (3 slots of 25 units each). These projected onto a set of 25 phonological cleanup units (less units were used than in the semantic space because the phonological space has such fewer units than semantics. Early piloting revealed that 25 units were more than sufficient. See HS98 for more information). These cleanup units project back onto the phonological units. Here, again, an attractor network can be created which will repair partial or degraded phonological patterns. HS98 demonstrates the effect of this attractor, and damage to it, on the learning process. Briefly, the better the phonological attractor is at repairing or completing partial or degraded patterns, the less work a system which maps onto that representation has to do. This provides relief for the system mapping onto that representation, which allows for better learning (more robust, better generalization, etc.)

Because the presence of semantic and phonological attractors have been found to have important consequences in previous work, the present study incorporates these ideas.

The semantic space maps onto the phonological space via a set of 150 hidden units. There is feedback in both directions.

4.3.2 Training Regime

Online learning was used, with words selected for training according to their probability of presentation (Equation 4.6). There were four different training tasks which the network was exposed two, all of which were randomly interleaved. Once a word was selected for training, one of the four tasks was chosen, again at random.

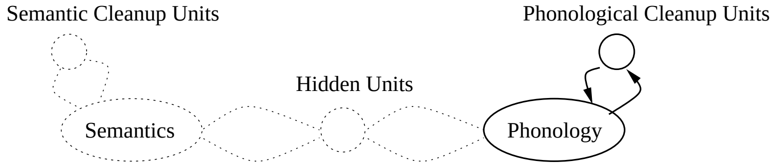


Figure 4.3: The Phonological Task. Untrained connections/units are shown in dashed lines.

1. The Phonological Task. Ten percent of the time, a word was trained on the Phonological task. This task develops the phonological attractor in the absence of its usage in an explicit, overt production or comprehension task; it is analogous to a child reflecting on the sound patterns of the utterances around him. There is evidence that children are in fact quite sensitive to such things. Nazzi, Bertoncini, and Mehler (1998) demonstrated that infants exhibit sensitivity to their language's rhythmic structure as early as such measurements can be made, suggesting that learning takes place in utero. Saffran and colleagues have demonstrated that children can pick up on the statistical regularities of streams of tokens presented for very short (2 minute) durations (Saffran, Aslin, & Newport, 1996), and when not explicitly attending to the content of the sounds they are exposed to (Saffran, Newport, Aslin, & Tunick, 1997).

The task used to train the model on the sound patterns of English was qualitatively identical to that used in HS98, except that it has been modified slightly to accommodate continuous time networks. The phonological form of the target word was clamped onto the phonological units for 3.5 units of time. Then a target signal was provided for the next 3.5 units of time, in which the network was required to retain the phonological pattern in the absence of external clamping. In HS98, auto connections were used to give the units a tendency to retain their value, but gradually decay. To accomplish the task, the network had to learn enough of the statistical regularities of the representations to prevent this decay. In the current simulations, the idea is the same, but since continuous time units are utilized, auto connections are not necessary to provide the units with a tendency to gradually decay; this was part of the units' normal processing dynamics.

When trained on the Phonological task, only the weights from the phonological units to the phonological cleanups and back were modified; the other units in the system that are responsible for semantics and the mapping from semantics to phonology were quiescent.

It was found in HS98 that training a phonological attractor on this task (which was called the Pattern Retention task) formed attractors which allowed the phonological representation to reliably repair corrupted phonological patterns. It also gave rise to categorical perception phenomena (see Repp, 1984, for a review) and phoneme restoration phenomena (Warren, 1970), and internal onset-rime distinctions (Treiman, 1986). Hence, the network was forced to absorb knowledge about the sound structure of English which had interpretable effects on other tasks.

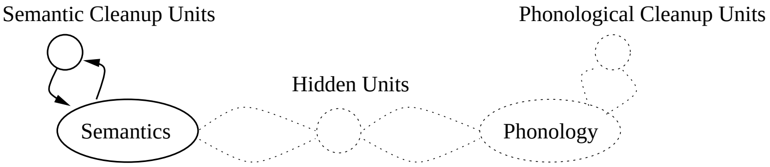


Figure 4.4: The Semantic Task. Untrained connections/units are shown in dashed lines.

2. The Semantic Task. Another 10 percent of the trials were devoted to training the semantic attractor. This task was constructed to be identical to the Phonological Task: the pattern of semantic units corresponding to the selected word was clamped onto the units for 3.5 units of time, and the network is allowed to cycle. Then the semantic units were unclamped, and the network must preserve their activity in the face of the tendency of the units' activity to decay away. To accomplish the task, the network must learn about intercorrelations between units, and form attractors which can preserve these states.

This task is more difficult than the Phonological Task, because there are far more semantic units than phonological units, and the mean intercorrelatedness of the units is far lower. As mentioned before, phonology is constrained by a physical process, which strongly delineates the representational space. Concepts have no such hard constraint. The notion that phonological attractors are stronger than semantic ones is of some theoretical importance, as will be shown. Suffice it to say that it was expected that the Semantic Task would not be as easily performed by the network as the Phonological Task, but that training it to its asymptotic level of performance was considered important nonetheless.

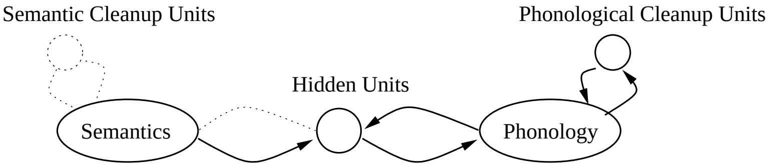


Figure 4.5: The Production Task. Untrained connections/units are shown in dashed lines.

3. Production (speaking). This task involved training the semantics to phonology pathway (sem→pho). It was used for 40% of the trials. It is loosely based on the task of producing an utterance; either a word in picture naming or free speech. No attempt is made here to model the development of early articulation, through babbling, mimicry or any other infant speech (see Plaut & Kello, In Press, for work in this direction). Nor is any attempt made to look at the use or formation of sequences of utterances or grammatical relations between words (see Allen & Seidenberg, In Press, for work on this topic). In its barest form, this task involves the production of the appropriate phonological form for a word given its semantic representation.

In training, the semantic pattern of a word was clamped onto the semantic units for the full 7 units of time. The output of the phonological units for the final 1.0 unit of time was compared with the target values; error was injected into the network according to the standard backprop equations. All weights were updated, except those leading back into semantics (since the values of the semantic units were clamped, no weight changes would have resulted here anyway). Importantly, the weights in the phonological attractor were trainable during this task.

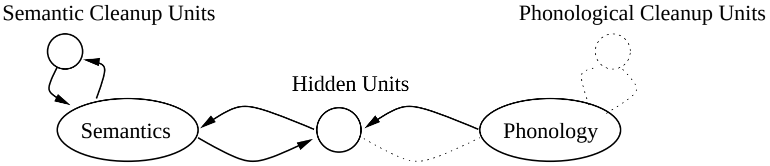


Figure 4.6: The Comprehension Task. Untrained connections/units are shown in dashed lines.

4. Comprehension (hearing). Finally, words were trained on this task 40% of the time. This is the mirror image of the Production task. The phonological form of a word was clamped onto the phonological units for the full 7 units of time. During the final 1.0 unit of time the output of the semantic units was compared with their targets.[1]

As a side note, it should be mentioned that the training of the phonological and semantic spaces independent of production or comprehension tasks may seem odd. Certainly, the phonological and semantic attractor spaces undergo training during production and comprehension, and the resultant representations are influenced by their being marshaled into these two tasks. HS98, for example, explored the influence of reading on the phonological representation. However, as mentioned earlier, there is empirical evidence that children attend to the sound structure of their language independent of its explicit usage in production or comprehension. As for semantics, it is largely a philosophical question whether one knows things about objects in the world independent of communication, but I think that view is defensible. Certainly one would not want to take the strong position that the only way you can know that [goats] have the property <+*smell_bad*> is by being told this, or telling someone this. There is an entire visual object recognition system that presumably maps onto semantics which is not implemented here, and the semantic space is probably structured by its usage in the different tasks of visual and auditory perception, much in the same way that phonology has been found to be structured based on the dual tasks of auditory comprehension and reading. Explorations of the domain of visual cognition is clearly outside of this work, but can provide a (rough) justification of the Semantic Task and the use of training it independent of an explicit production task.

The model was trained for 700,000 word presentations (approximately 280,000 Production, 280,000 Comprehension, 70,000 Semantic and 70,000 Phonological trials). The most frequent word (THE) was presented about 947 times, the least frequent words ($p_i = 0.05$; e.g., CUD) an

[1] No attempt was made here to distinguish optional from necessary or defining semantic features. Certainly, upon hearing a word one does not necessarily activate all of one's knowledge of the relevant concept; such things have been found to be dependent on factors such as the context of presentation. For example, the concept [piano] would activate <+*heavy*> in the context of moving, but maybe not in the context of a recital (Barclay et al., 1974). Such variation is interesting, and of some importance in various kinds of ambiguity resolution, but beyond the scope of this work.

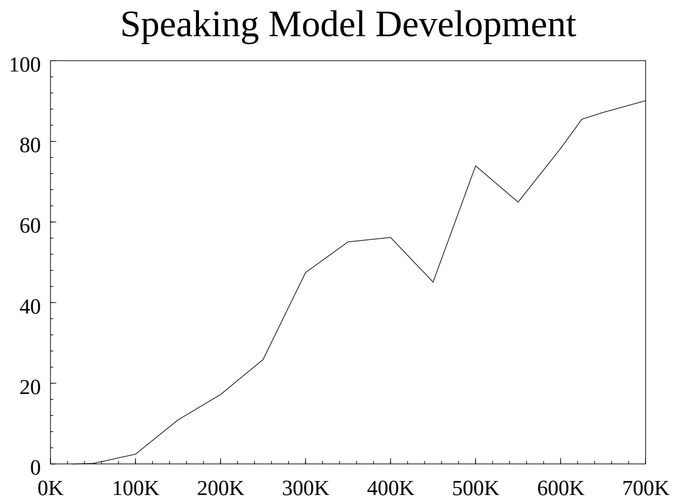


Figure 4.7: The development of the speaking model.

average of 47 times each. A learning rate of 0.2 was used for 500,000 word presentations, then lowered to 0.1 for the remaining 200,000 word presentations.[2]

4.3.3 Results of Training

Figure 4.7 depicts the development of accuracy for the speaking model. At asymptote, the model correctly generates pronunciations for 90% of the items by type (99% by token). The performance of the hearing model is shown in Figure 4.8. Because there are a large number of homophones in the training corpus (565 items are members of a homophone group), the performance is broken out by the whole corpus, the non-homophonous (i.e., unambiguous) items, and the homophonous (i.e., ambiguous) items. At asymptote, the model scores 87% of the non-homophonous items perfectly. The items it makes mistakes on are limited to one or two incorrect semantic features (for example, it recognizes the item BAY as having the features <*object*> and <*body_of_water*>, but not <*R36*>, which is the randomly generated feature which distinguishes BAY from COVE). The model therefore is scored as incorrectly computing the full semantics of BAY, by producing a representation that is identical to COVE. Virtually none of the non-homophonous items are totally incorrect.

For the homophones, about 25% of them are correctly recognized, meaning that the network selected a dominant member and computed its representation exactly. This corresponds to half of the homophone pairs, with one member perfect and the other totally wrong. For the remaining homophone items, the network produces a mix of features, splitting the difference between homophonous items. For example, ALE is interpreted as <*foodstuff*> at activity level 0.59, and an ailment, with <*suffer*> at activity level 0.8. This is characteristic; the network's semantic units are

[2] Beginning with a high learning rate and then lowering it during training often results in faster convergence than either maintaining a high learning rate (which can lead to network oscillations), or starting with a lower one (which can dramatically slow initial learning.)

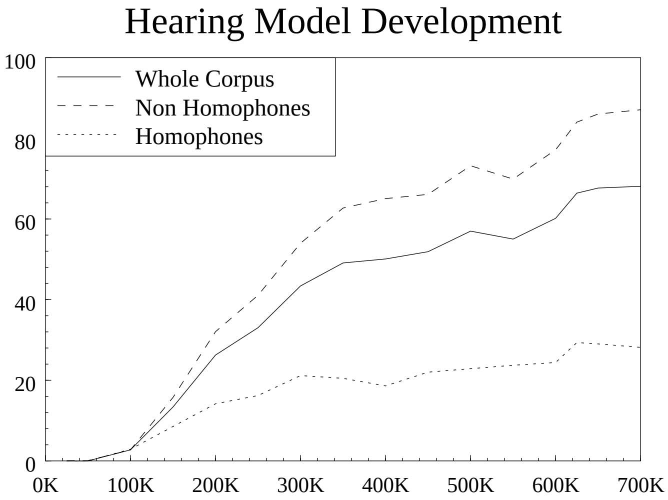


Figure 4.8: The development of the hearing model.

not driven to extreme values for either interpretation. Such activation reflects the ambiguity of the phonological form; the network is "on the fence" as to which interpretation is correct.

The addition of a small amount of context will resolve the ambiguity, however. If the <*alcohol*> feature is activated, and all other semantic features are left unspecified, then upon presentation of the phonological form of ALE the semantic output is the unambiguous representation of ALE. If instead the <*be*> feature is activated (indicating a state of being), then all features relevant to ALE are suppressed and the output is a veridical representation of AIL (consisting of <*bad*>, <*suffer*> and <*be*>). The model therefore will produces uncertain or borderline activations for ambiguous words presented in isolation, but when small amounts of context or expectation is presented (which could derive from earlier items in a sentence, or a discourse context) then the item is able to be correctly identified.

While an exploration of the use of context in lexical ambiguity resolution is far beyond the scope of this work, this behavior of the model is promising. The model shows sensitivity to frequency differences in ambiguous items (this will be explored in greater detail in Chapters 5 and 6), but also exhibits sensitivity to the effect of prior context. Empirical work in sentence processing has identified the similar and differential effects of these two sources of constraint (MacDonald, 1993) and their respective time courses of utilization (Stevens, Harm, Schuster, & MacDonald, 1995).

4.4 The Reading Model

4.4.1 Architecture

Figure 4.9 shows the architecture of the reading model. The top section is the hearing/speaking model described in the previous section. A set of 87 orthographic features, representing a slot based localist representation were used to represent the spelling of a word. The position of each

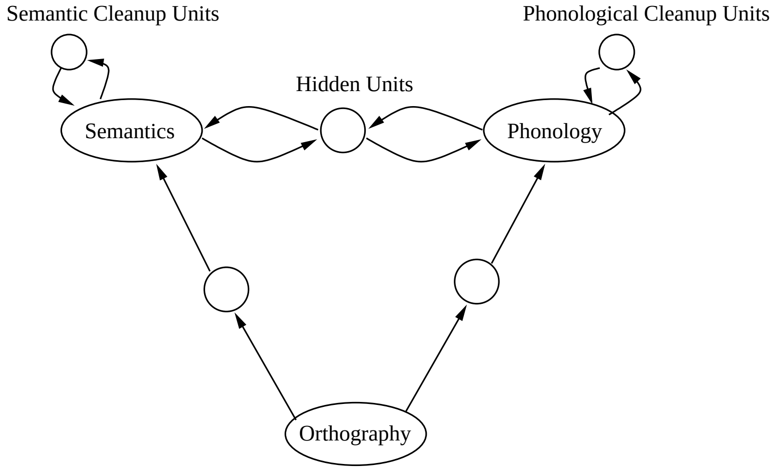


Figure 4.9: Implemented reading model. The top section is taken from the model shown in Figure 4.2.

of these units was computed based on the set of words in the training set. One hundred hidden units mediate the mapping from these orthographic units to semantics, forming the orth→sem pathway. Similarly, a set of one hundred other hidden units mediate the orth→phon pathway. The architecture and processing dynamics of the hearing/speaking model was identical to that used in the hear/speak prestructuring phase.

4.4.2 Training Regime

At the conclusion of hearing/speaking training, the weights from the hear/speak model were frozen and added to the larger reading model. Freezing the weights is not strictly necessary; earlier work (Harm & Seidenberg, 1997) used a process of *interleaving* in which hearing trials were used along with reading trials. Such a scheme prevented *catastrophic interference* (McCloskey & Cohen, 1989),[3] a phenomena whereby a network trained on one task and then used on a second task can tend to "forget" the original task. Interleaving tasks (as is naturally seen in the child learning to read; children who begin formal literacy instruction do not cease to hear and speak) blocks this tendency. Weight freezing is another way, which is less computationally burdensome and has very similar end results. Other approaches to preventing catastrophic interference, which involve hippocampal playback of items are also possible (McClelland, McNaughton, & O'Reilly, 1995).

Items were presented to the network according to the same online learning scheme as before, with the same frequency distributions. A simplifying assumption was made that the distribution of words does not change during the development of reading. This is of course unrealistic; a

[3] To demonstrate catastrophic interference, McCloskey and Cohen used an arithmetic task, with a bank of units representing one digit, another representing an operator (addition, subtraction, etc), and one last bank for the final digit. Training was blocked such that their network had to learn to add 1 to various numbers. Then they trained it to learn to add 2, and they found that it "forgot" how to add 1. But this is an artifact of the training regime and representation. The network, having been trained on adding 1, had no way to know that 1 is an *argument* and the + symbol was an *operator*. It had, in fact, no way to know that those units were not simply biases, which were constantly on. This bears no resemblance to any task a human faces, and so its force as a criticism of connectionist learning principles has little force (see also Hetherington & Seidenberg, 1989).

model which more tightly followed the development of children's literacy would provide frequency distributions of words according to the curriculum being modeled.[4] Future work will focus directly on the size and composition of a child's vocabulary and the ways in which the distribution of items affects the acquisition of literacy.

Error signals were provided for both the phonological and semantic representations of a word. Again, this is unrealistic. Different curricula emphasize or de-emphasize pronunciation of words versus the comprehension of connected text. Within this framework, one could explore the impact of different degrees of feedback on pronunciation or meaning; such an exploration is beyond the scope of the current project but is a topic to be investigated in the near future.

To computationally instantiate the principle that the reading system is under pressure to perform as rapidly as possible, error was injected into the semantic and phonological representations early; from timestep 2 to 14. The network therefore received an error signal not only if it did not produce the correct pronunciations and meanings of words, but if it did not *rapidly* produce them.

4.4.3 Results of Training

The network was trained for 3 million word presentations. At the conclusion of training, the network produced the exact correct semantic representations for 93% of the items. For items it did not compute the exact correct semantics for, it was off by an average of 1.2 features. It produced the correct phonological representation for 98.2% of the items.

I now will turn to more detailed analyses of the reading model's behavior and performance throughout training.

[4] It is an interesting policy issue how children's initial lessons should be structured. Many researchers believe that children should be provided with very simplistic words initially; others say children need a rich, interesting set of things to read and "see Jane run" basals only serve to bore the child. See Adams (1990) for discussion.

Chapter 5

DOL To Semantics

In this chapter, the question of the model's division of labor in the computation of a word's semantics is considered. The factors to be considered are: effects of skill level, the word's frequency, the word's regularity, interactions of these effects, and homophony. Empirical research reviewed in the Introduction has suggested the importance of these factors, so they each will be considered in turn, and discussed in terms of the model's behavior.

