Why do children learn to say “Broke”?

A model of learning the past tense without feedback

Niels A. Taatgen

University of Groningen

John R. Anderson

Carnegie Mellon University

Draft 12/4/00

Abstract

Learning the English past tense is characterized by U-shaped learning in irregular verbs. Existing cognitive models rely on a sudden increases in vocabulary, a high token-frequency of regular verbs, and feedback in order to model this phenomenon. All these assumptions are at odds with empirical data. In this paper a hybrid ACT-R model is presented that shows U-shaped learning without feedback, changes in vocabulary, or unrealistically high rates of regular verbs. It can also explain why there is a distinction between regular and irregular verbs in the first place, by examining the costs and benefits of both types of verbs.

Why do children learn to say “Broke”?

A model of learning the past tense without feedback

Learning the past tense has been the subject of debate in cognitive science since Rumelhart and McClelland first modeled it as part of their PDP-effort (1986). Numerous authors have contributed to the issue since, criticizing the original model (e.g., Pinker & Prince, 1988) or offering alternatives, either connectionist (Plunkett & Marchman, 1991; MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1993) or symbolic (Ling & Marinov, 1993). Although each of these models offers contributions to the debate, they leave some issues unaddressed, and sometimes make assumptions that are not entirely realistic. Fortunately, more empirical data have become available on the topic, mainly through a detailed review of the available data by Marcus, Pinker, Ullman, Hollander, Rosen and Xu (1992).

One of the main topics in learning the past tense is U-shaped learning. Traditionally, three stages are distinguished. In the first stage, when the child starts using past tenses, irregular verbs are used correctly. In the second stage, the child develops a sense for the regularity in regular past tenses. As a consequence, it will now sometimes construct past tenses of irregular verbs in a regular way (e.g., go-goed as opposed to go-went). In the third stage, this overregularization diminishes until performance is without errors. Since performance on irregular verbs is worst in the second stage, the performance-curve has a U-shape, hence the name of the phenomenon. The interesting question is what causes this U-shape.

One account focuses on a dual representation of knowledge: on the one hand past tenses are memorized as separate cases and on the other hand a rule is learned that can

produce regular past tenses (Pinker & Prince, 1988; Marcus et al., 1992). According to the dual-representation explanation, in the first stage only separate cases are memorized. This means that in the first stage producing past tenses is only partially successful, because if a past tense has not been memorized it cannot be reproduced. This changes in the second stage, because at that moment a rule is learned that can produce regular past tenses. The assumption is that this rule is so strong initially that it replaces the strategy of retrieving past tenses from memory, producing overgeneralization. Only when the right balance between the rule and the examples is regained in the third stage, overgeneralization diminishes and performance becomes perfect eventually. The dual-representation explanation leaves a number of questions. How is the regular rule learned? Why isn’t the regular rule used in the right way immediately in stage two, and why does it take so much time before the right balance between examples and the rule is found?

The single-representation explanation only uses a single system, usually a neural network, to explain past-tense learning. In this explanation, U-shaped learning is mainly initiated by changes in the vocabulary size and constitution. When the vocabulary is still small, the network is large enough to accommodate separate cases, but as the vocabulary grows the need for regularization increases. According to the single-representation explanation, U-shaped learning in the networks is initiated by an increase in vocabulary but also by an increase in the proportion of regular words in it. The proportion of regular verbs has to be low initially to prevent generalizations and then large to enable generalizations. The problem with this assumption is that increases in the proportion of regular words (the type frequency) in the vocabulary of children do not correlate with the tendency to overregularize (Marcus, 1995). Moreover, the proportion of regular words

used in actual speech (the token frequency) is reasonably stable around 30%, which is too low for the networks to pick up (Marcus, 1995).

Another important issue that is left unaddressed by the existing models is the question why there is a distinction between irregular and regular verbs in the first place. Why are not all verbs either regular or irregular? Since the existing models all learn by getting feedback on their accomplishments, this question does not have to be asked to get a functioning model. But this leads to another problem, the well-known fact that parents in general do not correct children’s language with respect to syntactical errors (Pinker, 1984). The domain of past tense is no exception to this (Marcus et al., 1992). Perhaps there is something about language usage that encourages the development of regular and irregular verbs which would occur even in the total absence of any feedback. Although children receive no direct feedback on their own production, they do receive input from the environment: they hear and interpret past tenses produced by others. Interpreting a past tense requires a different process than generating one: instead of producing a past tense given the stem, now the stem has to be produced given the past tense.

