A Distributed Model of the English Past Tense Formation

Steve R. Howell (steven.howell@keystone.edu)
Department of Psychology, Keystone College, 1 College Green, La Plume, PA 18440 USA

Maryellen C. MacDonald (mcmacdonald@wisc.edu) & Mark S. Seidenberg (seidenberg@wisc.edu)
Department of Psychology, University of Wisconsin-Madison, 1202 W. Johnson Street, Madison, WI 53706 USA

Abstract

We developed a version of the Joanisse and Seidenberg (1999) past-tense model to address two issues: whether the model's performance depended on the use of localist semantic representations, and the challenges to this account presented by a patient who was impaired in generating irregular past tenses despite apparently intact semantics. The model also demonstrates the frequency by regularity interaction from Patterson et al (2001), and shows that a single-mechanism connectionist model can perform realistically on the past-tense task.

Keywords: Past tense; neural network; language

A Distributed Model of the English Past Tense Formation

For more than twenty years, the process of past-tense formation in English has served as a battleground between competing theories of language processing (see Seidenberg and Plaut, in press, for a review). In English, the past tense of a verb is usually formed by adding the inflectional morpheme spelled –ed to the end of the verb. However, for about 180 irregular verbs, the past tense is irregular, formed not by adding –ed but via a vowel change or some other mechanism (e.g. run-*ran,* keep-*kept,* go-*went).* The differences between regulars and irregulars are standardly taken as indicating that they involve different types of knowledge and processing mechanisms. Regular forms are generated by applying a rule, whereas the irregulars are stored in memory. Much evidence has been marshalled in support of this "dual-mechanism" account (e.g. Pinker and Ullman, 2002).

An alternative viewpoint holds that past tenses are generated by a common lexical system encoding relations among phonology, semantics, and (in literate individuals) orthography. The system (which underlies many lexical phenomena, not just verb morphology) picks up on statistical regularities in the mappings between codes. A central claim is that generating irregular forms tends to require greater input from semantics compared to regulars. This theory and its application to patients with semantic impairments is summarized elsewhere (McClelland & Patterson, 2002; Seidenberg and Plaut, in press). The approach holds that the distinction between regular forms and irregular forms is invalid because of their overlapping structure, and that the past tense rule is an idealization that abstracts away from details of the actual processing mechanism.

One of the more successful recent models of the past tense formation was developed by Joanisse and Seidenberg (1999) . Earlier models of the past tense (e.g. Rumelhart and McClelland, 1986) had focused on tasks involving mappings between the phonological forms of the present and past tense. By design such models could not address any semantic phenomena, and thus they could not distinguish between homophones with different past tenses such as ring-*ringed* and ring-*rang*. The Joanisse and Seidenberg (1999) model (hereinafter J&S) was an advance in two respects: it incorporated both semantics and phonology, and it acquired this knowledge in the course of performing several language functions or tasks. These tasks included hearing or comprehending (mapping input phonemes to semantics), speaking (mapping semantics to output phonemes), repeating (mapping input phonemes directly to output phonemes) and transforming (past tense formation – mapping input phonemes of a verb to the output phonemes of the past tense of the verb).

The J&S model was trained on the present and past tenses of 600 monosyllabic verbs, of which 64 had irregular past tenses. The repeating tasks included an additional 594 English verbs to increase the model's exposure to English phonology. After training, the model performed quite well, exhibiting correct performances after training of 99.8%, 99.5%, 98.2% and 99.3% on speaking, hearing, repeating, and transforming respectively. The model was tested on the 20 nonce verbs from Ullman et al. (1997) in order to determine its capacity to generalize to verbs which had not been included in the training set. Test performance was 90%, and even the errors were of a type that people occasionally produce.

The trained model was then lesioned in two ways, phonologically and semantically. Phonological damage affected performance on all three types of verbs, but had the largest effect on nonwords. Semantic damage also affected all three types of verbs, but the effect was largest for irregular verbs. The conclusion from the model's performance was that the "double dissociation" observed across patient groups can be replicated by different types of lesions in a system that does not include separate 'rule' and 'exception' mechanisms (Joanisse and Seidenberg, 1999). Further, the model's errors were also broadly consistent with the patient data.