5.1 Method: Lesion Studies

The primary methods of analysis were lesion studies. To evaluate the efficiency of the orth→sem route, a lesion was applied to the orth→phonroute and the accuracy of the network's semantic output was measured. Similarly, to evaluate the orth→phon→semroute, the orth→semroute was lesioned and the accuracy of the remaining network was measured.

This method is not perfect. In any complex, interactive system, the behavior of an isolated component is not always inferable from the behavior of the system with that component withdrawn. Imagine trying to reverse engineer a computer motherboard by systematically breaking off one chip at a time, and observing the behavior of the system. However, the method has a long tradition in behavioral neuropsychology (see Shallice, 1988, for a historical perspective). Further, the picture is not as bleak as the motherboard example would suggest. Here we are not really trying to infer the function of the component which we are removing, but rather the capacity of the remaining system. If the floppy drive of a computer works perfectly when the video card is removed, we can infer that the video card is not necessary for the functioning of the floppy drive. The approach taken here is similar. More subtle analyses will follow, which will be directed at questions such as "well, if component X is not necessary for the accomplishment of task Y, does having component X *help* in the accomplishment of task Y?" In particular, while it is important to consider the accuracy of a given path, it is also important to consider its potential role in the speed of a computation.

5.2 Effect of Skill Level

The accuracy measures of the intact model, a model with only a direct/semantic path to semantics and one with only a phonological path to semantics were computed along varying stages of development. The results are presented in Figure 5.1.

The accuracy of the intact model rises rapidly, then flattens out, growing more slowly for the bulk of training time. Initially, the accuracy of the intact model and the model with only a phonological route parallel each other, indicating that phonological processing is doing most of the work. Quickly, however, the performance of the intact model surpasses that of the phonology-only model, whose performance reaches asymptote and does not improve further. At this point early in development (about 250K iterations), the increase in competence of the intact model matches the increases in the competence of the semantics-only model.

Importantly, the development of the orth→sem path continues even after the intact model has essentially reached asymptote, at about 1.5M iterations.

One thing that is not clear from Figure 5.1 is the degree of overlap in the words that are read by the different pathways. Is it the case that the words are partitioned into orth→phon→sem versus orth→sem, or are some words redundantly supported by either route, and some require both routes? It is impossible to tell. So, words correctly read by the intact network were further broken down into four distinct subgroups: those that require both pathways to be read (cannot be read by either path in isolation), those that can be read by either pathway, those that can be read by orth→sem but not orth→phon→sem, and those that can be read by orth→phon→sem but not orth→sem. Figure 5.2 shows this breakdown over the course of development.

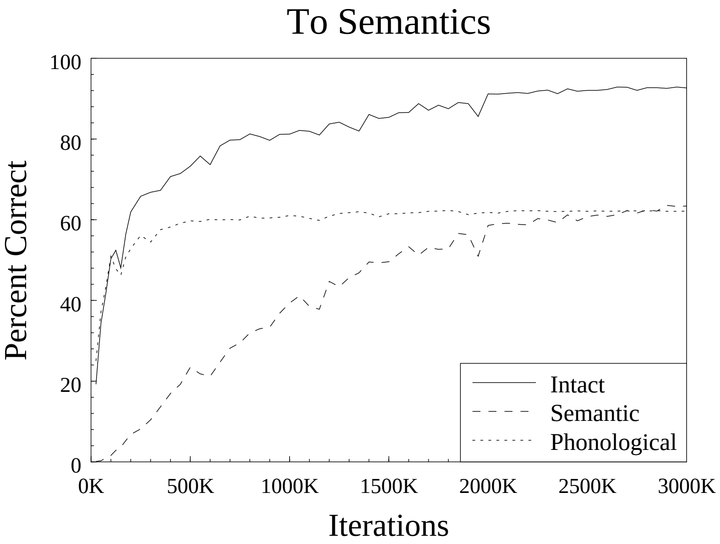


Figure 5.1: Accuracy of activation of semantics, by intact model, model with only semantic and phonological paths.

As expected, there is an initial burst of words that can be read only by the phonological route. Early on, the phonological route is the dominant method of reading. This begins to fall off by 1M iterations, at which point more words can be read by either route. Interestingly, at that point about 10% of the items can only be read by the orth→sem route. This number grows to about 15%, where it tops out. Asymptotically, about half of the words are redundant; they can be read accurately by either route. Approximately equal numbers of items can be read only by the conjoined cooperation of both routes, by orth→phon→sem, or by orth→sem.

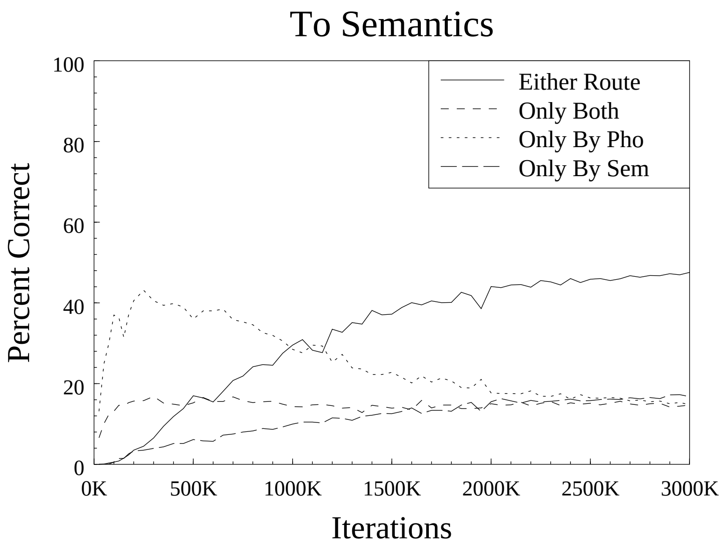


Figure 5.2: Accuracy of activation of semantics. Percent of items which can be read only by both routes, by either route, by the phono but not semantic route, and the semantic but not phonological route are shown.

This effect provides a nice accounting of disparate findings that have been seen in the literature: the importance of phonology and phonological skills on early reading acquisition, and its relatively reduced importance for skilled readers. The picture that emerges is similar in spirit to the "phonological recoding" idea of Share (1995), but realized in a different way. It is not the case that the phonological loop "trains up" the direct recognition system, but rather that the two systems are responding to the task according to their inherent computational properties: orth→phon is correlated, phon→sem is known, and orth→sem is difficult but fast. The system does not "decide" to switch strategies, but rather responds to the task it is assigned: computing the meaning of the word as rapidly as possible, subject to intrinsic computational constraints.

5.3 Speed Effects

The time course of activation of semantics was next measured. An item was assumed to be correctly recognized when all semantic features were within 0.2 of their target values; the amount of time until a word was correctly recognized was computed for all items, and averaged. This measure was taken at various points in development, and is shown in Figure 5.3. The time it takes the network to recognize the average item decreases over time, and eventually flattens out. The network is pressured not only to recognize items but to do so rapidly.

As noted in the previous section, a number of words can be read by either route in isolation. This fact masks a subtle but important point that is revealed by the latency analyses. Within this model, the effect of the two routes working together is different from the effect of the two in isolation (contrast with the "horse race" model of Paap & Noel, 1991).

The 728 words which are redundantly read (i.e. they can be recognized by either pathway in isolation) were analyzed for their latencies in the intact model, the orth→sem model and the

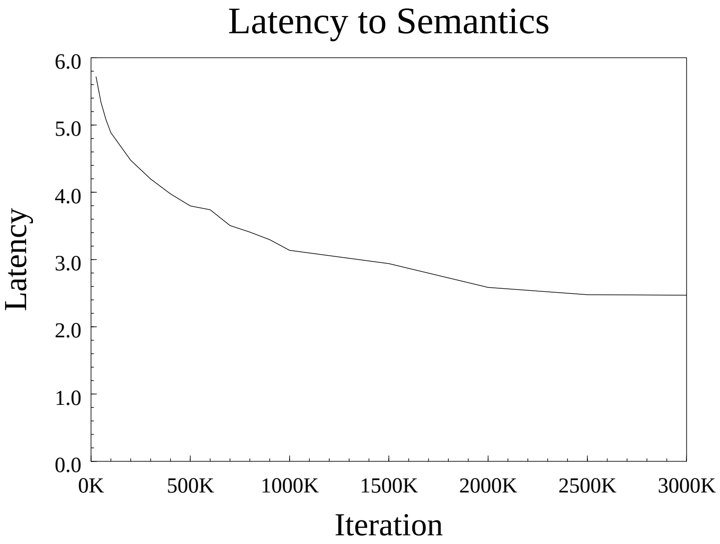


Figure 5.3: Semantic latencies.

orth→phon→sem model. The results are depicted in Figure 5.4. Clearly, the latencies for the intact model are much lower than either pathway operating independently. The aggregate, cooperative contribution of both processing routes drives the response time of the semantic representations.

5.4 Frequency Effects

As reviewed in the introduction, there is evidence that the relative contribution of a given pathway is modulated by the frequency of the item. The division of labor analysis was conducted, splitting items by their frequency into low or high frequency. Items with a probability of presentation of 0.7 or higher were coded as high frequency; items lower were low frequency.

Figure 5.5 shows the accuracy of the direct, orth→sem route, broken down by item frequency. Both items grow in proficiency over time. However, throughout development, the high frequency items are strongly dominant over the low frequency items. The direct, orth→sem route favors high frequency items, by a large degree.

Figure 5.6 shows the frequency breakdown for the phonological route. Here again, high frequency items are dominant. But the difference is not so dramatic; the low frequency items enjoy a much larger stature in the phonological route.

So is there a frequency modulation of the division of labor to semantics? It is difficult to see from Figures 5.5 and 5.6. If the asymptotic state of the network is analyzed, and plotted by frequency, it is easier to see an interesting effect. Figure 5.7 shows the breakdown of items by high and low frequency, and by percentage correct by path.

Overall, as expected, the high frequency items are more accurate than the low frequency ones. However, a partial interaction is seen which modulates this effect. The low frequency items lean on the two paths to approximately the same degree. The high frequency items show a different effect; they rely more on the orth→sem path than the orth→phon→sem path.

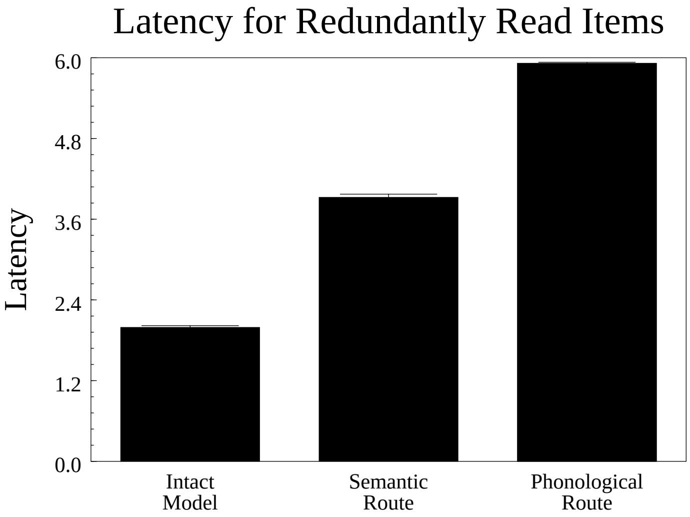


Figure 5.4: Semantic latencies for redundantly read words.

Recall that the model is pressured to produce the semantics of the word as rapidly as possible. Over the course of training, this error has an additive effect on the network weights. Words are presented probabilistically, so the network is under much greater pressure to optimize the high frequency items over the low frequency items. This drives the network to read high frequency items by the rapid orth→sem route.

Such a view can be taken as a counter to Smith's (1973) argument about efficiency. Smith argued that reading is accomplished too fast to accommodate phonological recoding. However, Zipf's law (Zipf, 1935) states that there is a constant relationship between the number of words at a given frequency range and the square of that frequency range; i.e. the frequency histogram for any language follows a curve $y = k/x^2$, for some constant k. Only the most highly frequent items tend to violate this relationship. What it means is that there is a very small set of words which occur very frequently, and a very large number of words which are much more infrequent.

Let's take an example. Suppose that phonological reading (via orth→phon→sem) takes twice as long as direct, orth→sem access of meaning. If 10,000 monosyllables are read by direct orth→sem encoding, it would take 10,000 units of time; 20,000 if they are read phonologically. But if just the most frequent 5% of words (by type) are read directly by orth→sem, and the remaining 95% were read by the slower orth→phon→sem route, then it would only take 11,089 units of time. If only the most frequent 2% of the items were read this way, it would take 12,195 units of time. The point is simple: arguments about efficiency of encoding do not necessitate orth→sem reading for *all* items if one considers the frequency distribution of items in the language. Reading just a few high frequency items by the rapid orth→sem route has a dramatic effect on the overall latency of comprehending a message.

This analysis considered the effects of frequency on the DOL to semantics. However, frequency is not independent of regularity, another factor which has been found empirically to be relevant to the computation. The effects of regularity will be explored next, and finally, interactions between frequency and regularity will be analyzed.

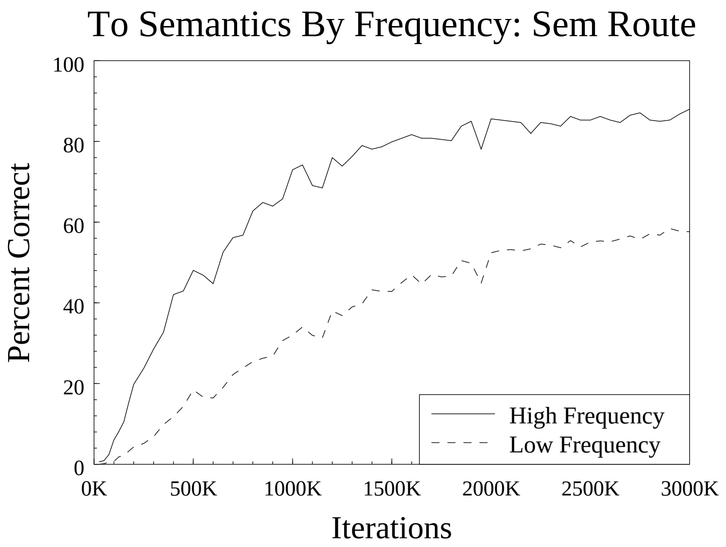


Figure 5.5: DOL to Semantics by Direct route, by Frequency.

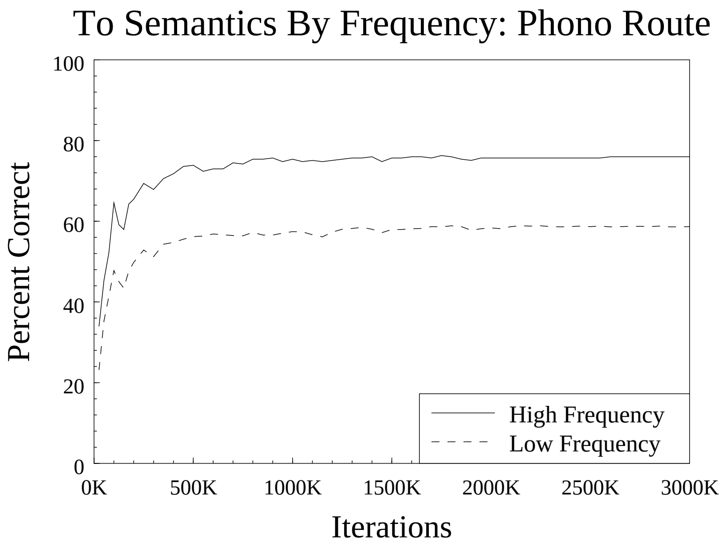


Figure 5.6: DOL to Semantics by Phono route, by Frequency.

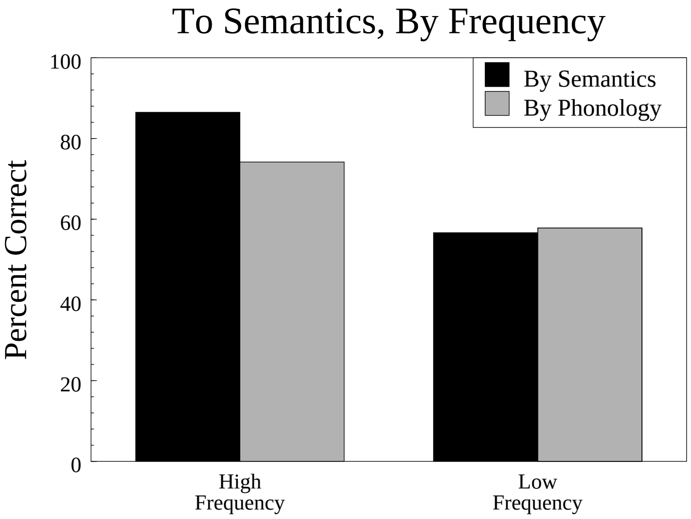


Figure 5.7: Asymptotic frequency effect on computation of semantics.

5.5 Regularity Effects

Words in the training set were divided up according to their regularity. Regularity was assessed using a *friends and enemies* measure (Jared, McRae, & Seidenberg, 1990). Each word's orthographic rime was analyzed. All items with that orthographic rime which have the same phonological pronunciation are counted as *friends* of that word. Words with that rime but a differing pronunciation are *enemies*. If a word has no enemies, it is *regular* (e.g., BAT). If it has some enemies, but more friends than enemies, it is *regular-inconsistent* (e.g., GAVE; contrast with HAVE). If a word has more enemies than friends, it is an *exception* (e.g., the ubiquitous HAVE). Finally, if a word has no friends nor enemies, it is *strange* (e.g., YACHT). For the purposes of this analysis, exception and strange words are considered *irregular*; they do not get collateral benefit from a set of other items. Regular and regular-inconsistent items are grouped under the heading *regular*.

An analysis similar to the previous section was conducted. Figure 5.8 shows the results by the direct route. Here, it is clear that exception items enjoy a slight benefit over the regulars by this route.

Contrast with Figure 5.9. Here, a mirror image is seen: the regular items are more readily read by the phonological route than exceptions. This effect is robust throughout the course of training.

Finally, Figure 5.10 shows an interesting interaction of the effect of regularity on the efficiency of pathway. Regular items are more efficiently processed by the orth→phon→sem pathway. Exceptions show a reciprocal effect; they are more efficiently processed by the orth→sem pathway. The explanation of such effects in the orth→phon→sem route is straightforward: while both exception and regular items can be read by orth→phon in the model, the regular items are more easily read by this route, because they are more able to exploit regularities in the mapping.