In this paper we will address a number of this issues by presenting a model using the ACT-R cognitive architecture (Anderson & Lebiere, 1998). The model mimics a real learning situation, where the model receives no feedback on the past tenses it produces itself, but where the environment supplies examples of correct past tenses. We will also look at what happens if we deprive the model of all its input, and force it to make up its own past tenses. Before discussing the models, we will first introduce the relevant aspects of the ACT-R cognitive architecture.

The ACT-R architecture

The basic theoretical foundation of the ACT-R (Adaptive Control of Thought, Rational) architecture is rational analysis (Anderson, 1990). According to rational analysis, each component of the cognitive system is optimized with respect to demands from the environment, given its computational limitations. The main components in ACT-R are a declarative (fact) memory and a production (rule) memory. ACT-R is a so-called hybrid architecture, in the sense that it has both symbolic and sub-symbolic aspects. We will introduce these components informally. Table 1 provides a formal specification of some critical aspects of the subsymbolic level. Further details about the architecture can be found in Anderson and Lebiere (1998).

Items in declarative memory, called chunks, have different levels of activation to reflect their use: chunks that have been used recently or chunks that are used very often receive a high activation. This activation decays over time if the chunk is not used. Activation represents the probability (actually, the log odds) that a chunk is needed and the estimates provided for by ACT-R’s learning equations represent the probabilities in the environment very well (see Anderson, 1993, chapter 4, for examples). The level of activation has a number of effects. One effect of activation is that when ACT-R can choose between chunks, it will retrieve the chunk with the highest activation. Activation also affects retrieval time. As the activation of a chunk decreases, its retrieval time grows exponentially. At some point it is no longer feasible to retrieve a chunk: it would just take too much time. Because of this ACT-R is not able to retrieve chunks with an activation below a certain threshold.

Chunks cannot act by themselves, they need production rules for their application.

In order to use a chunk, a rule has to be invoked that retrieves it from declarative memory and does something with it. Since ACT-R is a goal-driven theory, chunks are always retrieved to achieve some sort of goal. In the context of learning past tense the goal is simple: given the stem of a word, produce the past tense. One strategy to produce a past tense is to just retrieve it from memory, using a rule like:

```
IF    the goal is to produce a past tense of a word
      AND there is a chunk that specifies the past tense of that word
THEN  the set the answer of the goal to the past tense
```

If the goal is to produce a past tense of a certain word, this rule will attempt to retrieve a chunk from declarative memory that specifies what the past tense is. Of course this rule will only be successful if such a fact is present.

The behavior of rules is also governed by the principle of rational analysis. Each rule has a real-value quantity associated with its expected outcome. This expected outcome is calculated from estimates of the cost (in time) and probability of reaching the goal if that rule is chosen. The unit of cost in ACT-R is time. ACT-R’s learning mechanisms constantly update these estimates based on experience. If multiple rules are applicable for a certain goal, the rule is selected with the highest expected outcome.

In both declarative and procedural knowledge, selections are made on the basis of some evaluation, either activation or expected outcome. This selection process is noisy, so the item with the highest value has the greatest probability of being selected, but other items get opportunities as well. This may produce errors or suboptimal behavior, but also allows the system to explore knowledge and strategies that are still evolving.

In addition to the learning mechanisms that update activation and expected outcome, ACT-R can also learn new chunks and rules. New chunks are learned automatically: each time a goal is completed it is added to declarative memory. If an

identical chunk is already present in memory, both chunks are merged and their activation values are combined. New rules are learned on the basis of specializing and merging existing rules. The circumstance for learning a new production rule is that two rules fire one after another with the first rule retrieving a chunk from memory. A new rule is formed that combines the two into a macro-rule but eliminates the retrieval. The macro-rule is specialized to contain that information that was retrieved. This process is reasonably safe, in the sense that it never produces rules that are completely different from the rules already present, but can nevertheless produce radical changes in behavior if the new rule outperforms the old rule[1].