Later work highlighted some of the limitations of this model, chiefly that it used an arbitrary localist representation for semantics that did not include any real word meaning or allow for different degrees of similarity among semantic representations of words. While neuropsychological data on anterior lesions was interpreted by Pinker and Ullman (2003) as a challenge to the J&S model, work by Bird et al (2003) refuted Ullman's own findings (Ullman et al., 1997) by identifying a confound of phonological complexity in the experimental materials with their word representations.

More challenging evidence was introduced by Miozzo (2003). Miozzo presented evidence of a neuropsychological patient (AW) with acquired brain damage who encountered problems accessing phonology in speech production, but seemed to have intact ability to access word meaning. In the context of J&S, this would mean damage to the phonological units, but intact semantic units. AW was better able to produce the past tenses of regular verbs than irregular verbs, the opposite of what one would expect if J&S were correct. That is, damage to semantics (posterior lesions) is thought to impair irregular performance more than regular, with damage to phonology (anterior lesions) affecting regulars more.

The appearance of a selective deficit for irregulars when lexical access is impaired was argued to be more in line with a dual-mechanism account, since it specifies that irregular forms are specified in the lexicon while regulars are processed via a "rule" mechanism. Specifically, Miozzo claimed that J&S could not account for such a deficit, a problem for the single-mechanism viewpoint.

The critical issue about the Miozzo patient is not whether semantics is "intact" but rather his ability to use this information in performing different tasks. The claim that semantics was well-preserved derived from performance on one type of task, word-picture matching. This task speaks to properties of the mapping from phonology to semantics which was well-enough preserved to allow performance at a high level. However, the patient was also severely anomic, unable to generate names of objects. This task speaks to properties of the mapping from semantics to phonology, which was highly impaired. The past tense generation task used to assess verb knowledge involves speech production; given the present tense as input, produce the past tense. In the J&S model, semantics is relevant to generating past tenses for irregular verbs. The patient's poor performance on irregulars follows from the inability to use semantics to generate phonology.

In short, performance on irregular past tenses can be impaired by damage to semantic representations (as in semantic dementia) or in the use of semantic information to compute phonology (as in anomia). The computation from sound to meaning is not immediately relevant to the past tense generation task and neither is AW's ability to perform tasks involving this computation. The computation from meaning to sound *is* highly relevant to past tense generation, especially for irregular words, and so is AW's severe anomia.

Although this account is consistent with the J&S model and later models emphasizing the roles of semantics and phonology in performing various tasks (e.g. Patterson et al, 2009), it is important to determine if the model will perform in the expected ways when implemented, while continuing to account for other phenomena.

Another interesting finding related to the past tense, is the frequency by regularity interaction reported by Patterson, Lambon-Ralph, Hodges, and McClelland (2001). They found that, in patients with semantic lesions, more damage occurred to irregulars than regulars, and especially to lower-frequency irregulars. Furthermore, McClelland and Patterson (2003) argue that the frequency effect for exceptions cannot be accounted for by the dual route model, only by the interacting operation of a connectionist model. Hence, we believe this effect is central to a complete model of the past tense. The J&S model did not demonstrate such an interaction, but we believe that a similar model which incorporates richer semantics, thus reflecting human processing more accurately, should also demonstrate something like this frequency by regularity effect.

Thus, in the present work, we revise and extend the J&S model to incorporate richer, more meaningful distributed semantic representations, and test it to two purposes. One, this revised model should be able to demonstrate the alternative account of AW's deficit, and so we will simulate both types of damage to our revised PT model and compare the results. We expect to find no difference, and thus demonstrate that Miozzo's patient is no challenge to the connectionist account of the past tense formation.