These reciprocal effects in the orth→sem path are difficult to interpret, given that exceptions tend to be higher in frequency than regulars (Bybee, 1988). Thus, the superiority of exceptions by the orth→sem route could simply be a duplication of the effects of Figure 5.7. Analogously, the

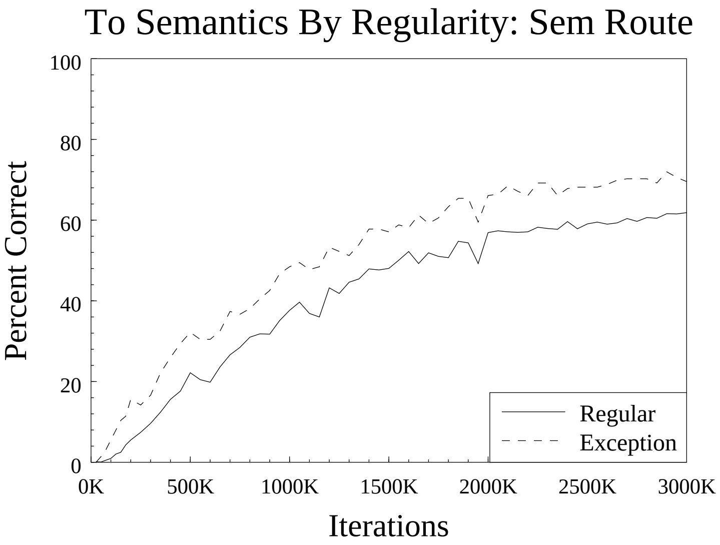


Figure 5.8: DOL to semantics by regularity, by direct route.

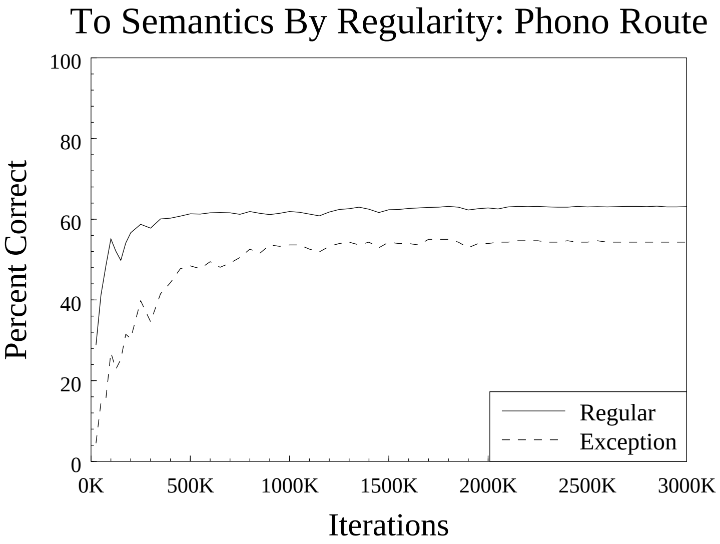


Figure 5.9: DOL to semantics by regularity, by phonological route.

effects of frequency are difficult to interpret in isolation when regularity is not factored in. The conjoined effects of frequency and regularity are therefore considered next.

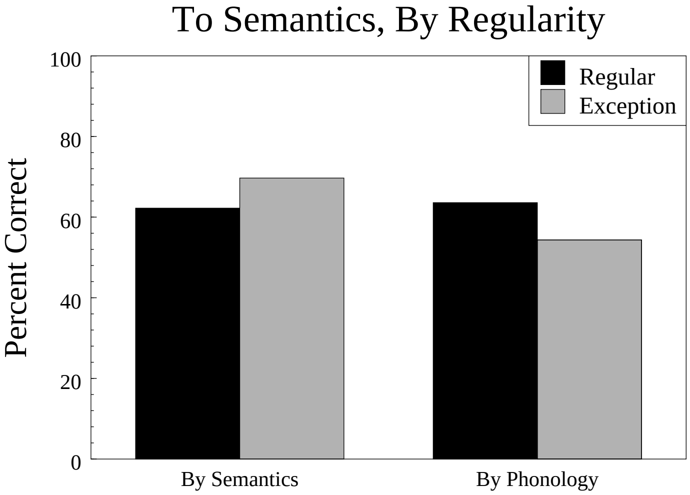


Figure 5.10: Activation of semantics by regularity, by pathway.

5.6 Interaction of Frequency and Regularity

Using the definitions of frequency and regularity adopted above, words were broken into four cells: high frequency regulars, high frequency exceptions, low frequency regulars and low frequency exceptions. The asymptotic analyses were conducted on the model.

Figure 5.11 shows the effects of frequency and regularity along the phonological route. For the phonological route, high frequency items are more accurate than low, as is normal. Within both groups, the regulars show and advantage over the exceptions. In Figure 5.12, however, we see the frequency by regularity effect for the orth→sem route. Here, there is no regularity effect for the high frequency items. However, the exceptions show a small advantage over the regulars for the low frequency items. High frequency items get a boost overall; for the low frequency items, it is the items that are least easily read by the phonological route that depend most on the orth→sem route.

These results reflect the conjoined effects of regularity and frequency on the division of labor. Regularity shows a strong effect on the orth→phon→sem pathway for both low and high frequency items. The orth→sem pathway sees a reciprocal regularity effect, but only for low frequency items (the high frequency ones showing a ceiling effect). The effect of regularity on the low frequency items reflects the orth→sem pathway making up for the orth→phon→sem pathway's limitations: it reads exceptions better than regulars, because the other pathway shows the reverse effect. It is picking up the slack left by the orth→phon→sem pathway.

The notion of orth→sem accomodating the weaknesses of the orth→phon→sem pathway plays a crucial role in understanding the effect of homophony on the DOL, as will be shown in the next section.

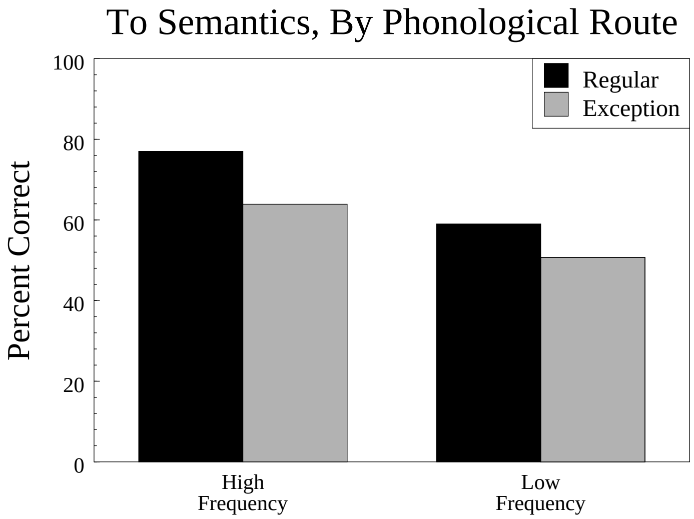


Figure 5.11: Frequency by regularity effects, to semantics, along the phonological path.

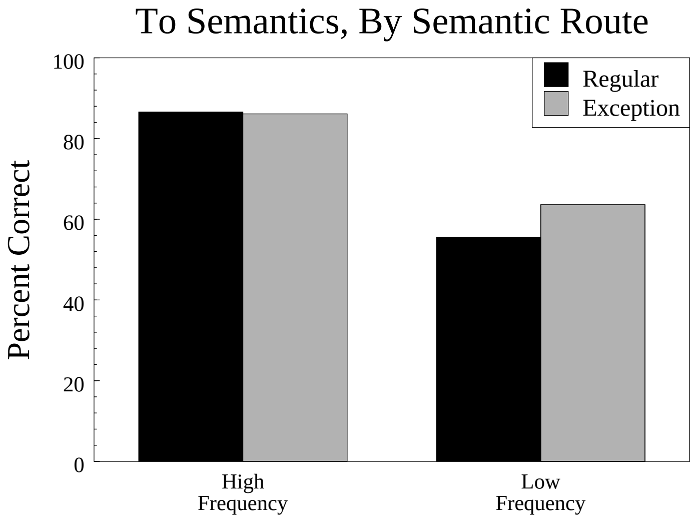


Figure 5.12: Frequency by regularity effects, to semantics, along the semantic route.

5.7 Homophones: Jared and Seidenberg 1991

As discussed in the introduction, a study by Jared and Seidenberg (1991) (JS91) provided evidence for the modulation of the division of labor by the frequency of words. In their study, subjects performed a semantic decision task (e.g., “is it an object?”). In one condition, target items were either exemplars (MEAT), a homophonous foil (MEET), or a spelling control (MEAN). In a second, target items could also be pseudohomophones of a target (e.g., TABUL for TABLE). Items were broken down by the frequency of the exemplar, and the frequency or word status of the foil (hf foil, lf foil, or pseudohomophone foil). The number of false positives for each foil above and beyond the spelling controls was measured.

Figure 5.13 shows the results. It was found that the only conditions which gave a significant number of false positives were ones where the exemplar was low in frequency, and the foil was either low in frequency as well, or was a pseudohomophone.

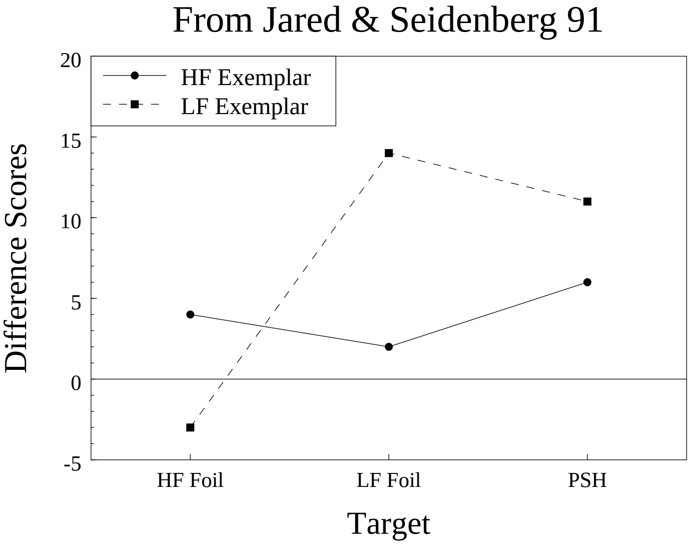


Figure 5.13: The Jared and Seidenberg (1991) results. False positives occur only when a target is either a low frequency foil or pseudohomophone, and the relevant exemplar is also low in frequency.

These results are a bit mysterious. From the previous results, it is easy to see why only a low frequency foil, and not a high frequency one, would result in false positives. High frequency items are more likely to benefit from the direct orth→sem route than low frequency ones; the orth→sem route is not “fooled” by homophony the way the orth→phon→sem route would be. As more orthographic information is available to the semantic system, the chance of a false positive for a homophone decreases. What is mysterious is the effect of the exemplar frequency on the tendency of homophone and pseudohomophone foils to produce false positives. Why, exactly, should the frequency of MEAT modulate whether or not one gets a false positive for MEET? Why would the frequency of TABLE affect whether a pseudohomophone like TABUL produces a false positive? JS91 were not able to provide a definitive answer, instead emphasizing that the finding of a frequency effect argued against the strong position taken by Van Orden and colleagues (Van Orden,

		Exemplar		Foil	Spelling	
Category	Exemplar	Frequency	Foil	Frequency	Control	Pseudohom
Object	ale	LF	ail	LF	ace	eil
Object	cot	LF	caught	HF	cat	cought
Object	load	HF	lode	LF	loud	lowed
Life Form	bear	HF	bare	HF	beer	bair

Table 5.1: Sample stimuli for JS91 replication

1987; Van Orden et al., 1988, 1990) that orth→sem does not initially influence the computation of semantics for *any* word, regardless of its frequency or any other factor.

To explore why these results obtain, a replication of the JS91 effect was attempted. Stimuli were chosen algorithmically as follows. All items in the training set were divided into the categories of object, living thing, or other, based on the presence or absence of the semantic features <*object*> and <*life_form*>. Items which are objects or living things are candidates for exemplars. Items which are not objects are candidates to be a foil or spelling control for object exemplars. Those which are not living things are candidates to be foils or spelling controls for living thing exemplars.

For each word which was a candidate exemplar, the set of words which were candidates for homophone foils was scanned, to see if an item with the same pronunciation was found. If an item was found, it was recorded as a homophone foil for the exemplar. Then the set of candidates were also scanned for items with the same number of letters as the exemplar, the same initial letter, and whose spelling differed by at most one letter from the exemplar. If a match was found, that was the spelling control for the exemplar. Finally, a pseudohomophone was generated algorithmically, if possible, by taking the exemplar's orthographic onset, and swapping in a different orthographic rime which has the same pronunciation. All foils and exemplars were then coded according to frequency based on a median split of the frequencies of items; those above the median were high frequency, those below were low.[1] Table 5.1 shows a sample set of items; a total of 93 matched items resulted.

The JS paradigm was simulated by presenting the foils, spelling controls and pseudohomophones to the intact model, and observing the activation on the semantic feature for the exemplar. For example, if CAUGHT was presented to the model, the <*object*> feature would be monitored. The activity of the semantic feature was recorded when the network had proceeded for 4 units of time; short of the normal, full recognition time cycle of 7 units. This was to mimic the pressure for a rapid response which subjects in the experiment would be subject to. The idea was that spurious activity on the inappropriate semantic feature is analogous to a false positive response by a subject; if <*life_form*> was active when the network was presented with BARE, then that corresponds to a false positive.

As per JS91, false positives for pseudohomophones and foils was subtracted from their matched controls. The resulting items were subjected to a 2x3 analysis: high and low frequency exemplar by hf foil, lf foil or pseudohomophone foil.

[1] This method differs from the preceding sections, because the standard frequency split resulted in far too few high frequency items.

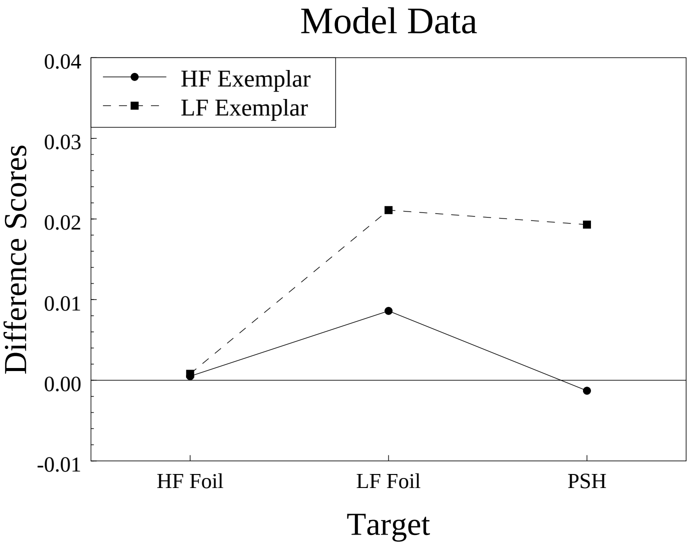


Figure 5.14: Replication of Jared and Seidenberg (1991) results. Model exhibits spurious activation on incorrect semantic feature only when a target is either a low frequency foil or pseudohomophone, and the relevant exemplar is also low in frequency.

The results for the high and low frequency foils by high and low frequency exemplars successfully replicated the JS91 results (Figure 5.14). However, the pseudohomophones produced far too many false positives. Considering the architecture of the implemented model, it is easy to see why this is. When a pseudohomophone such as BAIR is presented to the model, very little activation proceeds along orth→sem. The nature of the mapping of orth→sem leads to point attractors; items nearby in spelling space do not produce generalizations the way they do for orth→phon. So BAIR activates the phonological form /bεr/, which activates the <*life_form*> feature, and a false positive results. The model has no mechanism by which it can reject BAIR, that is, it could not decide that BAIR is not an living thing, because it could not decide that BAIR is not a word.

5.7.1 Interlude: A Stab at Lexical Decision

Intuitively, subjects can reject BAIR on the basis of a conscious spelling check. They read BAIR, the meaning of [bear] becomes active, but they know that [bear] is not spelled BAIR. Connections from semantics to orthography were not part of the implemented model, yet this is where such knowledge would be contained.

To allow for simulation of the JS91 task, then, a simple addition was made to the triangle framework. An additional mapping from semantics to orthography was implemented (Figure 5.15). This was trained up so that semantic activity would produce an orthographic representation of the spelling of that item.

At this point, the decision process for determining a false positive needed to be augmented. A "no" response to the question "is it an object" could be obtained by low activity on the <*object*> feature. If the <*object*> feature was active, however, the spelling for the semantic representation at that moment in time (again, after 5 units of time), as computed by the sem→orth pathway, was

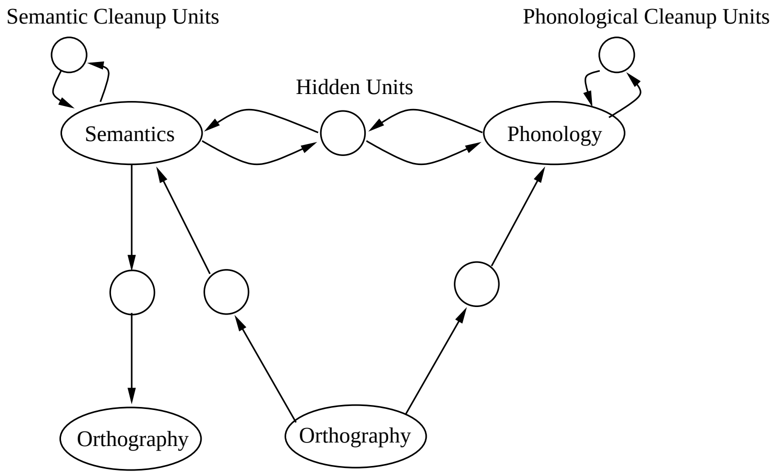


Figure 5.15: Reading model with spelling check.

determined. This was compared to the input spelling pattern. If the euclidean distance between these two spelling patterns was greater than 1.0, but forms a coherent pattern (values reasonably extremal), then a rejection occurs (that is, a “no” response). The test for extremal values is necessary so the model does not reject as a nonword an item which it does not actually know the spelling of (see below for more).

Hence, the model can respond “no” to a question about whether a word is an object based on computed semantics, or on the basis of a decision process modeled with a simple spelling check.

Such a mechanism may provide the beginnings of a full computational implementation of the lexical decision task. Previous work (Plaut, 1997) simulated lexical decision by measuring the stress on semantic units. This method does not work here for two reasons. First, the semantic attractor tends to draw non-extremal values to extremal points. Plaut (1997) did not have this problem because he did not incorporate a semantic attractor in his simulations. Simple activity in semantics is not sufficient, because pseudohomophones produce such activity, and if a network is trained on phon→sem prior to reading, then pseudohomophones would, in all likelihood, produce strong activity in semantics. Similarly, activity along orth→sem is not adequate, because many words are not read along that pathway. We need to distinguish words whose sound we know but which we may have never seen in print, and hence do not know the spelling of, from words which we *do* know the spelling of, which create a mismatch. If I had never seen the word CHUTNEY, but had heard it, and my grocery store suddenly began carrying cans of chutney, I would not reject the label CHUTNEY as a nonword on the basis of my failure to map sem→orth for that word, because I don’t *know* the spelling of CHUTNEY. Contrast with BAIR, where I do know the sem→orth relation for the concept [bear], and hence am able to make a rejection of BAIR on the basis of this knowledge. The implementation of the sem→orth pathway, and the test for extremal values in spelling is a way to operationalize the meta-knowledge that seems to be necessary for this task (that is, the decision “I know the spelling of that word, and that’s not it!”).

Having added this machinery to the simulated decision process, the experiment was repeated. This method managed to correctly reject 94% percent of the pseudohomophones, while only falsely rejecting 3% of the actual words. Figure 5.14 shows the full data for the JS91 replication. There is a close qualitative match to the study (Figure 5.13).