A rational account of regular and irregular past tense

Why is there a distinction between regular and irregular verbs in the first place? Both the regular and the irregular forms of past tense represent possible ways of modifying a present tense verb to mark its tense. The regular rule adds an extra morpheme to the stem, and in the case of irregular verbs the stem itself is changed (except for cases like hit-hit, where the past tense is identical to the present tense). Using a regular rule seems a very economical solution from the viewpoint of memory: only one rule is needed to produce past tenses for all verbs. Since there is no systematic way in which the stem itself can be changed, irregular verbs have to be memorized separately. So what is the advantage of using an irregular past tense? Why not have a single rule for the past tense and be done with it? There are several possible reasons. First, a regular past tense is usually slightly longer than an irregular past tense. A regular rule always adds a morpheme to the stem that sometimes has to be pronounced as a separate syllable. A second reason why

irregulars may have an advantage is that they are actually more regular from a phonetic viewpoint (Burzio, in press?). For example, *keeped is phonetically irregular (in English) as opposed to kept. This phonetic disadvantage suggest that the use of a regular rule requires some phonetic postprocessing that makes it less attractive than just storing and retrieving an irregular form.

A third alternative is to use no past tense at all and just use the present tense. Although this may be the cheapest strategy from a memory perspective, it has the risk of miscommunication, which also carries its costs.

In this micro-economy of knowledge, the optimal choice of strategy depends on the frequency a word is used. High-frequency words benefit more from the irregular strategy, because the cases memorized turn up quite often. For low-frequency words the use of a rule is more optimal, since maintaining a case in memory for the few occasions the word is used does not overcomes the disadvantage of using a rule. In ACT-R, this trade-off is already built into the basic mechanisms of the architecture. Due to activation learning, low-frequency knowledge receives lower activation than high-frequency knowledge. This activation difference translates into retrieval time and success: low frequency items take more time to retrieve or cannot be retrieved at all.

Figure 1 illustrates that this is the case: the 478 verbs (89 irregular, 389 regular) that children or their parents use reported in Marcus et al. (1992) are sorted with respect to their frequency according to Francis and Kucera (1982). The curve shows the number of occurrences in the Francis and Kucera corpus, while a bar indicates an irregular verb. As can be seen in the graph, most irregular verbs are high frequency words: the first regular verb is no. 13 (use). According to this distribution, only 25% of the words used (the token-

frequency) are regular, which is close to the 30% Marcus et al. (1992) found in children’s speech.

In summary, the rational-analysis theory of ACT-R predicts that even in a situation where the cognitive system can choose between maintaining distinct past-tense forms (irregular past tenses) and adding a suffix to a word (regular past tenses), it will end up with high-frequency irregular verbs and low-frequency regular verbs. This will be the basis for the model. An indirect form of feedback is added to this basis: although no feedback is given on the behavior of the model, it does perceive correct forms in the environment.

A model of learning the past tense

The model initially has to choose between the following rules to construct a past tense given the stem of the verb:

- Attempt to retrieve the past tense from memory
- Attempt to generate a new past tense by analogy: retrieve an arbitrary past tense from memory and use it as a template to find a past tense for the current word
- Just use the stem as past tense

None of these strategies is very good initially. Analogy involves more than one reasoning step, does not have any examples initially, and is not always successful. The retrieve strategy also needs examples before it can be successful. Just using the stem as a past tense works all the time, but does not produce a past tense that can be distinguished from the present tense. Before the model can do anything useful beyond producing a past tense that is identical to the stem, it has to perceive some examples in the environment. Note

that there is no regular rule yet, ACT-R will learn it later on as a specialization of the analogy strategy. These initial strategies are similar to those proposed by MacWhinney (1978), who also suggested that the regular rule is formed on the basis of analogy.

The input for the model consists of the 478 words from Marcus et al. (1992). For each trial, a word is randomly selected from the set based on the frequency distribution in figure 1. This word is presented to the model with the goal to find the past tense. After the model has come up with a past tense, 2000 simulated seconds pass, after which the next word is presented. The model receives no feedback on the accuracy of the answer. Although the model gets no external feedback, it can update the expected outcomes of its rules based on internal feedback, caused by different execution times of the different strategies. Besides internal feedback, the model perceives past tenses in the environment. For every past tense the model produces, it perceives two correct past tenses in the environment. Also, the model perceives 2000 past tenses initially.

Both the past tenses it perceives and produces itself are added to ACT-R’s declarative memory, resulting in a growing library of past-tense examples. Not all of these examples will be available, however, due to activation decay with time. Furthermore, not examples are necessary correct, since incorrect forms produced by the model itself are also maintained.