Second, the model should exhibit the frequency by regularity interaction when subject to a posterior (semantic) lesion, a la Patterson et al. (2001). This sort of effect in models is often dependent upon the use of a very realistic training set that closely mimics the language that a human is exposed to, and hence we will pay close attention to incorporating the particular stimulus sets (in this case verbs, matched for various dimensions such as frequency, phonological complexity, etc.) that have recently been used in human experiments (e.g. Ullman et al, 1997, Patterson et al, 2001, Bird et al. 2003).

Method

We attempted to extend the J&S connectionist model of past tense processing to the use of distributed representations in order to account for the findings of Miozzo (2003) and Bird et al (2003). The revised model is shown in Figure 1.

Words were represented in various codes depending on the subsystem involved in the past-tense generation task: speech input (a phonological representation of the sounds of words), speech output (a similar code used in generating speech), and distributed semantic representations (the meanings of words). In humans, speech input consists of continuous acoustic patterns that are recognized and parsed into sequences of discrete phonemes by a process of categorical perception. Speech output, on the other hand, consists of sequences of articulatory gestures subserved by the motor cortex that result in the physical production of sound. Some have suggested that, with experience, humans incorporate both auditory and motor cortex representations of sound into a single phonemic representation. As a simplifying assumption, we have followed this evidence and used exactly the same distributed phonological code for both input and output.

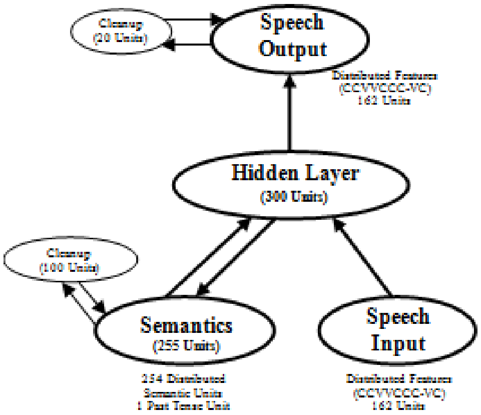

Figure 1: Model Architecture.

These representations employed a CCVVCCC-VC template (C=Consonant, V=Vowel), with each phoneme represented via 18 binary phonological features. Each words vowel was aligned with the first V [VV was used to represent dipthongs such as the /oy/ in (BOY)]. Initial consonants were aligned with the C slots from right to left, and following consonants were aligned with C slots from left to right. The final VC was used represent the /Id/ syllable in words such as TASTED. Any units in empty slots were set to 0.0 activation.

As pointed out in J&S, the use of a distributed code for the phonological representations allowed the model to represent degrees of similarity between words. A knowledge of the similarities among words is essential to the model's ability to generalize. Units on the speech output layer were connected to and from a series of "cleanup" units (Hinton & Shallice, 1991; Plaut & Shallice, 1993). These units provided a way of representing nonlinearly separable (and hence more complex) phonological dependencies and made the processing of phonological output a dynamic process in which the model settles into a final pattern over a series of time steps (McClelland and Rumelhart, 1981; Plaut, McClelland, Seidenberg, and Patterson, 1996; Harm & Seidenberg, 1999).

In the semantic layer, each verb was represented as a distributed pattern of activation over a set of 254 semantic "features". These semantic representations were provided by David Plaut (Plaut, 2004, Personal Communication) who created them via an LSA-like process (e.g. Landauer, 1997). Of course, the semantic bits used in this sort of distributed semantic representation are not meaningful in themselves, and hence are not really "features" in the sense that, for example, the feature sets of Howell, Jankowicz, and Becker (2005) are. The similarity of one verb to another is only found in the overall pattern of overlap of the 254 bits, including their covariance. However, at present such meaningful feature sets for words have limited vocabularies, such as 450 words or so, and hence would not be capable of representing the 1300+ verbs used in the present model. One additional node (the 255th) was used in the semantic representations to indicate present or past tense. The semantic layer was also connected to a cleanup layer of 100 units. This size is dramatically increased from the J&S model, due to the increased complexity of the semantic relationships in the Plaut semantic representations