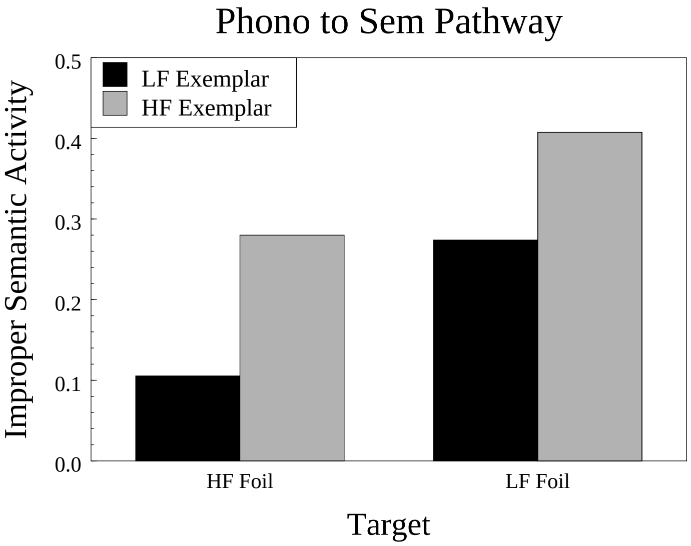


Figure 5.16: Spurious semantic activation from homophones on Pho to Sem. Items which trigger the most false positives are low frequency foils with high frequency exemplars.

Why do these effects obtain? The earlier analysis of frequency effects demonstrates that high frequency items are better able to be read via orth→sem than low frequency ones, so the finding that high frequency foils do not result in false positives is simple to explain. However, it is less clear why low frequency foils and pseudohomophones of high frequency exemplars do not also show false positives. The subject, and the model, does not see the exemplar in the trial; hence, why should its frequency matter?

Consider the foils first. Figure 5.16 shows the spurious activation of incorrect semantic features in the phon→sem model when the phonological form of a word is presented. For example, if the phonological form of MEET is presented, the activity of the <*object*> feature is recorded.

High frequency foils with low frequency exemplars cause the least degree of spurious activation, as would be expected. Similarly, the most problematic items are the low frequency foils with a high frequency exemplar (e.g., GAIT, which causes activity on <*object*> due to GATE). Items where the foil and exemplar are matched, either both low frequency, or both high frequency form the intermediate cases.

So phonologically, the items with high frequency exemplars and low frequency foils are the most problematic by sem→phon, and by extension, by orth→sem→phon. Recall that learning is error driven; the development of the orth→sem pathway is in part driven by the degree to which orth→phon→sem yields incorrect patterns. Hence, while broadly the orth→sem pathway does not develop as well for low frequency items as for high, the low frequency member of a pseudohomophone pair (e.g., GAIT versus GATE) is under extreme pressure to repair the erroneous activation from orth→phon→sem. Put plainly, the accuracy of orth→sem is driven by the errors of orth→phon→sem. Low frequency members of a homophone pair which have a low frequency enemy need to overcome low levels of spurious activity from orth→phon→sem (Figure 5.16), and hence have small demands on orth→sem. Low frequency members of a homophone pair which have a high frequency enemy need to overcome much greater spurious activity from orth→phon→sem, and hence place much greater demands on orth→sem. Figure 5.17 shows the accuracy of orth→sem for varying kinds of homophones; note that the accuracy of low frequency foils with high frequency exemplars is higher than that of low frequency foils with low frequency exemplars. For the low frequency foils, the accuracy of orth→sem in Figure 5.17 mirrors the *errors* of those items in Figure 5.16.

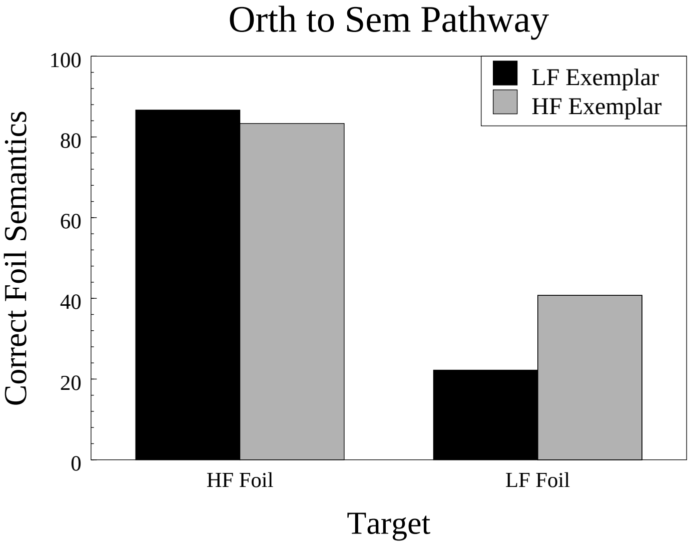


Figure 5.17: Accuracy of orth to sem for homophones. Highly frequent foils are more accurate than low frequency foils. However, this effect is modulated by the frequency of the exemplar: the higher the frequency of the exemplar, the more accurate the orth→sem computation. Low frequency foils with low frequency exemplars are the most error prone.

This explanation cannot apply to pseudohomophones, however, because it relates to the pressures of learning on the low frequency item in a homophone pair. Pseudohomophones, by definition, are not learned; they are novel stimuli. The explanation for these items must derive elsewhere.

As stated earlier, pseudohomophone items that are homophonous with low or high frequency exemplars produced many false positives, because the model had no means to reject these items. The introduction of the spelling check mechanism was necessary to quell these false positives. With the spelling check mechanism in place, effects that match the empirical study obtain. Why? The reason, it turns out, is that in order for a false positive to occur, two things need to happen. The incorrect semantic unit needs to be activated, and the spelling check system needs to make a mistake, that is, it needs to fail to reject the spelling of the item. Why is this spelling check driven by the frequency of the exemplar, when the exemplar is not actually processed by the model in the experiment? It is because the spelling check for pseudohomophone items such as BAIR (compare with the exemplar BEAR) is driven by the exemplar's semantics. The nonword BAIR activates the phonological form /bɛr/, which activates the semantics of [bear]. It is the sem→orth pathway *for this concept* which is used in the spelling check; the concept [bear] activates the spelling BEAR, which doesn't match the spelling of the input item BAIR. For a false positive to occur, this system needs to fail. It will fail for items which are most difficult by sem→orth: exemplars which are low in frequency. The accuracy of the sem→orth computation is in part dependent on the frequency of the item. Hence, pseudohomophones with a low frequency exemplar are more prone to false positives.

This explanation could also in principle apply to the homophone foils, obviating the previous discussion of effects on them. However, the correct results obtain in the model for homophone foils with or without the spelling check apparatus in place. High frequency exemplar / low frequency foil items are rejected not by the spell check mechanism (as suggested must be the case by Van Orden and colleagues), but by failure to activate the incorrect semantic unit in the first place.

The explanation of the phenomena derived from the model places two different causes for the JS91 results. High frequency foils are competently read by orth→sem and hence do not generate false positives. Low frequency foils with high frequency exemplars are also competently read by orth→sem because they need to defeat the incorrect activation of the exemplar by orth→phon→sem. Low frequency foils of low frequency exemplars do not have to overcome such strong input from orth→phon→sem and hence have weaker activity along orth→sem. Pseudohomophones of high frequency exemplars are easily rejected by the spelling check mechanism, which is highly developed for high frequency exemplars. Pseudohomophones of low frequency exemplars are not, because the spelling mechanism is less well developed for the low frequency items.

Chapter 6

DOL, Semantics, and Masking

6.1 The Challenge of Masking Studies

Chapter 5 outlined a theory of the division of labor to semantics and the factors that modulate it. A view emerges in which factors such as the frequency of a word, its regularity and its homophone status all play a part in the division of labor, in complex and subtle ways. In line with the work of JS91, we have evidence against the strong Van Orden hypothesis, that word recognition is invariably, initially phonological.

However, this package of results and insights is challenged by a body of work which claims to support and extend the Van Orden hypothesis, and provide unequivocal evidence for the primacy of phonological coding in word recognition. A study by Lesch and Pollatsek (1993) utilized a *masked priming* paradigm to examine semantic priming effects on naming. The study used two conditions: one in which a prime was presented very briefly to a subject (50ms), then masked with a pattern mask, then a target was presented which the subject was to name. In the crucial trials, the prime was a homophone of an appropriate prime word (e.g., TOWED preceding the target FROG). A second condition was also used in which the prime was presented longer (200ms), then masked, then the target was presented. Their finding was that at the short presentation of the prime, both appropriate primes (TOAD for FROG) and homophones of appropriate primes (TOWED for FROG) primed the target item above and beyond that of spelling controls. At longer presentation times of the prime, only the appropriate item primed the target.

Lukatela and Turvey (1994b) extended these results by including into their design a frequency manipulation of the prime. In one condition, the appropriate prime was more frequent than the inappropriate homophone prime. In the other, the inappropriate homophone prime was more frequent than the appropriate one. Additionally, they included a pseudohomophone condition, and explored whether the pseudohomophone of an appropriate prime (e.g., TODE for FROG) primed the target or not.

The Lukatela and Turvey (1994b) results are summarized in Table 6.1. Following Lesch and Pollatsek, they found that at short presentations, an appropriate prime (TOAD) and a homophone prime (TOWED) facilitate naming of the target item (FROG). In the longer condition, they found that the appropriate word prime (TOAD) but not the inappropriate word (TOWED) prime the target, providing a replication of Lesch and Pollatsek (1993). Additionally, Lukatela and Turvey (1994b) found that the pseudohomophone item (TODE) primes the target at both the short and long conditions. Crucially, they found no interactions of prime type and sublist (the manipulation of whether the appropriate homophone item is more or less frequent than its distractor).

		Priming?	
Prime	Target	Short	Long
TOAD	FROG	y	y
TOWED	FROG	y	n
TODE	FROG	y	y
TORD	FROG	n	n

Table 6.1: Lukatela and Turvey (1994b) results. In the long condition, only a relevant item primes. In the short condition, homophones and pseudohomophones prime the target, but not spelling controls (e.g., TORD).

Lukatela and Turvey argue that their results indicate that word recognition is initially phonological, and that there is a time course to the "cleanup" process which disambiguates homophonous items, such that the short condition has not had time to disambiguate the homophones, while the long one has. The finding that pseudohomophones prime in both conditions lead them to argue for a modification to the standard idea of a spelling check. They claim that the spelling check must proceed by activating the lexical representations of all candidate items regardless of frequency. This activation accesses the stored spelling of these items. They are in turn checked against the spelling of the input. The spelling check must proceed such that a positive match allows that item to suppress its enemies. So, for example, when the word TOWED is read, the lexical entries for [toad] and [towed] are activated. When a match is found between the activated lexical representation of TOWED and the input spelling, it receives a boost of activation which allows it to suppress the competing item TOAD. When the input is a pseudohomophone such as TODE, it too activates both the lexical entries for [toad] and [towed]. However, in this case, neither item receives a positive match to the input, so neither item receives a boost, so neither item is suppressed. In order to accommodate their results, Lukatela and Turvey must argue that [toad] receives no suppression despite its accessed spelling not matching the input of TODE; they explicitly argue that only a positive match of addressed spelling and input spelling can generate the competition required to suppress the inappropriate item. In the absence of this suppression, both lexical items ([towed] and [toad]) are active equally irrespective of their frequency; hence you get the same degree of priming for TODE as you do for TOAD.

This scheme has some problems. The most obvious is: if failing to match addressed spelling to input spelling results in the same lexical activation as an appropriate word that matches, then how do I know that I'm looking at the nonword TODE and not a word? There is no mechanism in their scheme that would tell me that TODE is not, in fact, TOAD. This is fine for accounting for why both TODE and TOAD prime FROG, but ignores 20 years of research in lexical decision, which has shown that people can reject inappropriately spelled items quite rapidly. Subjects tend to be slower at rejecting pseudohomophones in a lexical decision task than ordinary nonwords, but they still can do the task. The Lukatela and Turvey scheme, like other schemes for lexical decision based on positive activation of semantic representations has the problem that it is at pains to explain how we reject pseudohomophones. They would have to appeal to a conscious, homonculous-type decision process to allow for the rejection of pseudohomophones, but it is not clear what mechanism or knowledge that homonculous could draw on in their theory. Furthermore, the assertion that the rejection of pseudohomophones in a lexical decision task is not possible by any kind of automatic process goes against nearly 20 years of research in lexical decision.

This point is important and bears emphasis. To put it simply: when the subject sees TOWED, the visual form activates the concepts [toad] and [towed]. Given sufficient time, a positive match is obtained between the concept [towed] and the input spelling. This positive match drives the activation of [toad] down, which is why TOWED does not prime FROG at long latencies. At shorter times, this spell check cannot progress and so TOWED primes FROG. When the input is TODE, both [toad] and [towed] are activated. Neither activated lexical representation matches the input, so both items are equally active. In particular, [toad] is active to the same degree as when the input is TOAD, hence the same degree of priming for TODE and TOAD. The fact that the addressed spelling of [toad] does *not* match the input TODE does *not*, in their scheme, lead to suppression; only a positive match does.

The authors feel this bizarre scheme is necessary to explain their pattern of results. They invoke this scheme to explain differential time courses of priming for homophonous words (TOWED) and pseudohomophones (TODE). But recall from the previous section the discovery that the rejection of homophonous words can proceed via different mechanisms than that of pseudohomophones. In that model, pseudohomophones are rejected by activating a semantic representation and then finding a mis-match between the accessed spelling and the input spelling. Homophones are generally rejected prior to this being necessary; they do not activate the incorrect meaning in the first place, because the orth→sem route is in place. Lukatela and Turvey are theoretically committed to the idea that orth→sem is not available to disambiguate items, and hence need to apply the same mechanism to explain the disambiguation of homophones as pseudohomophones.

6.2 Insights from the Model

In the model presented in the previous chapter, pseudohomophones often activate the incorrect semantic representations, and hence a spell check mechanism must be used to reject the pseudohomophone. Homophones are disambiguated directly before this is necessary. Hence, the fact that TODE primes FROG at 250ms can be accounted for by arguing that the spelling check has not had time to be applied. Lukatela and Turvey must argue that the spelling check *has* had time to be applied at 250ms, because that is their explanation for how TOWED is suppressed. In the model presented here, TOWED would not prime because orth→sem would suppress the inappropriate semantic representation.

At this point, this model accounts for the Lukatela and Turvey results at 250ms quite more elegantly than their model, and further is capable of rejecting pseudohomophones in a lexical decision task. It makes the further prediction that at much longer SOAs (say, 800ms), the spelling check would have time to work, and so at 800ms, TODE would not prime FROG. Their model makes no such prediction. This could be empirically verified, but it intuitively seems very unlikely that once the prime was actually consciously recognized as a nonword it would prime the target item. Lukatela and Turvey could argue that a consciously recognized item wouldn't prime anyway, but that misses the point that the framework they are proposing *can't* recognize that TODE is a pseudohomophone.

However, we have failed so far to explain their pattern of results at short SOAs; that is, why exactly does TOWED, presented briefly, prime FROG?

The answer can be found by carefully examining the assumptions of the masked priming paradigm. Recall that the idea here is that a word is presented on the screen for a brief amount of time. Then, the word is masked, either by a random pattern, or by a new word (such as the target). The implicit assumption behind these studies is that the effect of the mask is to halt processing of the input item; this allows the experimental manipulation to interpret the state of lexical processing at that and subsequent moments as a snapshot frozen in time of the processing of that item. As such, masking is considered a powerful tool for discovering the time course of lexical processing.

But what if this implicit assumption is wrong? It is not, of course, necessarily true that processing must stop when input is removed. While some computational metaphors support such an idea (it would be hard to imagine a hash table continuing its processing in the absence of input), there are a host of others which do not. The interactive activation (IA) model of McClelland and Rumelhart (1981), introduced almost 20 years ago, is one such example. The current model under consideration is another.

The model presented in this work contains attractors in phonology and semantics. These attractors are capable of performing pattern completion, repairing or altering partial or degraded input (Plaut & Shallice, 1993; Harm & Seidenberg, 1998). The masking paradigm assumes that the effect of the mask halts processing on both (putative) pathways; evidence of phonological orth→phon but not orth→sem activation is interpreted by Lesch and Pollatsek and Lukatela and Turvey as evidence for the primacy of orth→phon. But what if the effect of the mask turned out to have a more debilitating effect on orth→sem than on orth→phon? The result would be quite devastating for the interpretation of the masked priming results; the "finding" of the primacy of orth→phon would in fact be an artifact of the paradigm used to investigate such effects.

6.3 The Model's Behavior When Masked

The JS91 experiment from the previous chapter were repeated, under conditions analogous to masking. Simulating the effect of masking is difficult in the model; orthographic units correspond to discrete letters. Noisy activation presented to the orthographic units resulted in noisy activation via orth→sem, which is not exactly the desired behavior. The idea behind masking studies is that information is no longer available to the reading system; the real function of pattern masks is to obliterate any iconic visual image that may remain in the visual system. To directly emulate the removal of useful visual information from the reading model, then, masking was modeled by shutting off input to the semantic and phonological units, respectively, from orth→sem and orth→phon.[1]

The JS91 simulation was conducted with input masked after 2.0 units of time (out of 7), and allowed to continue cycling in the absence of orthographic constraint. The state of the network was sampled after 4.0 units of time, as before. Table 6.2 shows the activities of the inappropriate unit in the masked and unmasked conditions.

The results are illuminating. Overall, a much greater degree of false positive responses was detected. Whereas in the normal, intact model, the only conditions which yielded even an occasional spurious activation of the incorrect unit were the low frequency foils of low frequency exemplars,

[1] A more accurate simulation of masking would require input of the form of visual features, similar to the IA model, which collected and formed letter representations over time in an attractor network analogous to the semantic and phonological attractors. Such a simulation of low level letter recognition is beyond the scope of this work but remains an interesting avenue of future work.

	Normal			Masked		
	Foil			Foil		
Exemplar	HF	LF	PSH	HF	LF	PSH
HF	0.0005	0.0086	-0.0013	0.14	0.097	0.21
LF	0.0008	0.0211	0.0193	0.097	0.23	0.23

Table 6.2: JS91 simulation results, normal and masked. The normal condition matches the empirical findings; only low frequency foils and pseudohomophones of low frequency exemplars show spurious activation. When masking is introduced, dramatically more spurious activation occurs.

and pseudohomophones of low frequency exemplars (as was found in the JS91 study), the masked model sees many false positives in all conditions.

The reason for the increase in false positives for the pseudohomophone conditions is straightforward. When the orthographic input is removed, a spelling check, and hence a rejection based on incorrect spelling, is not possible.

The reason for the increase in false positives for the foils follows from the processing dynamics of the network. When orthographic input is removed, the semantic system has only its own self activation, and activity from phon→sem. Similarly, phonology only has its own activity and activity from sem→phon. Intuitively, it seems possible that the phonological attractor ought to be able to retain a pattern over time more easily than the semantic attractor. There are far fewer phonological features than semantic ones, and the degree of redundancy is much greater. The semantic space is more sparse than the phonological space, and the average unit is active far less often.