As described in the Appendix, analogy applied to the English past tense system results in the learning of two new rules for past tense. One of these is a rule that just uses the stem. This is identical to the "use stem as past tense" rule and plays no significant role in the behavior of the system. The other learned is the rule that uses the "ed" inflection and this is the one that will produce the overgeneralizations. If we had a richer phonological

system we might also learn rules for the few other semi-regular vowel-change patterns in English like "ring-rang". However, they are such a smaller part of the English past tense system their omission does not change our ability to capture the basic U-shaped learning curve.

The analogy strategy is not very successful in general, because it retrieves the verb with the highest activation as an example. These verbs are irregular most of the time, so only when a regular example is retrieved will the analogy strategy produce a regular past tense. It will therefore take some time to learn the regular rule. The regular rule itself is very successful: it will always produce a past tense given a stem. Only retrieving a highly activated past tense is more efficient. Once it is learned, it takes some more time before it is used frequently, because its expected outcome still has to grow.

Figure 2 shows how the expected outcomes of the different strategies develop in the first 20 months of the simulation. Note that rule selection is a noisy process: the rule with the highest evaluation only has the highest probability of being selected. Furthermore, if a certain rule fails (e.g., the retrieval rule cannot retrieve anything) the next best rule is selected (e.g., the regular rule). In each month approximately 1300 past tenses are produced. Initially the model will mainly use the stem as a past tense. As it gains experience in storing past tenses from the environment and producing them itself, however, it will be able to retrieve previous past tenses from memory using the retrieve rule. The expected outcome of the retrieval rule increases gradually over time as the model learns more past tenses.

Around the fourth month of the simulation, the model learns its first new rule, the rule that use the stem as the past tense. As this rule duplicates the behavior of an already

existing rule, the “use stem”-rule, it does not play an important role in behavior. Between the sixth and seventh month the model learns the regular rule. Its expected outcome initially increase rapidly, because it is able to produce a past tense out of any stem, as opposed to the analogy strategy that it originated from. Soon the basic behavior of the model is to first try to retrieve an irregular past tense from memory, and if this fails to create a regular past tense by adding a suffix. The reason that retrieval is tried first is that it is phonetically more efficient than the rule for the reasons noted early. Note that it need not be the case that all irregulars are phonetically more efficient. Rather, retrieval has to be on average more efficient that applying the rule. It is also worth noting that frequent regular can be generated by retrieval as well as by the regular rule.

Figure 3 plots how often the model chooses an irregular form for an irregular verb as opposed to a regular form. The plot also shows how overregularization is usually presented in the literature: 100% means perfect performance, anything beneath it is overregularization. The results show U-shaped learning. The downward slope coincides with the learning of the regular rule. At this point in the simulation the model has not memorized all irregular past tenses yet at a level that they can be retrieved without errors. If it fails to retrieve an irregular past tense it will use one of the regular rules, producing overregularization. The regular rules may also win the competition with the retrieve rule because of the noise, so the model will not even try to retrieve an irregular past tense. A third source over overgeneralization occurs if the retrieve rule retrieves a previous overgeneralization from memory. Gradually the model will master the irregular past tense, producing the upwards slope in the U-curve.

Input from the environment is integrated with the examples the model has produced

itself. When past tenses from the outside world concur with past tenses produced by the system itself, they strengthen each other because identical chunks are collapsed in declarative memory. If we assume forms from the outside world are always correct and forms produced by the model are occasionally correct, incorrect past tenses in declarative memory eventually lose the competition.

The predicitions that model produces can be compared to empirical data from Marcus et al. (1992). From the children that they studied most extensively, Adam and Sarah show the pattern associated with U-shaped learning, that is, they show a reliable period without overregularization followed by a period with overregularization. A third child, Abe, shows extensive overregularization over the whole period that he is studied. He shows no signs of U-shaped learning, presumably because his overgeneralization had already started at the beginning of the study. He is of particular interest because Marcus et al. report on his behavior on individual words. We will look at how the model handles individual words later on.

Figure 4 show the overregularization rates of Adam and Sarah. Both the children and the model show the initial stages of U-shaped learning, from no overregularization (first stage) to overregularization (second stage). Although overregularization in the model gradually diminishes, Adam and Sarah do not shows any signs of diminished overregularization during the period studied. Nevertheless Marcus (1996) reports that overregularization gradually trails of in children, 4.2% in preschoolers, 2.5% in first graders and 1% in fourth graders. Even adults sometimes overregularize, but this is very rare. In both Adam and Sarah overgeneralization seems to increase more gradually than in the model. A possible explanation for this is the fact that the model is tested on the full

vocabulary, even words it has not encountered very often yet. An irregular verb that the model has not yet encountered will almost always lead to an overgeneralization. Children on the other hand will presumably tend to avoid words they do not know well. As their vocabulary grows they will tend to use less frequent irregulars more often, increasing overgeneralization.