One major advantage of the J&S model was the way it incorporated a variety of language tasks into the learning of the past tense. That is, people acquire their knowledge of language by using it for different purposes. What is learned from one task, such as speaking, may affect the ability to perform other tasks, such as hearing (especially if the phonemic representations of the two overlap, as discussed above). We approximated this aspect of human learning by interleaving training on three tasks. Speaking involved taking the semantic representation of a present or past tense verb as input and generating its phonology. This maps to a person with a semantic meaning in mind who then must articulate it. Hearing involved taking the phonological code of a word as input and activating its semantic meaning. This is of course the opposite process to the above. Transforming, the task most specific to the issue at hand, involved taking the phonology of a verb and an indication of past-tense semantics (turning the past tense bit ON) as input and generating past tense phonology. The model had to find a set of weights on its connections that allowed it to perform all of these tasks accurately.

The J&S model actually incorporated a fourth task, repeating, which was added to give that model more experience with the structure of English phonology. For the present model we eliminated this task, as it was less central than the other three, and tended to cause repetition errors on the past tense task. Also, in the present model we were able to include all 1365 words in all three tasks, rather than having a large subset used only for the repeating task.

The model was trained on the present and past tense of 1365 monosyllabic English verbs, consisting of the vocabulary used in the J&S model plus the non-overlapping words drawn from Patterson, Lambon-Ralph, Hodges &McClelland (2001), and Bird et al. (2003). These additional words were included to make it possible to more closely map to the human data reported therein by using the same stimulus words.

The verbs were presented to the model with a probability equal to their logarithmic frequency. Task probabilities were set at: Speaking, 20%, Hearing, 40%, Transforming (present-past), 40%). Other simulations indicated that the model's performance was not highly sensitive to the exact proportion of trials of each type. The network was trained using the backpropagation-through-time algorithm (Williams and Peng, 1990) and the MikeNet simulator created by Mike Harm. Each trial began with the random selection of an item (verb) and a task. The input appropriate for a given task was presented, and activation was propagated throughout the network for seven time steps. Weights were then adjusted based on the discrepancy between the observed and the expected patterns. Initial (pre-training) weights were randomized to small values between -0.01 and 0.01. The learning rate was set to 0.001, a smaller value than used in J&S which seemed to allow for smoother settling in this more complex model. A logistic activation function was used, and error was calculated using the cross-entropy measure (Hinton, 1989).

Results

Training was halted after 1.7 million training trials, at the point where overtraining seemed to be becoming an issue. At this point the accuracy of the trained network was assessed on all three tasks (Speaking, Hearing, and Transforming), over all of the words in the training set. For the Speech Input and Speech Output layers, words were scored phoneme by phoneme, using a Euclidian Distance metric to select the phoneme closest to the network's output. If the closest phoneme differed from the target phoneme, it was scored as incorrect. If any of a word's several phonemes were incorrect, the entire word was scored as incorrect. Semantically, the word closest to the network's output was selected via the same Euclidian Distance metric. If this selected word differed from the target word, it was scored as incorrect. Accuracy on all tasks on the training set was quite good: Speaking, 1288 correct out of 1365, or 94.4%; Hearing, 1289 correct of out 1365, or 94.4%; and Transforming, 658 correct out of 685, or 96.1%. Note that there are fewer trials (685) for the Transforming task, since only present tense verbs can be transformed into the past tense. However, each trial uses a present tense verb as phonological input, and the past tense as phonological output, so all 1365 words are still being used in all tasks.

The ability of the model to generalize and produce the past tense for words on which it had not been trained was assessed (as in J&S) using the 20 nonce words from Ullman et al (1997). As these are meaningless nonwords without semantics, the model was given only the phonological code of the nonce verb as input, and the past-tense semantics bit as input (indicating the model should perform the transformation task). Using the same scoring criteria as above, the network generated acceptably 'correct' past tenses on 17 of the 20 nonce verbs, or 85% correct.