A test was devised to measure the phonological and semantic attractors' ability to retain patterns over time. The normal model was lesioned such that the sem→phon and phon→sem pathways were cut. This allowed examination of the semantic and phonological attractors in the absence of collateral support from each other. Each word in the training corpus was presented to this lesioned model, and over the timecourse of processing, the percentage of correctly active features was measured. These percentages were averaged for each timestep and the results are presented in Figure 6.1 (left). The experiment was then repeated, using the masking paradigm described above. At 2.0 units of time, input from orthography to semantics and phonology was totally shut off, and the networks allowed to continue cycling. These results are shown in Figure 6.1 (right).

Clearly, the effect of the mask is much more debilitating on the semantic attractor than the phonological attractor. The phonological attractor is, on average, much better able to retain its correct activity in the absence of orthographic input.

If, when a mask is applied, phonological activity perseverates but semantic activity (driven from orth→sem) decays rapidly, then one would naturally expect to see increased phonological effects under conditions of masking. These effects were seen in the masked version of the JS91 experiment; all conditions showed an increase in phonologically based false positives.

This predicts that the intact model ought to show different patterns of behavior for homophone items depending on the introduction of masking. One homophone pair, BALM and BOMB was examined. The probability of presentation of BOMB during training was 0.58; BALM was presented with a probability of 0.21. Hence BOMB was presented almost three times as often. Correspondingly, when the hearing network was presented with the phonological form /bam/ (corresponding to both words), it unambiguously activated semantics for an exploding device, not a medicine.

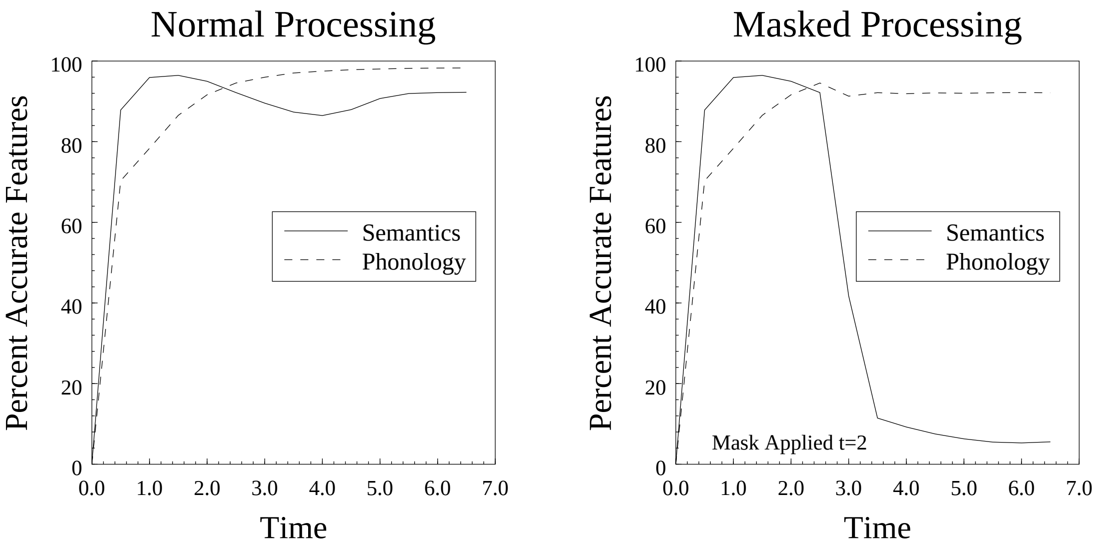


Figure 6.1: Normal and masked timecourse of feature activation, for semantics and phonology. Semantic units ramp up as rapidly as phonology, but are more disrupted by the mask.

The word BOMB contains the semantic feature <*weapon*>, while BALM contains the semantic feature <*medicine*>. By measuring the activity of these units when the printed form of these words was presented to the normal and masked network, one can see the effect over time of the mask on the different homophones and their semantic interpretation.

Figure 6.2 shows the results. The top row shows the behavior of the normal and masked network when BALM is presented normally, and masked. The bottom row shows the results for BOMB. For the normal network, the results are straightforward: when presented with BALM, the <*medicine*> feature is activated, and the <*weapon*> feature is suppressed. This is due to input from orth→sem; if the orth→sem pathway is lesioned the network generates a positive for <*weapon*> (0.60) and virtually no activity for <*medicine*> (0.09). When BOMB is used as input to the normal network, the reverse pattern is seen: <*weapon*> is activated and <*medicine*> is suppressed.

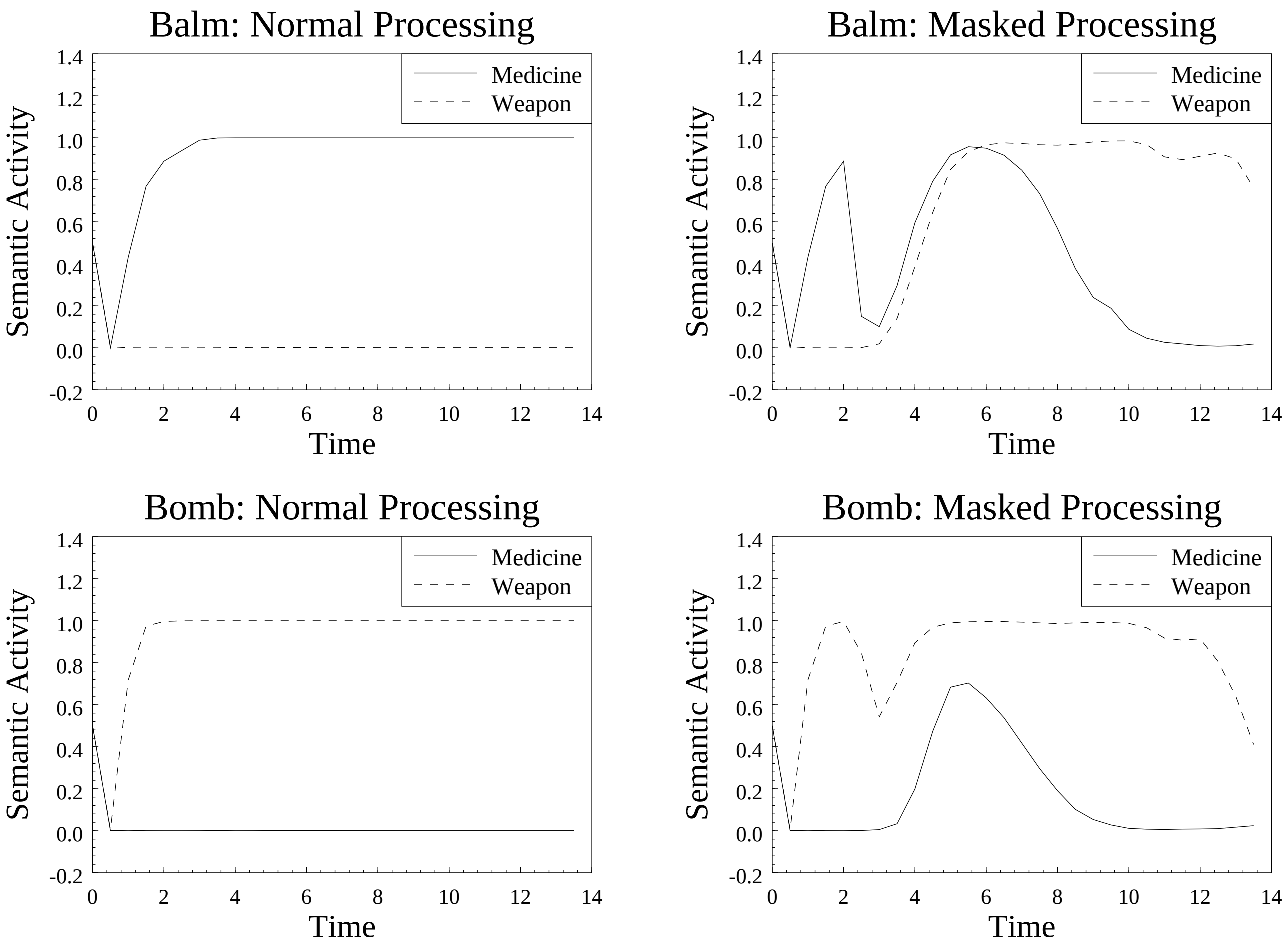
Under conditions of masking (as before, at t=2), the pattern becomes more complex. Here we see that BALM, when masked, shows a time when both <*weapon*> and <*medicine*> are activated. Over time, the inappropriate feature <*weapon*> becomes the dominant feature, showing that under masking conditions BALM would activate semantic features similar to BOMB.

Of greater interest is the behavior of the network when BOMB was presented. Recall that BOMB was the dominant homophone; when the phonological units were clamped with /bam/, the semantics were unambiguously the weapon interpretation. When the spelling of either BALM or BOMB was presented to the network with no orth→sem pathway, the semantic units reflect the weapon interpretation. However, when BOMB was masked, there was a brief period of activation of the <*medicine*> feature, before it was suppressed. The effect of the mask was to throw the semantic attractor into a period of confusion in which it straddles the fence between the two interpretations of /bam/.

This is an important observation. When Lukatela and Turvey (1994b) found no effect of sublist (that is, if the distractor homophone was the more or less frequent of the pair), they conclude that this implicates a model of word recognition in which the meanings of both homophones become active *to an equal degree*, and remain so until a spell-check process can suppress the incorrect item. Aside from seeming unintiutive (it would be quite surprising to find that people activate the concept [ewes] to the same extent as [use] when they encounter the word USE), the result now seems to be an artifact of the masking paradigm. This point deserves emphasis: at no point in the normal operation of the model does BOMB prime the concept [balm]. The normal system is subject to frequency differences, as one would intuitively expect. The introduction of masking, however, introduces enough noise into the semantic attractor that allows spurious activation of both meanings of the phonological form /bam/. This is a property of complex dynamical systems: small perturbations can lead to wildly divergent intermediate states, before the system returns to equilibrium. The state of the system during this settling process is difficult to interpret; but that's the point. It is the mask, and not the normal dynamics of the intact system, that throws the semantic attractor into this state.

6.4 The Frequency Manipulation

Lukatela and Turvey (1994b) manipulated the frequency of the primes (TOWED versus TOAD) and found no effect of sublist; that is, it did not matter if the appropriate prime was higher or lower in frequency than the homophone distractor. As noted above, they conclude that the lexical access


Balm: Normal Processing

Balm: Masked Processing

Bomb: Normal Processing

Bomb: Masked Processing

Figure 6.2: Normal and masked timecourse of activity for BALM and BOMB.

mechanism must not be frequency sensitive. The simulations from the previous section cast doubt on this interpretation, showing how the masked priming paradigm can produce effects that overrun the normal frequency biases.

But there is further reason to reject the conclusion of Lukatela and Turvey. Examination of their items reveals that their frequency manipulation was in fact quite weak. For the items listed in Appendix A of Lukatela and Turvey (1994b) the mean pairwise difference between the frequency of the more frequent item and its less frequent distractor is 150, which seems like a strong manipulation. However, this is carried by a small number of items with very large differences in frequency; the median difference is only 24. Further, there are 18 pairs (out of 60) for which the difference in frequencies is $<= 10$. The items from Appendix B have the same problem: while the mean difference is 130, the median is only 29, and 23 pairs (out of 84) have a difference $<= 10$. The failure to find an effect of sublist may well be due to the weakness of the frequency manipulation.

Lesch and Pollatsek (1993) also failed to find an effect of relative frequency of the prime, but their items also had a very weak manipulation of frequency. While the mean frequency difference for their items was 111, the median was only 29, and 8 out of 32 of their items had a frequency difference $<= 10$.

The Jared and Seidenberg (1991) items, in contrast, have a mean difference of 82, and a median difference of 50. No items had a difference $<= 10$. The differences were not carried by a few large items, but were robust across all stimuli.

The studies by Lesch and Pollatsek and Lukatela and Turvey (1994b) had very weak manipulations of frequency and failed to get an effect of frequency. The Jared and Seidenberg study had a robust frequency manipulation, and obtained the effect. Hence the strong conclusion of Lukatela and Turvey (1994b) that the relative frequency of does not influence homophone pairs does not influence lexical access is not supported by the evidence provided.

6.5 Summary

Lukatela and Turvey's description of the reading system is couched in the language of classical cognitive science: access to a lexicon of stored items, the atomic tokening of different lexical entries, etc. Lacking metaphors such as partial semantic activation, chaotic dynamics, and attractor networks, they are forced to construct a theory of lexical access with an extraordinary amount of otherwise totally unmotivated theoretical baggage: false matches in the spelling check have no effect, the word recognition system has no early access to a word's spelling, the phon→sem computation is insensitive to frequency. The current simulations show that none of these findings are necessitated by the data. The account presented here works by appeal to computational principles in a straightforward system that attempts to compute the semantics of a word as rapidly as possible and from as many sources of information as possible. Hitherto unexamined (and, in fact, unconsidered) assumptions of the masking paradigm are shown to be demonstrably false within such a framework. The patterns of behavior noted in the literature that have been supposed to implicate an initially exclusive role of phonology in the computation of a word's meaning actually derive as an artifact of the masking paradigm used in these studies.

Chapter 7

DOL To Phonology

In this section, the question of the division of labor in the computation of a word's phonology is considered. While not as contentious a policy issue as the DOL in the computation of meaning, the question of the DOL to phonology is of significant theoretical importance. Much of the data which has driven the development of models of word recognition derive from the performance of normal and brain damaged people's word naming times. As noted earlier, the classical dual route model is fundamentally a dual route model to the *pronunciation* of a word; access to meaning is of secondary importance in the framework. Similarly, while SM89 outlined a framework for the computation of both the meaning and pronunciation of a word, the implemented model computed the pronunciation, not the meanings, as has the bulk of work which has followed in this tradition (e.g. Plaut et al., 1996; Harm & Seidenberg, 1998).

The division of labor to phonology is different from that to semantics in that the factors more sharply favor the orth→phon route over the orth→sem→phon route. Whereas the computation of meaning had the speed requirement encouraging learning along orth→sem, the correlated nature of English orthography made orth→phon easier. With the computation of a word's phonology, the orth→phon pathway is both the faster and (broadly) easier one.

One might be tempted to think, then, that the DOL to phonology would be quite uninteresting; words are read according to orth→phon, which is both the direct, one-step route, and the easy, correlated route. However, there is evidence from various sources that there is more to the story.

7.1 Evidence from Acquired Dyslexias

7.1.1 Deep Dyslexia

The syndrome of deep dyslexia (Coltheart et al., 1980) is characterized by a complex pattern of behavior in brain damaged populations.

a. Poor nonword naming. Patients are very impaired at pronouncing novel nonwords.

b. "Visual" errors. Patients make visual errors, that is, pronouncing SYMPATHY as SYMPHONY.

c. "Semantic" errors. Patients additionally make semantic errors, e.g., reading the word SYMPHONY as ORCHESTRA. Interestingly, on occasion, patients combine semantic and visual errors, yielding patterns such as reading SYMPATHY as ORCHESTRA (presumably by visual misanalysis of SYMPATHY to SYMPHONY, then a semantic shift from SYMPHONY to ORCHESTRA.)

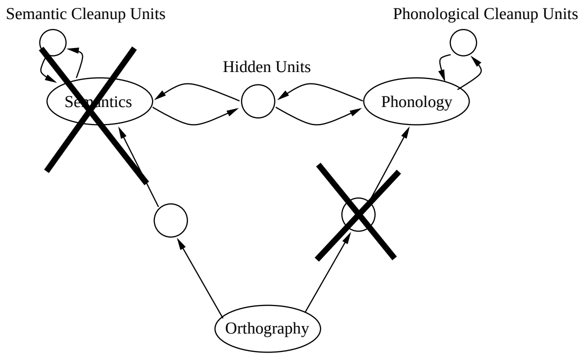


Figure 7.1: Putative sites of damage in deep dyslexia.

d. Word class effects. In addition to being impaired in the reading of nonwords, deep dyslexic patients have greater difficulty with closed class words than open class words. Within the open class words, those with abstract meanings are named more poorly than concrete ones.

Hinton and Shallice (1991), Plaut and Shallice (1993) and Plaut (1991) have explored an explanation of these patterns of results by appeal to attractor networks. They implemented a network with the orth→sem→phon pathway, and left orth→phon unimplemented. The exclusion of orth→phon was meant to reflect the hypothesized devastation of this pathway by brain damage. They then used varying degrees of damage to the semantic attractor, an intermediate stage in the orth→sem→phon computation. By varying the degree of damage, or whether the damage was just before semantics or just after it in the orth→sem→phon path, there were able to account for the various patterns of damage within this syndrome. The extreme damage to orth→phon requires all reading to be done semantically. This semantic reading explains the word class effects, and the semantic shifts. By introducing damage in the orth→sem computation, this process is made sufficiently noisy that visual errors result. By damaging the semantic attractor itself, semantic errors result.

Crucial to this explanation is the notion that orth→sem→phon can support reading for a reasonably broad (but by no means exhaustive) set of words. Additionally, it is assumed that the semantic system is organized in such a way that closed class items are more semantically vulnerable than open class items, and that effects of concreteness have an effect on the vulnerability of open classed items (see Chapter 3 for more detail).

7.1.2 Surface Dyslexia

Surface dyslexia is characterized by a patients' pronounced failure to read exception words. As noted earlier, the dual-route model explains this pattern by positing that exception words cannot be read by rule, and hence require a lexical access mechanism. These patients are said to have a problem with their lexical access, with preserved rule application (Patterson et al., 1985).

However, there appear to be two distinct forms of surface dyslexia that have been reported, both fitting the broad characteristic of impaired exception word reading, but differing in other regards. The first, termed "dysfluent" surface dyslexia, exhibits impaired nonword reading as well as word reading. Additionally, their naming latencies are quite long. Patient JC, studied by Marshall and Newcombe (1973) is a characteristic patient of this sort.

In contrast, a “fluent” form of surface dyslexia has also been identified. Patient MP (Bub et al., 1985), unlike JC, exhibited normal naming latencies, and normal nonword reading. MP, like most all other “fluent” surface dyslexics, had significant semantic impairment, while JC’s semantics were normal.

An account of both varieties of surface dyslexia within the computational framework of the SM89 model was provided by Plaut et al. (1996). The “dysfluent” pattern can be simulated by providing a partial lesion to the orth→phon pathway. In the SM89 framework, both regular and exception words can be read by this pathway. However, exception words are more computationally demanding, and hence are impacted more by impairments (see also Harm & Seidenberg, 1998, for discussion).

The “fluent” form, however, cannot be simulated in this way. Noting that virtually all patients of this form have a semantic impairment, and noting that the SM89 model as originally implemented failed to read many low frequency exception words by orth→phon, Plaut and colleagues hypothesized that such exceptions require support from semantics; that is, they are read via orth→sem→phon. Thus, the semantic impairments that the fluent patients exhibit is explained in this way: the patients were reading low frequency exceptions by orth→sem→phon before their brain injury, the brain injury impaired their semantic representations, which in turn impaired the reading of the items which need the semantic pathway operational in order to be read: low frequency exceptions.