In the children, the best predictor for the onset of overregularization is a sudden increase in the rate in which regular verbs are actually marked for past tense, as is indicated by the dotted lines in figure 4 (the sudden spike between months 27 and 32 in Sarah’s graph is due to the small number of observations.) This increase indicates the discovery of the regular rule. Figure 3 shows that this is true in the model as well: again the dotted line indicates the rate in which regular verbs are marked for past tense.

Another aspect of children learning the past tense is individual differences. Adam exhibits very little overregularization, while Sarah shows much more tendency to overregularize. A possible explanation can be found in the input from the environment. In the current model the two examples are perceived in the environment for every example produced. If fewer examples are perceived, because the environment supplies fewer or the child pays less attention to them, overregularization will increase. Figure 5 shows overregularization for different ratios of examples perceived and produced.

Although the vocabulary that serves as input for the model is fixed, the model itself only acquires these words over time. A good estimate of whether or not a certain word is part of the vocabulary is to look at its activation. If this activation is past a certain threshold[2], the word is assumed to be part of the vocabulary of the model. Figure 6 shows the result of the model, together with data from Adam and Sarah.

Marcus et al. (1992) also report behavior on individual words by Abe. The left side of figure 7 shows the four examples, the right side similar words from the model. For the word Say Abe shows very little overgeneralization. This turns out to be true for the model as well, and can be explained by the fact that Say is a high-frequency irregular verb (no. 4 in the word-frequency list). The second example, Eat for Abe and Spend for the model, is somewhat less frequent (spend is no. 79). Abe has some early problems with this word, but recovers later on. The model also shows recovery after early problems with Spend. For irregular words that Abe uses very little, his behavior is rather erratic, as exemplified by the words Draw and Win. Since draw and win are relatively high on the word-frequency list (no.’s 73 and 94), two other words that do have a low frequency for the model were selected in comparison: Eat and Spin (no’s. 114 and 247).

In general, overregularization is most common for infrequent words. In the model this can be explained by the fact that low-frequency irregular forms have a lower activation, increasing the probability of a retrieval failure and subsequent regularization. Marcus (1996) reports that this indeed the case: he found a correlation of -.34 between the frequency that parents use an irregular verb and the children’s overregularization rate. This correlation is -.21 for the model (using the Francis & Kucera frequencies for the parental frequencies). This correlation seems to be very low, but is mainly caused by the non-linearity in the frequency of words (as can be seen in figure 1). For example, the removal of to be from the data set, a verb with a very high frequency and almost no overregularization (so perfectly compliant with the correlation) nevertheless boosts it to -.34. Applying a log-transformation to the frequencies increases the correlation to -.83.

A model that without any input

An interesting variation on the model presented here is to deprive it of all input, in order to see what happens in a minimal system. Since the model has no access at all to correct past tenses, it will have to invent them itself. As a consequence, it not only needs strategies that are usually associated with past-tense learning, but also some rules that make up new past tenses. To the rules already present in the system the following rule is added:

> Generate a new past tense. Most of the time, this strategy will modify the stem into something new (corresponding to an irregular past tense), but sometimes adds a random suffix to the stem (corresponding to a regular past tense).

Generating something new is expensive, so this strategy will have a high cost associated with. The model will therefore prefer the other strategies as soon as they are reasonably productive.

Figure 8 shows the proportion of times the model uses an irregular or a regular past tense for a certain word. It clearly prefers irregular forms for high-frequency words and regular forms for low-frequency words, as predicted by the theory. Due to the probabilistic nature of both the model and the selection of words from the vocabulary, there are many spikes in the graph, indicating infrequent words for which the model nevertheless prefers an irregular past tense. Note that this particular model has no incentive to use the same past tense all the time, so it can use different forms of past tense without penalty. This also explains why the model can have several different regular past tenses.

Although this model receives no feedback at all, it is still interesting to look at the tendency of the model to use irregular forms for verbs that are irregular in reality.