Accuracy on a variety of other test sets was also calculated. First, two sets of test words (Regulars and Irregulars) that were NOT included in the training set were presented to the model. These verbs differ from the Ullman nonce verbs above in that they have actual semantic representations attached, but the model simply hasn't been trained on producing their past tense, only on the two other tasks. Thus, unlike with the nonce verbs, it is possible that semantic similarity between these novel verbs and other, known, verbs could influence the formation of the past tense form, in addition to phonological similarity as for the nonce verbs.

Performance on the Regular test was perfect (20/20 correct; 100%), demonstrating that the model has acquired the add –ed "rule" and is able to generalize well to novel regular verbs. Performance on the Irregular test set was good (13/16 correct; 81.2%), but understandably not as good as for the Regulars.

Irregular Deficits with Intact Semantics

To demonstrate that the kind of deficit that Miozzo's patient exhibited can be produced by damage to either semantic representations themselves, or to the connections from semantics to phonology, we performed two different lesions to the model and tested it under both conditions.

First, the semantic units themselves were lesioned (Semantic Layer lesions or SL) by adding a varying amount of noise to degrade their operation. This proportion was gradually increased to illustrate the progression of damage under lesions of varying severity. As expected, irregulars were hit hardest by this type of lesion (see Figure 2), which would in humans have corresponded to damage to the semantic association area in temporal cortex.

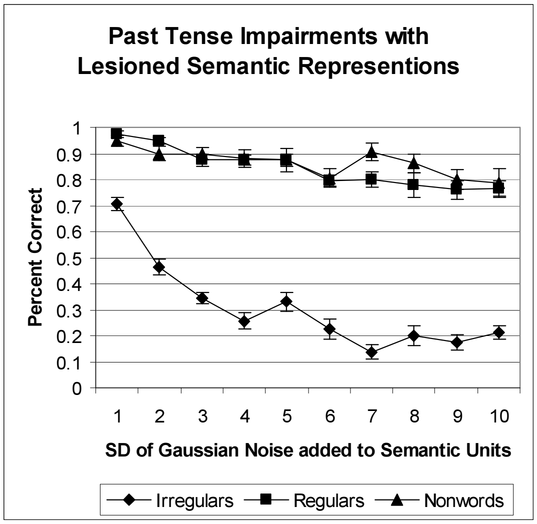

Figure 2: Performance after Semantic Lesion (SL)

Secondly, we tested the model after lesioning only the connections from the semantic layer to the hidden layer (SH lesions), and leaving the semantic layer's units intact. This would correspond to a patient with intact semantics (as in Miozzo's patient) who nonetheless had difficulty with retrieving the form of a word to go with a meaning (lexical retrieval deficit). Again, regular verbs and nonwords were largely intact even at high levels of damage, but irregular verbs showed a severe impairment that increased with the level of damage (See Figure 3).

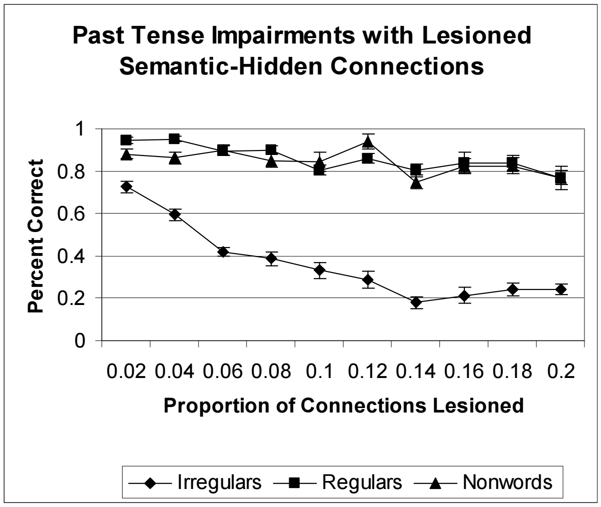

Figure 3: Performance after Semantic-Hidden Lesion

Frequency by Regularity Interaction

We simulated a semantic lesion by randomly severing connections from the semantic layer to the hidden layer with a probability of 0.2. After lesioning, the model was tested again, this time with a set of known words (words that had been included in the training set) that were derived from the test set used in Patterson et al. (2001). These verbs are divided into five lists, as discussed previously, and matched as closely as possible on phonological characteristics: low frequency regulars, high frequency regulars, low frequency irregulars; high frequency irregulars, and very high frequency irregulars. On this test set, the differential effect of the lesion can be clearly seen (Figure 4).