7.1.3 Phonological Dyslexia

The counterpart to surface dyslexia is “phonological” dyslexia; a syndrome defined by severely impaired nonword reading, with normal or relatively normal word reading (Beauvois & Derouesné, 1979; Derouesné & Beauvois, 1979). Here, the dual-route model explains the pattern of behavior in a straightforward way: whereas surface dyslexia involves normal rule application and impaired lexical lookup, phonological dyslexia is the reverse. The rules are impaired, and the lexical lookup is intact.

The SM89 framework can also account for phonological dyslexia in a straightforward way. Recall that the account of deep dyslexia involved two lesions: obliterated functioning of orth→phon as well as more minor impairments along the orth→sem→phon pathway. If phonological dyslexia is viewed as a form of deep dyslexia, without the semantic/visual errors (Friedman, 1996), then phonological dyslexia can be simulated by obliterating orth→phon while leaving orth→sem→phon intact.

This account predicts the same word-class effects seen in deep dyslexia. Such effects have been reported in cases of phonological dyslexia (Friedman, 1996). Further, it predicts exaggerated semantic effects on word reading. Patient MJ, reported by Howard and Best (1996), is grossly impaired at reading nonwords, failing to pronounce more than 45% of simple three letter nonwords such as BEM. Her word reading, however, is almost totally normal. In investigating MJ’s reading, Howard and Best found far lower effects of word regularity than is found in the normal population, and much exaggerated effects of word imageability, a semantic factor. This provides additional evidence that MJ is reading words aloud primarily by orth→sem→phon.

7.1.4 A Third Route to Pronunciation?

Coltheart (1996) presented an argument for a third route to pronunciation, which can be found in the dual-route model but not the “triangle” model of SM89. Evidence for such a route is the existence of cases such as patient WB (Funnell, 1983), who exhibited the classic symptoms of phonological dyslexia, but also exhibited severe semantic impairments. If the triangle model accounts for phonological dyslexia by positing reliance on the orth→sem→phon computation, then how can WB read words with high accuracy yet have a high degree of semantic impairments? Cases like WB, Coltheart argues, demand a route to pronunciation which is lexical (that is, not rule based), but not semantic. Coltheart claims that if the triangle model accounts for WB by positing that exceptions are read by orth→phon, then it loses its ability to explain the fluent surface dyslexic patients. But if they are not read in this way, then how can one account for WB? Some insights into how this can happen will be taken up in this chapter.

A similar argument is made for cases of patients who exhibit severe semantic impairments but do *not* exhibit surface dyslexia, such as patient DRN, studied by Cipolotti and Warrington (1995). If we are to believe that low frequency exceptions rely on orth→sem→phon, as the triangle model account states, then how could we find someone exhibiting severe semantic impairments but not surface dyslexia? Coltheart argues that a lexical, nonsemantic pathway devoted to exception reading is required to account for these patients.

This point is taken up by Plaut (1997), who argues that there must be individual variation in the DOL in patients, before the brain injury. By manipulating various model parameters, he was able to create models which relied very heavily on orth→phon for exception reading, and others which relied much more heavily on orth→sem→phonȦ patient whose reading system relies heavily on orth→sem→phon and then experiences semantic damage will be a fluent surface dyslexic. A patient who does not rely on this path, and reads virtually all exception words by orth→phon can withstand semantic impairment without becoming dyslexic.

7.2 Evidence from Normal Populations

As discussed in the introduction, Strain et al. (1995) found that normal subjects exhibit an effect of imageability, a semantic variable, in their naming latencies for words, but only for words which are low frequency exceptions. The standard interpretation of this effect is that the low frequency exceptions are most poorly read by the orth→phon route,

7.2.1 Evidence Bearing on DOL to Phonology: Summary

Within the framework of the triangle model, all three forms of acquired dyslexia demand some form of activation from orth→sem→phon. The items which are predicted to be the most likely to rely more heavily on orth→sem→phon are the exceptions, because they are most likely to be poorly read by orth→phon. The Strain et al. (1995) finding suggests that in the normal system, low frequency exceptions rely on orth→sem→phon. The fact that patients with putatively impaired orth→phon computation can still read many words suggests that more words can be read by orth→sem→phon. Why, then, are such effects not seen for all words in normal subjects? Perhaps

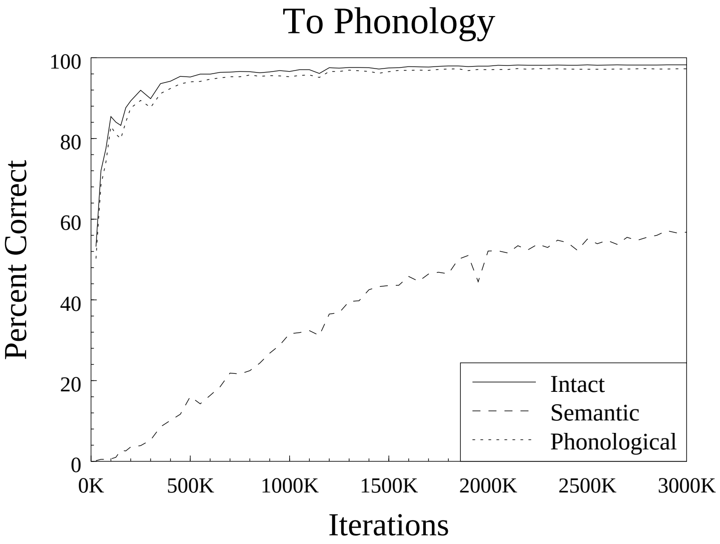


Figure 7.2: Accuracy of activation of phonology, by intact model, model with only semantic and phonological paths.

because orth→phon is sufficiently dominant in the normal system that effects of orth→sem→phon are washed away by orth→phon.

With such intuitions and evidence in hand, we can now explore how the model actually does compute the phonological form of a word.

7.3 Time Course of Development

Figure 7.2 shows the time course of development of the intact model, and the intact model with only input from orth→phon, and with only input from orth→sem→phon. As in the case of the DOL to semantics, the orth→phon pathway ramps up quite rapidly. The orth→sem→phon pathway takes much longer to develop. Interestingly, the performance of the normal model does not deviate far from that of the orth→phon pathway; the direct orth→phon pathway is driving the competence of the model throughout development. Nonetheless, approximately half of the word types are eventually read correctly by the orth→sem→phon path. Figure 7.3 shows the breakdown of items read by the orth→phon route, the orth→sem→phon route, or by either route, or requiring both routes. Unlike the DOL to semantics, here we see that initially a vast majority of items are read only by orth→phon. Over time, the set of items which can be read by either route increases. At asymptote, over half of all words can be read by either the initially dominant orth→phon route, or the orth→sem→phon route. Factors influencing the development of the orth→sem→phon route are considered next.

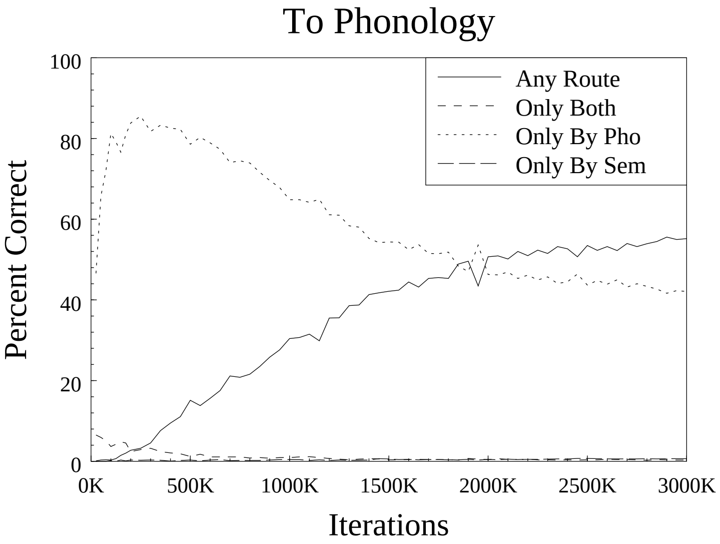


Figure 7.3: Accuracy of activation of phonology, by intact model, showing proportion of items read by orth→phon, orth→sem→phon, either route, or requiring both routes.

7.4 Frequency Effects

The DOL to phonology was next broken down by frequency. Figure 7.4 shows the effect of frequency on the orth→sem→phon pathway. Throughout development, high frequency items enjoy greater support from orth→sem→phon than low frequency items. This reflects the advantage of high frequency items for the orth→sem computation, explored in Chapter 5. As orth→sem is more accurate, orth→sem→phon naturally increases in accuracy as well.

The effect of frequency on the orth→phon route (Figure 7.5) is not seen; there is a ceiling effect for low and high frequency items.

Asymptotically, There is a frequency effect on the accuracy of orth→sem→phon but not orth→phon (Figure 7.6). The relative ease by which the model learns the orth→phon route washes away any effect of frequency. The much more difficult orth→sem→phon route shows effects of frequency.

7.5 Regularity Effects

The influence of regularity on the DOL to phonology is considered next. Figure 7.7 shows the breakdown of items to phonology, by the orth→phon route, as a function of regularity. Regular items are read more accurately than exceptions by the orth→phon route. A reverse is seen in Figure 7.8, where the exceptions are read more with greater accuracy than the regulars by the semantic route to phonology.

Asymptotically, we see that more words can be read accurately by the orth→phon route than orth→sem→phon, but this is modulated by regularity (Figure 7.9). Exception items are favored slightly by the orth→sem→phon route, while regular items are favored slightly by orth→phon.

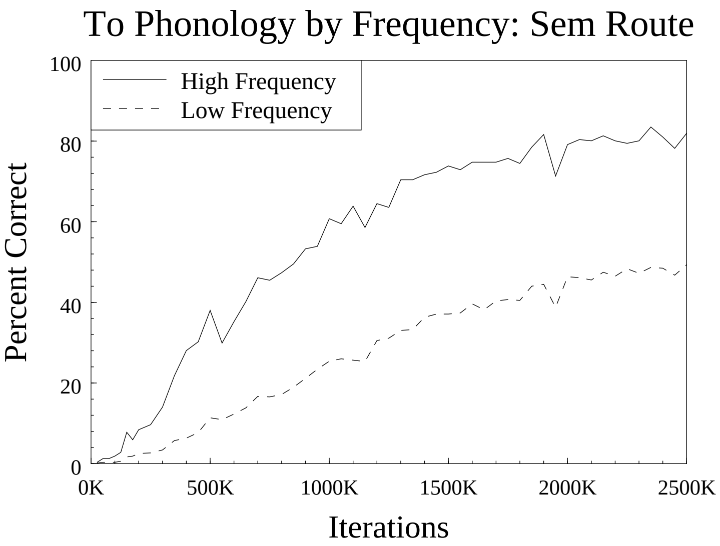


Figure 7.4: Frequency effects on orth→sem→phon over time. High frequency items show a large benefit over low frequency items.

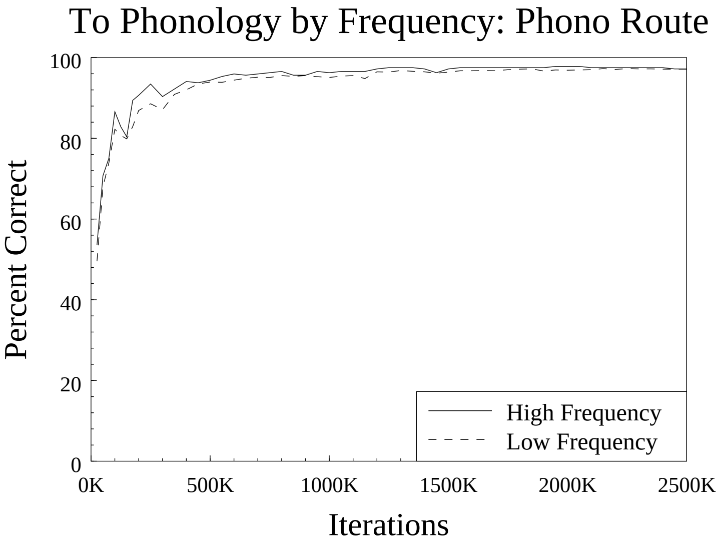


Figure 7.5: Frequency effects on orth→phon over time. A ceiling effect is clearly seen, where the frequency of a word does not impact its accuracy.

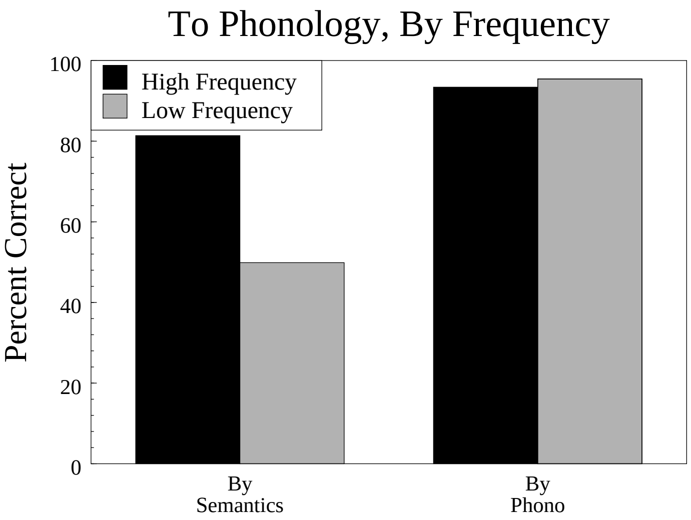


Figure 7.6: Frequency effects on computation of phonology. Frequency affects the orth→sem→phon pathway, which is much more difficult than the orth→phon pathway.

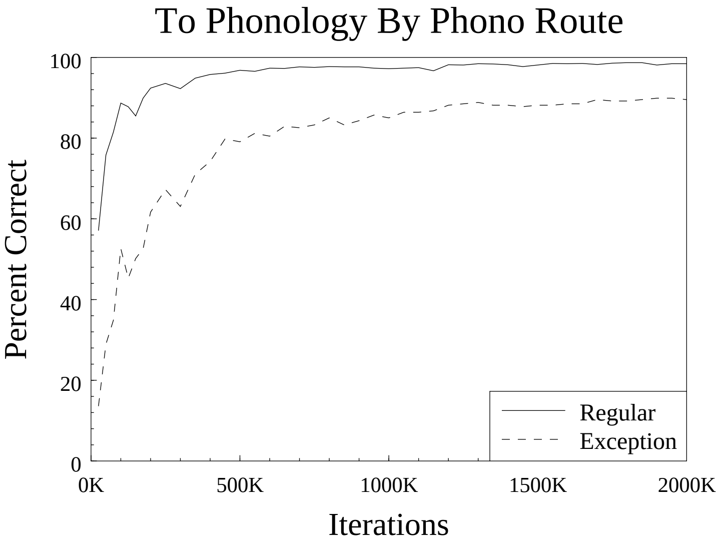


Figure 7.7: Regularity effects in the computation of phonology by orth→phon.

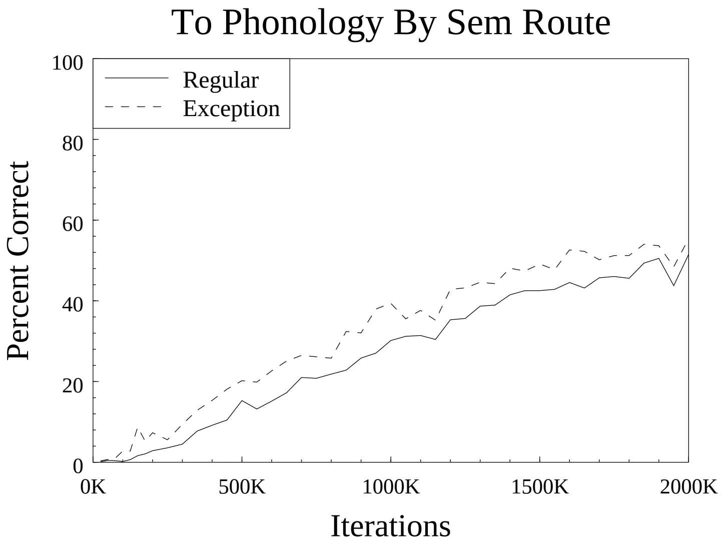


Figure 7.8: Regularity effects in the computation of phonology by orth→sem→phon.

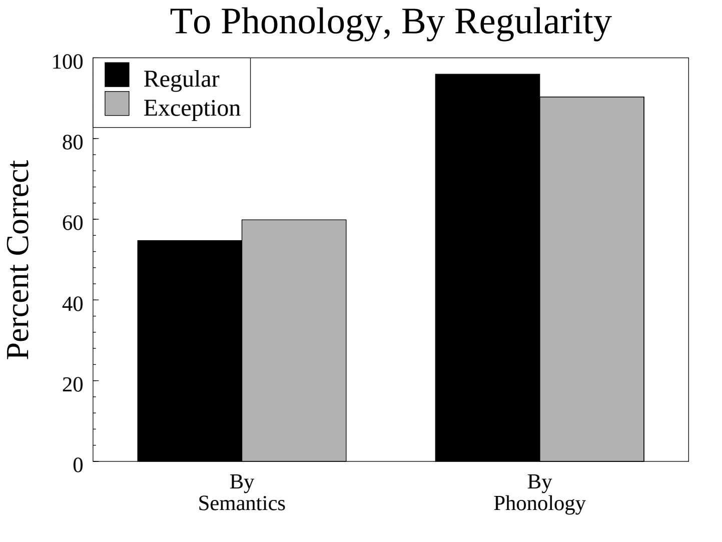


Figure 7.9: Overall regularity effects in the computation of phonology.

7.6 Interaction of Frequency and Regularity

As noted before, investigations of frequency and regularity independent of each other are of interest, but it is important to consider their conjoined effects, because in language they are in fact strongly correlated with each other.

Figure 7.10 shows the breakdown of frequency and regularity effects in the orth→phon route. For both the high and low frequency items, there is a regularity effect where regulars are read more accurately than exceptions. However, the expected disadvantage of low frequency exceptions over all other groups was not seen in the model's orth→phon accuracy.

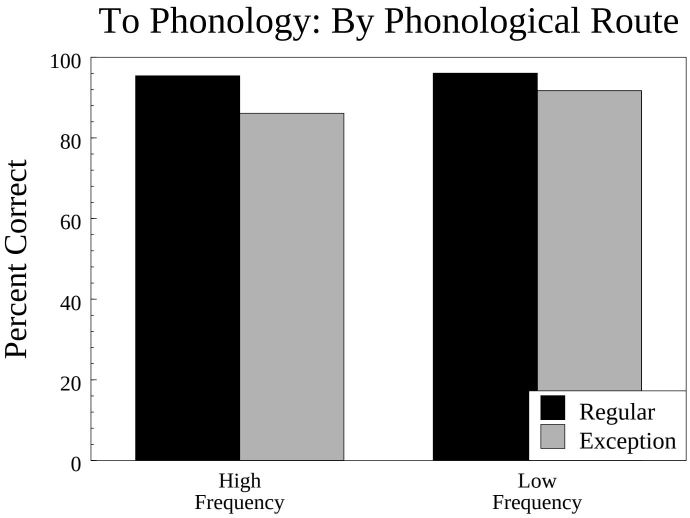


Figure 7.10: Frequency and regularity in the DOL to phonology, by the orth→phon route.