Although the model has no way of knowing what words are regular or irregular, irregular words generally have a high frequency, so the model will eventually favor irregular past tenses for these words. Figure 9 plots how often the model chooses an irregular form for an irregular verb as opposed to a regular form. The plot also shows how overregularization is usually presented in the literature: 100% means perfect performance, anything beneath it is overregularization. Interestingly enough, the model shows U-shaped learning. The downward slope coincides with the learning of regular rules. At this point in the simulation the model has not memorized all irregular past tenses yet at a level that they can be retrieved without errors. If it fails to retrieve an irregular past tense it will use one of the regular rules, producing overregularization. One of the regular rules may also win the competition with the retrieve rules because of the noise, so the model will not even try to retrieve an irregular past tense. Gradually the model will master the irregular past tense, producing the upwards slope in the U-curve. It will never reach 100% though, since it has not generated irregular past tenses for all the verbs, especially not for the low-frequency irregular verbs.

The no-feedback model is an oversimplification of reality. It is nevertheless interesting, because it shows that without feedback, without any change in vocabulary, and without an empirically incorrect high token-frequency of regular words, it can learn to inflect high-frequency words irregular and low-frequency words regular. Moreover, its U-shaped learning is caused by a model that only tries to learn rules and memorize examples at the same time.

Conclusions

The basic hypothesis underlying the models presented in this paper is that both regular and irregular forms each have their own advantages. Irregular forms are more effective as long as their use is frequent enough, while regular forms always work at a slightly higher cost. This hypothesis alone can explain many of the data associated with the learning of past tenses, without the need for further assumptions about the ratio between regulars and irregulars, feedback and the growth of the vocabulary. Instead the model is capable of generating the growth of vocabulary instead of prerequiring it. The models discussed in this paper do not model words at the level of phonetics. We believe that for this model, its simplicity is a virtue. It shows that the basic phenomena of learning irregular and regular past tenses can be explained by the small set of principles on which the model is based. It is however interesting to observe that words ending in a -d or -t are more often irregular (35% of the irregulars end with a -d or -t, as opposed to 11% of the regulars). This can be explained by the fact that for these words the additional costs of regularization are higher than other words, since the -ed has to be pronounced as a separate syllable instead of just an added phoneme.

The views expressed in this model largely coincide with the dual representation account as offered by Pinker and Marcus. It can fill in some of the gaps and problems of this account. It shows how the regular rule can be learned by analogy. It offers an explanation for why it takes so long to find the right balance between examples and the regular rule. The main reason is the lack of feedback. But another reason is the fact that the regular rule serves in important function at stage in learning where the child has only memorized a few of the irregular past tenses. A third reason is the fact that children learn their own errors (see also ?? & MacWhinney, ??). Correct irregular past tenses have to

compete in memory with overregularizations. This also offers an explanation for the phenomenon that correcting a child doesn’t seem to help, exemplified by the following exchange reported by Cazden (1972).

Child: My teacher holded the baby rabbits and we patted them.
Adult: Did you say your teacher held the baby rabbits?
Child: Yes.
Adult: What did you say she did?
Child: She holded the baby rabbits and we patted them.
Adult: Did you say she held them tightly?
Child: No, she holded them loosely.

In this exchange the number of past tenses produced by the child equals the perceived examples from the environment, so both the wrong and the correct form are strengthened equally.

The penalty for overregularization is very small: it is slightly less efficient than using the proper irregular form and it takes time to fully exploit this optimization. Pinker (1999) offers a slightly different view on irregular words, however. According to his account, irregular words stem from earlier versions of the language and survive because they have a high frequency, as opposed to low-frequency words that are regularized. As Pinker acknowledges, this cannot explain why new irregular verbs enter a language (e.g., dive-dove, catch-caught). If irregular rules would be a historical matter entirely, all irregular verbs would gradually disappear.

In general people seem to strive for short-cuts in language as long as this does not lead to communication problems. The first thing a novice in a new organization has to learn is all the acronyms his or her new colleagues use to refer to the various departments and services in the company (e.g., the Groningen University uses “DOOP” to refer to the department of research, education and planning). Within the company all these acronyms

are high-frequency words. The novice, however, will initially suffer from a version of overgeneralization by using the full title of the department as opposed to the acronym. Another recent abbreviation is using gonna for going to and wanna for want to. According to Boyland (1998) [other reference here!], the frequency of use of the verb want has increased in the past few centuries, especially in conjunction with to. The increased frequency of the combination has made memorizing the abbreviation a viable solution.