For regular verbs, there is no significant difference between low frequency and high frequency items. However, for irregular verbs, the low frequency items suffer the most after lesioning, with the high frequency items being more spared and the very high frequency items being even less affected. The difference between the irregular low frequency items and the irregular high frequency items is significant (t(18), p = 0.03), as is the difference between irregular high frequency items and irregular very high frequency items (t(18), p < 0.001). This matches the frequency by regularity interaction in the human data reported in Patterson et al., and suggests our model is capturing something even closer to the human experience than did the J&S model, thanks to its distributed semantics.

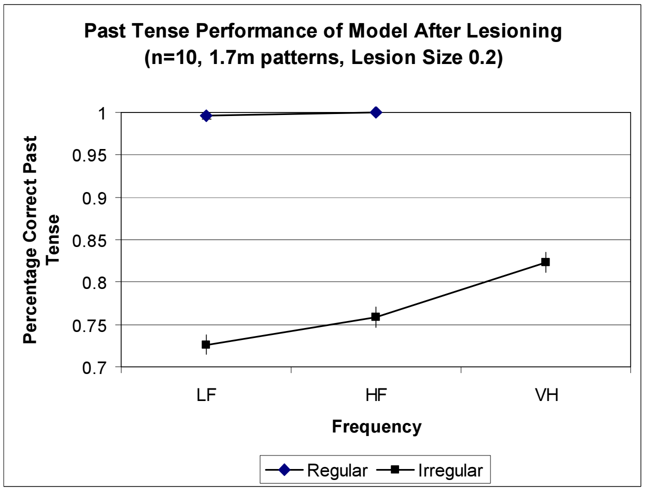

Figure 4: Performance after Semantic lesion, by regularity and frequency

Discussion

The J&S model (Joanisse and Seidenberg, 1999) of past tense performance was a significant advance, showing that a connectionist model could exhibit accurate, human-like performance on the past tense task while accounting for a variety of neuropsychological evidence. Furthermore, it did this at a subsymbolic level, simply by learning the statistical regularities between input phonology, output phonology, and semantics. However, its artificial localist representation of semantics prevented the J&S model from making use of semantic similarity in the past tense task, something that humans certainly could. Thus in the present model we used richer, distributed semantics. It is important to note that Woollams, Joanisse and Patterson (2009) have, since this present work was conducted, also developed an extension of the original J&S model that incorporates distributed semantics, replicating and extending earlier findings.

First, thanks to the richer, distributed semantics incorporated in this model, semantic similarity could now exert its effects, if any, on the past tense task. Second, the possession of richer semantics allowed the model to account for the patient data that was argued by Miozzo (2003) and Pinker and Ullman (2003) to be a challenge to the connectionist account of the past tense. For these tasks, we demonstrated that a lesion to the semantic representations themselves (posterior lesion), or to the semantic-to-lexical connections (perhaps some anterior lesions), will both result in a deficit for irregulars in the past tense task. Perhaps what Miozzo has demonstrated, however, is the importance of additional patient testing to determine where in the process the lesion has occurred. For example, if semantic access is intact (as in AW), but word production is not, then the lesion is likely along the neural pathways from semantics to phonology, rather than in semantic cortex.