7.6.1 Latencies, and the Frequency by Regularity Interaction

The disadvantage of low frequency exceptions is seen in the intact model's speed of naming, however. Naming latencies for the training set items were computed for the normal model, by measuring the time until the phonological attractor settled. The attractor was considered as having settled when none of its features had changed by more than 0.01 in the past 2 units of time. At that time, the accuracy was measured. Items which were settled but incorrect were discarded, as is the normal practice for naming studies.

The use of settling time in attractor networks as a measure of latency has been utilized in other work (e.g., Seidenberg & Plaut, In Press). The implicit theory is that a response can begin only when the attractor has stabilized on a representation. The extreme form of this hypothesis is probably not true; settling times for the initial phoneme, rather than an entire word can provided better matches to human reaction times (Seidenberg & Plaut, In Press). Still, the rationale for collecting human reaction times and model settling times is that both provide a standard measure of the processing difficulty of the item in question.

Table 7.11 shows the mean naming times for the different conditions. The standard frequency by regularity interaction was only partially observed in the model. The effect is twofold: the first is that frequency only affects the naming speed of exception items, not regulars. This effect was obtained. However, another aspect of the standard effect is that both low and high frequency regulars, and high frequency exceptions all have very similar naming times; the low frequency exceptions are the different items. This effect was not seen in the model, which shows exaggerated effects of exceptionhood.

I will return to the importance of the extended naming times for the low frequency exceptions in the discussion of the imageability effect.

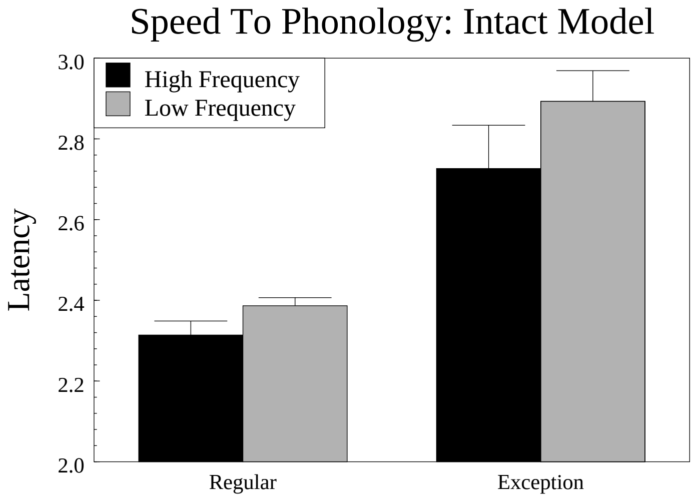


Figure 7.11: Naming latencies to phonology, by frequency and regularity. Most slowly named items are the low frequency exceptions.

Regarding the semantic route to naming, Figure 7.12 shows the breakdown of frequency and regularity by orth→sem→phon. Overall, the high frequency items show an advantage over the low frequency items, as expected. However, this advantage is modulated by regularity. For the low frequency items, the exceptions show an advantage over the regulars. For high frequency items, the pattern is reversed: regulars are more accurately read than exceptions.

7.7 Word Class Effects

Data from acquired dyslexias has suggested that when a patient is forced to rely on the orth→sem→phon path for reading words, not all words are read equally well. In particular, closed class words have been found to be at risk relative to open classed words.

Because closed class words are generally so high in frequency relative to the open class words (mean probability of presentation of 0.8, versus 0.36), the open and closed class words were broken out by frequency, so that frequency wouldn't bias the results.

Figure 7.13 shows the accuracy of the orth→sem→phon route for low and high frequency open and closed class words. For high frequency items, the difference between open and closed

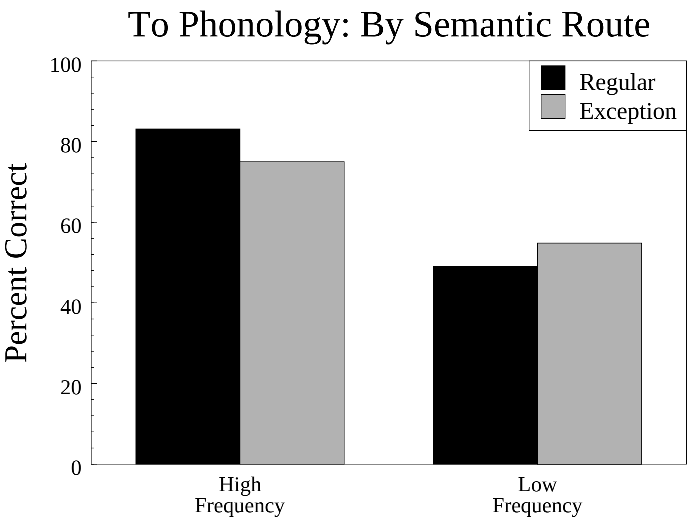


Figure 7.12: Frequency and regularity in the DOL to phonology, by the orth→sem→phon route.

class words is very small, with the open classed words enjoying a small advantage. A much more dramatic difference is seen for low frequency items, where the closed class words are at a severe disadvantage. Empirical studies tend to avoid using the very high frequency items, such as THE and YOU, so this result is not inconsistent with the behavioral data.

7.8 The Imageability Effect

The Strain et al. (1995) study is one of the few studies to find a significant semantic effect on the naming of isolated words. As noted earlier, the impetus for this study was the observation that the SM89 model had the most difficulty with low frequency exception words. The implemented SM89 model implemented only the orth→phon computation, and as such it predicted that these words, to the extent that a normal subject can name them, are more likely to be named by semantic mediation, that is, by orth→sem→phon. As such, they are the most likely to show effects of a semantic variable, imageability. The idea is that for these items, a normal subject is operating more like a deep dyslexic; computing the pronunciation of words semantically. Figure 7.14 shows the results of the Strain et al. (1995) study (Experiment 1). Imageability, while having a very small impact on all items, only had a reliable effect on the low frequency exceptions.

Coding items by frequency and regularity, I then split the remaining items according to imageability, from the MRC imageability norms. Items over the median value of imageability (450) were coded as highly imageable; those below 450 were coded as low. The naming latencies computed in Section 7.6 were analyzed with respect to the three variables of imageability, frequency and regularity.

Figure 7.15 shows the results. Statistically, the data are too noisy to yield reliable effects. Qualitatively, however, the main finding of the Strain et al. study was replicated: only for low frequency exceptions did the imageability variable have a large effect.

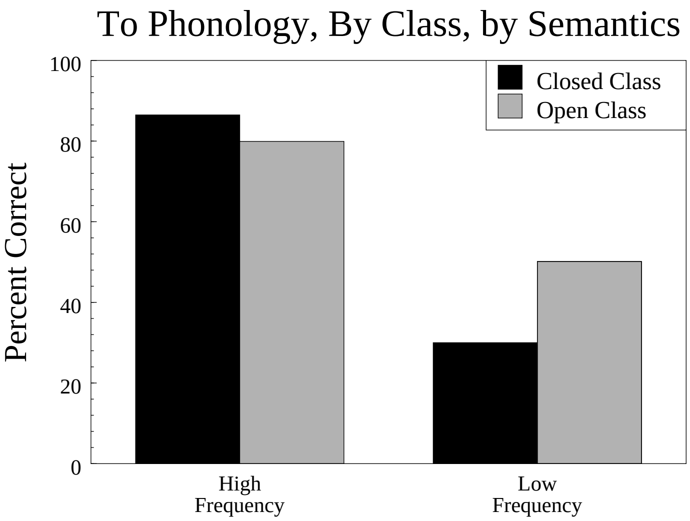


Figure 7.13: Accuracy of open and closed class, high and low frequency words by the orth→sem→phon pathway. Low frequency closed class words are most poorly read.

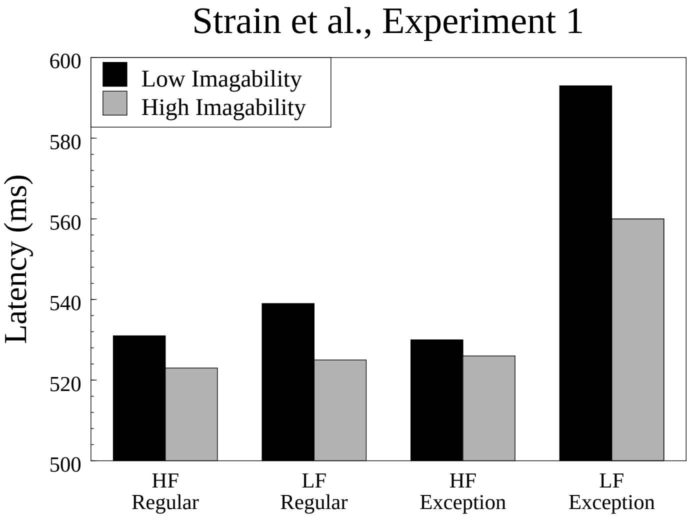


Figure 7.14: Data from Strain et al. (1995), experiment 1. The low frequency exceptions are the only items which show a statistically reliable effect of imageability.

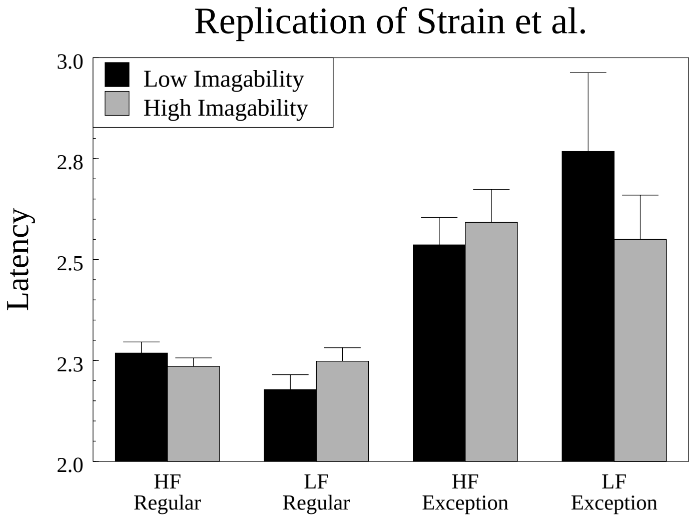


Figure 7.15: Qualitative replication of Strain et al. (1995), experiment 1. While no results are statistically reliable, the low frequency exceptions are the only items which approach a statistically significant effect of imageability.

Examination of the behavior of the model on these low frequency exceptions yields an interesting result. The standard explanation of the imageability effect is that the low frequency items cannot be read by orth→phon and hence rely on semantic mediation. However, all of the low frequency exception items used in the analysis (those having codings of imageability in the MRC database) could be correctly read by the isolated orth→phon pathway. Hence, it is not the case that these items rely on orth→sem→phon because they cannot be read by orth→phon, but rather that they receive the most benefit from orth→sem→phon in their speed of computation. It had been assumed that many low frequency exceptions must be inaccurately read by orth→phon in order to account for the imageability effect. Considering that the imageability effect is found even with perfect orth→phon accuracy for the crucial items, however, the failure in Section 7.6 to find salient accuracy effects for low frequency exceptions is not so surprising.

As shown in Section 7.6, low frequency exceptions take the longest time to settle. Because activity is pooled from all available sources, the slower items have the most opportunity to be affected by the slower activity from orth→sem→phon.

The naming experiment was repeated with only the orth→phon path operational. The naming latencies for the normal model were subtracted from those obtained from this isolated pathway. The differences represented the speedup owed to the orth→sem→phon route. If the orth→sem→phon route contributed little to the naming latency of an item, this difference would be close to zero; it would be greater if the orth→sem→phon had a greater impact on the naming time for the item.

Figure 7.16 shows the differences, broken down by the variables of interest. Broadly, the exception items benefit more from the orth→sem→phon route than the regular items. For the high frequency exceptions, both the high and low imageability items receive a benefit. For the

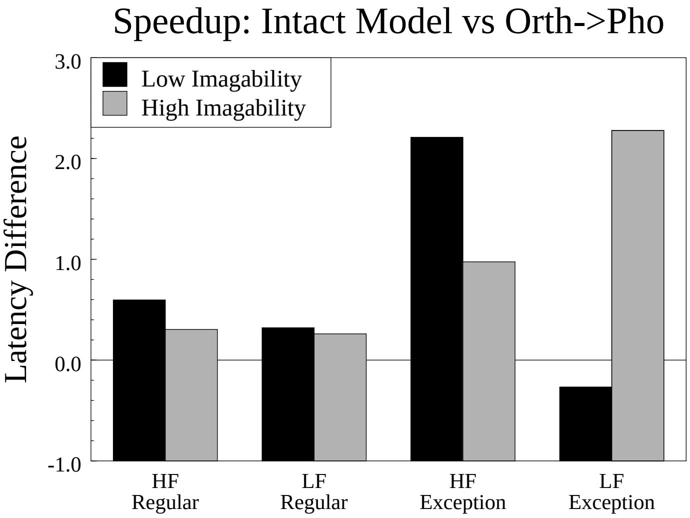


Figure 7.16: Difference in naming speed, orth→phon model versus intact model. Differences reflect the contribution of the orth→sem→phon path.

low frequency exceptions, the high imageability items receive a large benefit; the low imageability items receive virtually none.

The model therefore suggests a rather non-intuitive, yet quite sensible explanation for the imageability effect. Exceptions get a boost from orth→sem→phon more so than regulars. High frequency exceptions are robust; they can can ramp up semantic activations rapidly, regardless of the fragility of the semantic attractors (due to low imageability). As such, imageability is not a significant factor in their naming times. Low frequency exceptions are more fragile, due simply to their lower frequency. They are hence affected more by the dynamics of the semantic attractor, and hence by semantic variables such as imageability.

Chapter 8

General Discussion

8.1 Future Directions

This model is the first large-scale implementation of both the division of labor to phonology, and the division of labor to semantics. The results of this work are quite promising and important; a number of puzzling results have been clarified through exploration of this model. This work leaves open a wide range of future directions to take this research.

Thus far, the demonstrations of acquired dyslexia within the triangle framework have been accomplished via partial implementations, or through theorizing rather than actual demonstrations. A demonstration of acquired phonological dyslexia, for example, has not yet been accomplished. While one view of phonological dyslexia is that it results from the abolition of the orth→phon pathway, an alternative hypothesis is that purely phonological impairments drive the observed behavioral results. Currently, however, this is just theorizing; an explicit simulation could determine if such impairments could in fact give rise to the complex patterns of behavior seen in such patients. Such an explicit simulation could inform policies on the remediation of such patients.

Models of the development of reading have focused on the orth→phon part of the system, but orth→sem→phon provides an alternate method for pronunciation. Currently, there are no clear public policies for what to do with children who have impaired reading; schools typically either provide the child with more phonics-based drills, or more sight vocabulary drills. The computational framework presented here could be used to test different hypotheses about what to do with reading disabled children. Previous research has shown that there are two distinct forms of developmental dyslexia (Manis et al., 1996; Stanovich et al., 1997), but it is unclear what the proper remediation strategies ought to be for the different subgroups of reading impaired children.

Further policy implications for the teaching of reading could be extracted from this model. There is considerable controversy in the educational policy field as to the composition of reading basals (see Adams, 1990). One school of thought argues basals ought to be limited to simple words, while others emphasize the importance of meaningful, connected text (Smith, 1971). The model could be trained with training sets having different compositions, and their effects examined.

The general method presented here of creating large scale realistic semantic representations opens the door for serious simulation of a number of higher level phenomena. Smaller scale models of the phon→sem and sem→phon network presented here have been used to model the interaction of contextual and formal constraints in sentence processing (Allen, 1997), and the knowledge and processes utilized in grammaticality judgments (Allen & Seidenberg, In Press). Other, far simpler small scale networks have been used to model the conjunction of information used in various kinds of syntactic ambiguity resolution (e.g. McRae, Spivey-Knowlton, & Tanenhaus, 1998; Pearlmutter, Daugherty, MacDonald, & Seidenberg, 1994; Stevens et al., 1995). Armed with large scale semantic representations of words and a method of training sequences of items developed by Allen (1997), the interaction of syntactic, pragmatic and lexical constraints could be explored on a scale not previously possible.

The ability to model the learning of large sets of morphologically inflected items provides an opportunity to explore the relation between developmental dyslexia and developmental language deficits. Specific links between the two forms of impairments have been discovered in the empirical literature, but the exact nature of the common or disjoint causes of the two impairments is unclear (Joanisse, Manis, Keating, & Seidenberg, 1998). A model with explicit phonological and semantic attractors, and a realistic training set containing morphological regularities could shed light on such behavioral phenomena.

8.2 Conclusion

The current state of word recognition research is that the division of labor in reading is highly contentious. Researchers and policy makers have very differing intuitions about what factors would cause reading to rely more or less on one kind of processing. Empirical work has been largely contradictory and has not yielded clear answers.

One reason for this situation is the lack of an explicit model in which theoretical claims could be rigorously tested, and which empirical results could be explored to discover the computational principles underlying their genesis.

This work has identified a number of important factors which influence the division of labor. These factors guided the development of the first large scale implementation of the triangle model. This model provided insights into a number of puzzling or contradictory results in the literature. The picture that emerges from this work is that previous ways of thinking about the reading process (e.g., “is it with or without phonology?”), rooted in old-style assumptions about modular processing of information and atomic combining of information sources (cf. Fodor, 1983) are far too simplistic. A number of factors interact in complex ways to determine the relative weighting of different information sources in reading in an interactive and cooperative manner. Understanding such complex phenomena demands a more complex model of processing and stronger understanding of the underlying power and constraints of neurally inspired computation.

References

Adams, M. (1990). *Beginning to read.* Cambridge, MA: MIT Press.

Allen, J. (1997). Acquisition, processing and statistics. In *Tenth annual cuny conference on human sentence processing.* Santa Monica, CA.

Allen, J., & Seidenberg, M. S. (In Press). The emergence of grammaticality in connectionist networks. In B. MacWhinney (Ed.), *The emergence of language.* Mahwah, NJ: Erlbaum.

Baayen, R. H., Piepenbrock, R., & van Rijn, H. (1993). *The celex lexical database (cd-rom).* (Linguistic Data Consortium, University of Pennsylvania, Philadelphia, PA)

Balota, D. (1990). The role of meaning in word recognition. In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (p. 9-32). Hillsdale, NJ: Erlbaum.

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974). Comprehensive and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior*, *13*, 471-481.

Beauvois, M.-F., & Derouesné, J. (1979). Phonological alexia: Three dissociations. *Journal of Neurology, Neurosurgery and Psychiatry*, *42*, 1115-1124.

Bradley, L., & Bryant, P. (1983). Categorizing sounds and learning to read - A causal connection. *Nature*, *301*, 419-421.

Bub, D., Chancelliere, A., & Kertesz, A. (1985). Whole-word and analytic translation of spelling to sound in a non-semantic reader. In K. Patterson, M. Coltheart, & J. C. Marshall (Eds.), *Surface dyslexia* (p. 15-34). Hillsdale, NJ: Erlbaum.

Bullinaria, J. (1996). Connectionist models of reading: Incorporating semantics. In *Proceedings of the first european workshop on cognitive modeling* (p. 224-229). Berlin: Technische Universitat Berlin.

Bybee, J. L. (1988). Morphology as lexical organization. In M. Hammond & M. Noonan (Eds.), *Theoretical morphology* (p. 119-141). San Diego: Academic Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English.* New York: Harper & Row.