The hypothesis that U-shaped learning is initiated by large increases in vocabulary and/or the ratio between regular and irregular verbs as forwarded by various existing neural network models is not supported by the data. These models are only able to learn regularization if the majority of the past tenses is regular. So what is the big difference between the various neural networks and ACT-R? Learning in declarative memory has many similarities to learning in neural networks, and can even be implemented by one (Anderson & Lebiere, 1998, chapter 12). The main difference is that ACT-R singles out some particularly useful pattern and casts it into a production rule. Since this rule is rather successful in the system, it is protected from the catastrophic forgetting that neural networks tend to suffer from and that bans them from learning the regular rule if the proportion of regular past tenses is not high enough.

Availability of the models

The models are available on the internet by following the “published models” link on the ACT-R webpage: http://act.psy.cmu.edu/

Appendix

Apart from the initial strategies, ACT-R will learn a number of new rules due to proceduralization. The main new rule, the rule for regular past tenses, is learned on the basis of the analogy strategy. Analogy retrieves an arbitrary past tense from memory, and tries to establish a mapping between the word and its past tense. It then applies this mapping to the current word. The three types of patterns analogy can find are:

past tense of WORD1 = WORD1 + some suffix (regular)

past tense of WORD1 = WORD2 (irregular)

In the case of a regular verb, the mapping is clear: a suffix has to be added to the stem. In the case of an irregular word, the pattern is harder to establish, since the past tense is usually similar to the present tense, but not in a systematic way. The system does not model the phonetic level in detail, so it cannot establish a mapping in the case of irregular verbs. If a more extended representation were used, the analogy strategy would also be able to make analogies based on irregular forms, producing so-called blends like bring-*brang. Blends are rather rare [reference], so we saw no need to add the detail necessary to model them.

An example of a regular form that may serve as a basis for analogy is the word walk, represented in ACT-R’s declarative memory as:

```
PAST-TENSE2323
      ISA PAST
      OF WALK
      STEM WALK
      SUFFIX ED
```

This chunk represents that the past tense of walk is walk + ed. Now suppose the past tense of the word work is needed, represented by the following goal:

```
PAST-TENSE2638
      ISA PAST
      OF WORK
      STEM NIL
```

```
SUFFIX NIL
```

Note that goals have the same form as facts in declarative memory, except that they still have some empty slots (denoted by nil). In order to generate the past tense for work, we might retrieve walk from memory and use this as a template to generate a past tense for work. Examination of the past tense of walk reveals that the contents of the stem and of slots are identical and that the contents of the suffix slot equals ed.

Two rules implement the analogy strategy. The strategy used is not very sophisticated: it is basically a simple pattern matcher. A first rule looks at the target (e.g., work) and focuses on an empty slot, in this case the suffix slot. It then looks at the example to see what value is in this slot and copies this value to the target.

```
RULE ANALOGY-FILL-SLOT
IF    the target has a slot that is empty
      AND   in the example the slot is filled with some value that does not occur anywhere
            else in the example
THEN  set the slot in the target to the value in the example
```

The second rule notices slots that have the same value in the example, and makes these equal in the target.

```
RULE ANALOGY-COPY-A-SLOT
IF    the target has a slot that is empty and another slot that already has a value
      AND   in the example both these slots have the same value
THEN  set the empty slot in the target to the value of the slot that has a value
```

In the case of a regular past tense this rule can be used to copy the value in the of slot to the stem slot.

Combining these two rules while substituting a regular past tense like walk as an example produces a rule that derives a regular past tense without the need for an example:

```
RULE LEARNED-REGULAR-RULE
IF    the target is a past tense of a word and slots stem and suffix are empty
THEN  set the suffix slot to ED
      and set the stem slot to the word of which you want the past tense
```

Once the retrieval of the example is specialized into the regular rule, the process will always succeed in producing a past tense and be able to successfully compete with the other strategies.

A second rule that will also be learned is based on examples where the past tense equals the stem, producing the rule:

```
RULE LEARNED-EMPTY-SUFFIX-RULE
IF    the target is a past tense of a word and slots stem and suffix are empty
THEN  set the suffix slot to BLANK
      and set the stem slot to the word of which you want the past tense
```

This rule is identical to the “use the stem as past tense” rule already present in the model.

References

Anderson, J. R. (1990). The adaptive character of thought. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R. (1993). Rules of the Mind. Hillsdale, NJ: Lawrence Erlbaum.

Anderson, J. R., & Lebiere, C. (1998). The atomic components of thought. Mahwah, NJ: Erlbaum.

Boyland, J. T. (1998). How developing perception and production contribute to a theory of language change: Morphologization < Expertise+Listening < Development. Proceedings of the Thirty-fourth Meeting of the Chicago Linguistic Society. Chicago, IL: University of Chicago.