Patients like AW do require us to emphasize what we believe to be the central tenet of the connectionist approach to the past tense, at least insofar as it contrasts with the dual-route approach. That is, the connectionist approach emphasizes that there is a single *mechanism* or process that the brain uses to compute the past tense. That this process might be distributed and involve multiple physical "routes" should come as no surprise to those familiar with neural networks. Several areas of the brain (e.g. auditory phonological representations, articulatory phonological representations, semantic representations.) might be involved in the neural network that subserves this task, and they might be located in separate areas of cortex (e.g. auditory cortex, motor cortex, semantic association cortex, etc.) where it is possible that they will be damaged separately, and that component's contribution impaired. However, this is not damage or impairment of a distinct route, but rather of a part of the mechanism, and it will affect the processing of *all* words, regular, irregular, novel or nonword.

We believe that the difference is between a focus on delineated, non-interacting, modular or symbolic processes on the one hand (dual route), and a distributed, multiply-interacting, subsymbolic and statistical process on the other hand (connectionist). This is not a division that is specific to the past tense debate, but it is perhaps where the most obvious battleground has been. The present model provides additional evidence that the connectionist approach can continue to account for data that is argued by some to require the other approach.

References

Bird, H., Lambon Ralph, M. A., Seidenberg, M. S., McClelland, J. L., & Patterson, K. (2003). Deficits in phonology and past-tense morphology: What's the connection? *Journal of Memory and Language, 48*, 502-526.

Harm, M. W. & Seidenberg, M. S. (1999) Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological Review, 106,* 491-528

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence, 40,* 185-234.

Hinton, G. E., & Shallice, T. (1991). *Psychological Review*, 98, 74-95.

Howell, S. R., Jankowicz, D., & Becker, S. (2005). A Model of Grounded Language Acquisition: Sensorimotor Features Improve Lexical and Grammatical Learning, *Journal of Memory and Language, 53(2),* 258-276.

Joanisse, M. F., & Seidenberg, M. S. (1999). Impairments in verb morphology after brain injury: A connectionist model. *Proceedings of the National Academy of Sciences of the United States of America,* 7592-7597.

J. L. McClelland and D. E. Rumelhart. (1981) An interactive activation model of context effects in letter perception. *Psychological Review, 88*,375—407.

Landauer, T. K. & Dumais, S.T. (1997). A solution to Plato's problem: The Latent Semantic Analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211-242.

Miozzo, M. (2003). On the processing of regular and irregular forms of verbs and nouns: evidence from neuropsychology. *Cognition*, 87, 101-127.

McClelland, J. L. & Patterson, K. (2002). Rules or Connections in Past-Tense inflections: What does the evidence rule out? *Trends in Cognitive Sciences.*

Patterson, K., Lambon-Ralph, M.A., Hodges, J. R., & McClelland, J. L. (2001). Deficits in irregular past-tense verb morphology associated with degraded semantic knowledge. *Neuropsychologia, 39,* 709-724.

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

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

Pinker, S., & Ullman, M. (2002). The past and future of the past tense. *Trends in Cognitive Science, 6(11),* 456-463.

Pinker, S. & Ullman, M. T. (2003). Beyond one model per phenomenon. *Trends in cognitive science, 7(3*), 108:109

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tense of English verbs. In J. L. McClelland & D. E. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition,* 216-271. Cambridge: MIT Press.

Seidenberg, M. S., & Joanisse, M. F. (2003). Show us the model. *Trends in Cognitive Sciences, 7(3)*, 106:107

Seidenberg, M. S., & Plaut, D. C. (in press). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science.*

Ullman, M.T., Corkin, S., Coppola, M., Hicock, G., Growdon, J. H., Koroshetz, W. J. & Pinker, S. (1997). A neural dissociation within language: evidence that the mental dictionary is part of declarative memory, and that grammatical rules are processed by the procedural system. *Journal of Cognitive Neuroscience*, 9, 266-276.

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

Woollams, A.M., Joanisse, M., & Patterson K. (2009). Past-tense generation from form versus meaning: Behavioural data and simulation evidence. *Journal of Memory and Language*, 61, 55-76.

Acknowledgements

This work was supported by the Wisconsin Alumni Research Foundation. Thanks to David Plaut for providing the model's semantic representations.