Churchland, P. (1989). *A neurocomputational perspective.* Cambridge, MA: MIT Press.

Cipolotti, L., & Warrington, E. K. (1995). Semantic memory and reading abilities: A case report. *Journal of the International Neuropsychological Society*, *1*, 104-110.

Collins, A., & Loftus, E. (1975). A spreading-activation theory of semantic processing. *Psychological Review*, *82*, 407-428.

Collins, A., & Quillian, M. (1969). Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, *8*, 240-248.

Coltheart, M. (1981). The MRC Psycholinguistic Database. *Quarterly Journal of Experimental Psychology*, *33A*, 497-505.

Coltheart, M. (1996). Phonological dyslexia: Past and future issues. *Cognitive Neuropsychology*, *13*(6), 749-762.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589-608.

Coltheart, M., Davelaar, E., Jonasson, K., & Besner, D. (1977). Access to the internal lexicon. In S. Dornic (Ed.), *Attention & performance VI.* Hillsdale, NJ: Erlbaum.

Coltheart, M., Langdon, R., & Haller, M. (1996). Computational cognitive neuropsychology and acquired dyslexia. In B. Dodd, L. Worrall, & R. Campbell (Eds.), *Evaluating theories of language: Evidence from disordered communication* (p. 7-36). London: Whurr Publishers.

Coltheart, M., Patterson, K. E., & Marshall, J. C. (Eds.). (1980). *Deep dyslexia.* London: Routledge & Kegan Paul.

Derouesné, J., & Beauvois, M. F. (1979). Phonological processing in reading: data from alexia. *Journal of Neurology, Neurosurgery and Psychiatry*, *42*, 1125-1132.

Devlin, J. T. (1998). *The role of dynamic representations in understanding semantic impairments: Investigations from connectionist neuropsychology.* Unpublished doctoral dissertation, University of Southern California, Los Angeles, CA.

Devlin, J. T., Gonnerman, L. M., Andersen, E. S., & Seidenberg, M. S. (1998). Category-specific semantic deficits in focal and widespread brain damage: A computational account. *jcn*, *10*(1), 77-94.

Flesch, R. (1955). *Why Johnny can't read.* New York: Harper & Brothers.

Fodor, J., & McLaughlin, B. P. (1990). Connectionism and the problem of systematicity: Why Smolensky's solution doesn't work. *Cognition*, *35*(2), 183-204.

Fodor, J. A. (1983). *The modularity of mind.* Cambridge, MA: MIT Press.

Fodor, J. A. (1998). *Concepts: Where cognitive science went wrong.* New York: Clarendon Press.

Fodor, J. A., & Lepore, E. (1992). *Holism: A shoppers guide.* Oxford; Cambridge, MA: Blackwell.

Fodor, J. A., & Pylyshyn, Z. W. (1988). Connectionism and cognitive architecture: A critical analysis. Special issue: Connectionism and symbol systems. *Cognition*, *28*, 3-71.

Francis, W. N., & Kuçera, H. (1982). *Frequency analysis of English usage.* Boston: Houghton-Mifflin.

Frawley, W. (1992). *Linguistic semantics.* hil: lea.

Friedman, R. B. (1996). Recovery from deep alexia to phonological alexia: Points on a continuum. *bl*, *52*, 114-128.

Funnell, E. (1983). Phonological processes in reading: New evidence from acquired dyslexia. *britjp*, *74*, 159-180.

Gonnerman, L., Devlin, J., Andersen, E. S., & Seidenberg, M. S. (1995). "Morphological" priming without a morphological level of representation. *Journal of the International Neuropsychological Society*, *1*, 142.

Gonnerman, L. M., Andersen, E. S., Devlin, J. T., Kempler, D., & Seidenberg, M. S. (1997). Double dissociation of semantic categories in Alzheimer's disease. *Brain and Language*, *57*(2), 254-279.

Harm, M. W., & Seidenberg, M. S. (1996). Computational bases of two types of developmental dyslexia. In *Proceedings of the eighteenth annual conference of the cognitive science society* (Vol. 18, p. 364-369). Mahwah, NJ: Erlbaum.

Harm, M. W., & Seidenberg, M. S. (1997). *The role of phonology in reading: A connectionist investigation.* (Paper presented at the 1997 Computational Psycholinguistics Conference, Berkeley, CA)

Harm, M. W., & Seidenberg, M. S. (1998). *Phonology, reading acquisition, and dyslexia: Insights from connectionist models.* (Manuscript submitted for publication.)

Hebb, D. O. (1949). *The organization of behavior.* New York: John Wiley & Sons.

Henderson, J. B. (1994). *Description based parsing in a connectionist network.* Unpublished doctoral dissertation, University of Pennsylvania.

Hetherington, P., & Seidenberg, M. S. (1989). Is there "catastrophic interference" in connectionist networks? In *Proceedings of the 11th annual conference of the cognitive science society* (p. 26-33). Hillsdale, NJ: Erlbaum.

Hinton, G. E. (1989). Deterministic boltzmann learning performs steepest descent in weight-space. *Neural Computation*, *1*(1), 143-150.

Hinton, G. E., & Shallice, T. (1991). Lesioning an attractor network: Investigations of acquired dyslexia. *Psychological Review*, *98*(1), 74-95.

Hoeffner, J. H. (1996). *A single mechanism account of the acquisition and processing of regular and irregular inflectional morphology.* Unpublished doctoral dissertation, Department of Psychology, Carnegie Mellon University, Pittsburgh, PA.

Howard, D., & Best, W. (1996). Developmental phonological dyslexia: Real word reading can be completely normal. *Cognitive Neuropsychology*, *13*(6), 887-934.

Hummel, J. E., & Biederman, I. (1992). Dynamic binding in a neural network for shape recognition. *Psychological Review*, *99*(3), 480-517.

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, *29*(6), 687-715.

Jared, D., & Seidenberg, M. S. (1991). Does word identification proceed from spelling to sound to meaning? *Journal of Experimental Psychology: General*, *120*(4), 358-394.

Joanisse, M. F., Manis, F. R., Keating, P., & Seidenberg, M. S. (1998). *Speech perception, phonology, morphology: Heterogeneity of language deficits in dyslexic children.* (Manuscript submitted for publication)

Joanisse, M. F., & Seidenberg, M. S. (1998). *Dissociations between rule-governed forms and exceptions: A connectionist account.* (Poster presented at the 1998 Meeting of the Cognitive Neuroscience Society (San Francisco CA))

Jorm, A. F., & Share, D. L. (1983). Phonological recoding and reading acquisition. *Applied Psycholinguistics*, *4*, 103-147.

Keil, F. C. (1989). *Concepts,kinds, and cognitive development.* Cambridge, MA: MIT Press.

Kelly, M. H. (1992). Using sound to solve syntactic problems: The role of phonology in grammatical category assignments. *Psychological Review*, *99*(2), 349-364.

Koch, C., & Segev, I. (Eds.). (1989). *Methods in neuronal modeling: From synapses to networks.* Cambridge, MA: MIT Press.

Krashen, S. D. (1996). *Every person a reader.* Language Education Associates, P.O. box 7416, Culver City, California 90233.

Lesch, M. F., & Pollatsek, A. (1993). Automatic access of semantic information by phonological codes in visual word recognition. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *19*(2), 285-294.

Levin, B. (1993). *English verb classes and alternations.* Chicago: University of Chicago Press.

Liberman, A., Harris, K., Hoffman, H., & Griffith, B. (1957). The discrimination of speech sounds within and across phoneme boundaries. *Journal of Experimental Psychology*, *54*(5), 358-368.

Lukatela, G., & Turvey, M. T. (1994a). Visual lexical access is initially phonological: 2. Evidence from phonological priming by homophones and pseudohomophones. *Journal of Experimental Psychology: General*, *123*(4), 331-353.

Lukatela, G., & Turvey, M. T. (1994b). Visual lexical access is initially phonological: I. Evidence from associative priming by words, homophones, and pseudohomophones. *Journal of Experimental Psychology: General, 123*(2), 107-128.

Lund, K., Burgess, C., & Atchley, R. A. (1995). Semantic and associative priming in high-dimensional semantic space. In *Proceedings of the 17th annual conference of the cognitive science society* (p. 660-665). Hillsdale, NJ: Erlbaum.

Lundberg, I., Olofsson, A., & Wall, S. (1980). Reading and spelling skills in the first school years predicted from phonemic awareness skills in kindergarten. *Scandanavian Journal of Psychology, 21*, 159-173.

MacDonald, M. C. (1993). The interaction of lexical and syntactic ambiguity. *Journal of Memory and Language, 32*, 692-715.

Manis, F., Seidenberg, M., Doi, L., McBride-Chang, C., & Peterson, A. (1996). On the basis of two subtypes of developmental dyslexia. *Cognition, 58*, 157-195.

Mann, V. A. (1984). Longitudinal prediction and prevention of early reading difficulty. *Annals of dyslexia, 34*, 115-136.

Marcus, M., Santorini, B., & Marcinkiewicz, M. A. (1993). Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics, 19*, 313-330.

Marshall, J. C., & Newcombe, F. (1973). Patterns of paralexia: A psycholinguistic approach. *Journal of Psycholinguistic Research, 2*, 175-199.

McCarthy, R., & Warrington, E. K. (1986). Phonological reading: Phenomena and paradoxes. *Cortex, 22*, 359-380.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419-457.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part 1. An account of basic findings. *Psychological Review, 88*, 375-407.

McCloskey, M., & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 23). New York, NY: Academic Press.

McRae, K., de Sa, V. R., & Seidenberg, M. S. (1997). On the nature and scope of featural representations of word meaning. *Journal of Experimental Psychology: General, 126*(2), 99-130.

McRae, K., Spivey-Knowlton, M. J., & Tanenhaus, M. K. (1998). Modeling the influence of thematic fit (and other constraints) in on-line sentence comprehension. *Journal of Memory and Language, 38*, 283-312.

Miller, G. A. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography, 3*, 235-312.

Murphy, G. L., & Medin, D. L. (1985). The role of theories in conceptual coherence. *Psychological Review, 92*(3), 289-316.

Murphy, L., & Pollatsek, A. (1994). Developmental dyslexia: Heterogeneity without discrete subgroups. *Annals of Dyslexia, 44*, 120-146.

Nazzi, T., Bertoncini, J., & Mehler, J. (1998). Language discrimination by newborns: Toward an understanding of the role of rhythm. *Journal of Experimental Psychology: Human Perception and Performance, 24*(3), 756-766.

Olson, R., Wise, B., Conners, F., Rack, J., & Fulker, D. (1989). Specific deficits in component reading and language skills: Genetic and environmental influences. *Journal of Learning Disabilities, 22*(6), 339-348.

O'Reilly, R. C. (1996). Biologically plausible error-driven learning using local activation differences: The generalized recirculation algorithm. *Neural Computation, 8*(5), 895-938.

Paap, K. R., & Noel, R. W. (1991). Dual route models of print to sound: Still a good horse race. *Psychological Research, 53*, 13-24.

Patterson, K., Coltheart, M., & Marshall, J. C. (Eds.). (1985). *Surface dyslexia.* Hillsdale, NJ: Erlbaum.

Patterson, K., Seidenberg, M. S., & McClelland, J. L. (1989). Connections and disconnections: Acquired dyslexia in a computational model of reading processes. In R. G. M. Morris (Ed.), *Parallel distributed processing: Implications for psychology and neuroscience* (p. 131-181). London: Oxford University Press.

Patterson, K., Suzuki, T., & Wydell, T. N. (1996). Interpreting a case of Japanese phonological alexia: The key is in phonology. *Cognitive Neuropsychology, 13*, 803-822.

Patterson, K. E., Marshall, J. C., & Coltheart, M. (Eds.). (1985). *Surface dyslexia: Neuropsychological and cognitive studies of phonological reading.* London: Erlbaum.

Pearlmutter, B. A. (1989). Learning state space trajectories in recurrent neural networks. *Neural Computation, 1*(2), 263-269.

Pearlmutter, B. A. (1995). Gradient calculations for dynamic recurrent neural networks: A survey. *IEEE Transactions on Neural Networks, 6*(5), 1212-1228.

Pearlmutter, N. J., Daugherty, K. G., MacDonald, M. C., & Seidenberg, M. S. (1994). Modeling the use of frequency and contextual biases in sentence processing. In *Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (p. 699-704). Hillsdale, NJ: Erlbaum.

Perani, D., Cappa, S., Bettinardi, V., Bressi, S., Gornotempini, M., Matarrese, M., & Fazio, F. (1995). Different neural systems for the recognition of animals and manmade tools. *Neuroreport*, *6*, 1637-1641.

Pinker, S. (1991). Rules of language. *Science*, *253*, 530-535.

Plaut, D. C. (1991). *Connectionist neuropsychology: The breakdown and recovery of behavior in lesioned attractor networks.* Unpublished doctoral dissertation, School of Computer Science, Carnegie Mellon University. (Available as Technical Report CMU-CS-91-185.)

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, *12*, 765-805.

Plaut, D. C., & Kello, C. T. (In Press). The emergence of phonology from the interplay of speech comprehension and production: A distributed connectionist approach. In B. MacWhinney (Ed.), *The emergence of language.* Mahwah, NJ: Erlbaum.

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

Plaut, D. C., & Shallice, T. (1991). Effects of word abstractness in a connectionist model of deep dyslexia. In *Proceedings of the thirteenth annual conference of the cognitive science society* (p. 73-78). Hillsdale, NJ: Erlbaum.

Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377-500.

Putnam, H. (1989). *Representation and reality.* Cambridge, MA: MIT Press.

Repp, B. H. (1984). Categorical perception: Issues, methods, findings. In N. J. Lass (Ed.), *Speech and language: Advances in basic research and practice* (Vol. 10, p. 243-335). Orlando: ap.

Rips, L., Shoben, E., & Smith, E. (1973). Semantic distance and the verification of semantic relations. *Journal of Verbal Learning and Verbal Behavior*, *12*, 1-20.

Rosner, J., & Simon, D. P. (1971). The auditory analysis test: An initial report. *Journal of Learning Disabilities*, *4*(7), 40-48.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing, vol. 1.* Cambridge, MA: MIT Press.

Saffran, J., Newport, E., Aslin, R., & Tunick, R. (1997). Incidental language learning: Listening (and learning) out of the corner of your ear. *Psychological Science*, *8*(2), 101-105.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996). Statistical learning by 8-month-old infants. *Science*, *274*(5294).

Seidenberg, M. S. (1988). Cognitive neurospychology and language: The state of the art. *Cognitive Neuropsychology, 5*(4), 403-426.

Seidenberg, M. S. (1990). Lexical access: Another theoretical soupstone? In D. A. Balota, G. B. Flores d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (p. 33-72). Hillsdale, NJ: Erlbaum.

Seidenberg, M. S. (1995). Visual word recognition: An overview. In P. Eimas & J. L. Miller (Eds.), *Handbook of perception and cognition: Language.* New York: Academic Press.

Seidenberg, M. S., & Harm, M. (1995). Division of labor in a multicomponent model of visual word recognition. In *Proceedings of the 36th annual meeting of the psychonomic society.* Los Angeles, CA.

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

Seidenberg, M. S., & Plaut, D. C. (In Press). *Evaluating word reading models at the item level: Where has all the variance gone?* (To appear in Psychological Science)

Seidenberg, M. S., Waters, G. S., Barnes, M. A., & Tanenhaus, M. K. (1984). When does irregular spelling or pronunciation influence word recognition? *Journal of Verbal Learning and Verbal Behavior, 23*(3), 383-404.

Shallice, T. (1988). *From neuropsychology to mental structure.* Cambridge: Cambridge University Press.

Shankweiler, D., & Liberman, I. (Eds.). (1989). *Phonology and reading disability: Solving the reading puzzle.* Ann Arbor, Michigan: University of Michigan Press.

Share, D. L. (1995). Phonological recoding and self-teaching: *sine qua non* of reading acquisition. *Cognition, 55*, 151-218.

Share, D. L., Jorm, A. F., Maclean, R., & Matthews, R. (1984). Sources of individual differences in reading acquisition. *Journal of Educational Psychology, 76*(6), 1309-1324.

Shastri, L., & Ajjanagadde, V. (1993). From simple associations to systematic reasoning: A connectionist representation of rules, variables, and dynamic bindings using temporal synchrony. *Brain and Behavioral Sciences, 16*, 417-494.

Smith, F. (1971). *Understanding reading.* New York: Holt, Rinehart and Winston.

Smith, F. (1973a). The politics of ignorance. In R. Winklejohann (Ed.), *The politics of reading: Point-Counterpoint.* Newark, Delaware: IRA/ERIC.

Smith, F. (1973b). *Psycholinguistics and reading.* New York: Holt, Rinehart and Winston.

Smolensky, P. (1986). Foundations of harmony theory: Cognitive dynamical systems and the subsymbolic theory of information processing. In D. Rumelhart & J. McClelland (Eds.), *Parallel distributed processing, vol. 1.* Cambridge, MA: MIT Press.

Smolensky, P. (1990). Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, *46*, 159-216.

Stanovich, K., Siegel, L., & Gottardo, A. (1997). Converging evidence for phonological and surface subtypes of reading disability. *Journal of Educational Psychology*, *89*(1), 114-127.

Stevens, K., Harm, M., Schuster, S., & MacDonald, M. (1995). Aging and the use of context and frequency information during ambiguity resolution. In *Eighth annual cuny conference on human sentence processing.* Tucson, AZ.

Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory and Cognition*, *21*, 1140-1154.

Taraban, R., & McClelland, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory & Language*, *26*(6), 608-631.

Treiman, R. (1986). The division between onsets and rimes in English syllables. *Journal of Memory and Language*, *25*, 476-491.

Tunmer, W. E., & Nesdale, A. R. (1985). Phonemic segmentation skill and beginning reading. *Journal of Educational Psychology*, *77*, 417-427.

Van Orden, G. C. (1987). A ROWS is a ROSE: Spelling, sound and reading. *Memory and Cognition*, *15*(3), 181-198.

Van Orden, G. C. (1991). Phonologic mediation is fundamental to reading. In D. Besner & G. W. Humphreys (Eds.), *Basic processes in reading* (p. 77-103). hil: lea.

Van Orden, G. C., Johnston, J. C., & Hale, B. L. (1988). Word identification in reading proceeds from the spelling to sound to meaning. *Journal of Experimental Psychology: Memory, Language and Cognition*, *14*, 371-386.

Van Orden, G. C., Pennington, B. F., & Stone, G. O. (1990). Word identification in reading and the promise of subsymbolic psycholinguistics. *Psychological Review*, *97*(4), 488-522.

Warren, R. (1970). Perceptual restoration of missing speech sounds. *Science*, *167*, 392-393.

Waters, G. S., & Seidenberg, M. S. (1985). Spelling-sound effects in reading: Time course and decision criteria. *Memory and Cognition*, *13*(6), 557-572.

Williams, R. J., & Peng, J. (1990). An efficient gradient-based algorithm for on-line training of recurrent network trajectories. *Neural Computation*, *2*, 490-501.

Zipf, G. K. (1935). *The psycho-biology of language: An introduction to dynamic philology.* Boston, MA: Houghton Mifflin.