Francis, W. N., & Kucera, H. (1982). Frequency analysis of english usage: lexicon and grammar. Boston, MA: Houghton Mifflin.

Ling, C. X., & Marinov, M. (1993). Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs. Cognition, 49(3), 235-290.

MacWhinney, B., & Leinbach, J. (1991). Implementations are not conceptualizations: revising the verb learning model. Cognition, 40, 121-157.

Marcus, G. F. (1995). The acquisition of the English past tense in children and multilayered connectionist networks. Cognition, 56, 271-279.

Marcus, G. F. (1996). Why do children say “Breaked”? Current directions in psychological science, 5, 81-85.

Marcus, G. F., Pinker, S., Ullman, M., Hollander, M., Rosen, T. J., & Xu, F. (1992). Overregularization in language acquisition. Monographs of the society for research in child development, 57(4), 1-182.

Pinker, S. (1984). Language learnability and language development. Cambridge, MA: Harvard University Press.

Pinker, S. (1999). Words and rules: the ingredients of language. New York: Basic books.

Pinker, S., & Prince, A. (1988). On language and connectionism: Analysis of a distributed processing model of language acquisition. Cognition, 28, 73-193.

Plunkett, K., & Marchman, V. (1991). U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. Cognition, 38, 43-102.

Plunkett, K., & Marchman, V. (1993). From rote learning to system building: acquiring verb morphology in children and connectionist nets. Cognition, 48, 21-69.

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 (pp. 216-271). Cambridge, MA: MIT Press.

Author Note

Niels A. Taatgen, Department of Cognitive Science and Engineering and John R. Anderson, Department of Psychology.

This research was supported by a NATO-Science Fellowship awarded to Niels Taatgen by the Netherlands Organization for Scientific Research (NWO) and by ONR grant N0014-96-I-0491.

The authors would like to thank Brian MacWhinney, Jay McClelland, Gary Marcus and Steven Pinker for their comments on an earlier draft of the paper.

Correspondence concerning this article should be addressed to Niels A. Taatgen, Department of Cognitive Science and Engineering, University of Groningen, Grote Kruisstraat 2 / 1, 9712 TS Groningen, Netherlands. Electronic mail may be sent via Internet to niels@tcw2.ppsw.rug.nl

Footnotes

[1]The process of proceduralization used in this model is not part of ACT-R 4.0, the current newest version of ACT-R, but is part of a proposal for the next version of the architecture.

[2]The actual threshold is -3.5 for the base-level activation. Together with activation from the context, this is generally sufficient to retrieve the word.

Table 1

ACT-R equations. These equations are simplified versions of the original Anderson & Lebiere (1998) equations.

Equation	Description
Activation $A = B + \text{context} + \text{noise}$	The activation of a chunk has three parts: base-level activation, spreading activation from the current context and noise. Since spreading activation is a constant factor in the models discussed, we treat activation as if it were just base-level activation.
Base-level activation $B(t) = \log \sum_{j=1}^{n} (t - t_j)^{-d}$	n is the number of times a chunk has been retrieved from memory, and t_j represents the time at which each of these retrievals took place. So, the longer ago a retrieval was, the less it contributes to the activation. d is a fixed ACT-R parameter that represents the decay of base-level activation in declarative memory.
Retrieval time $\text{Time} = Fe^{-A}$	Activation determines the time required to retrieve a chunk. A is the activation of the chunk that has to be retrieved, and F is a fixed ACT-R parameter. Retrieval will only succeed as long as the activation is larger than retrieval threshold tau, which is also a fixed parameter
Expected Outcome $\text{Expected outcome} = P_p G - C_p + \text{noise}$	Expected outcome is based on three quantities, the estimated probability of success of a rule (P), the estimated cost of the rule (C), and the value of the goal (G)

Figure Captions

Figure 1. Frequencies of 478 words used by children or their parents from Marcus et al. (1992) according to the Francis and Kucera (1982) corpus. The curve denotes the number of occurrences in the corpus while the bars indicate irregular verbs.

Figure 2. Expected outcomes of the different strategies for the model

Figure 3. Overregularization and proportion of regular past tenses marked by the indirect-feedback model

Figure 4. Overregularization and proportion of regular past tenses marked by Adam and Sarah. Adapted from Marcus et al. (1992).

Figure 5. Overregularization for the model for different ratios of input from the environment and production.

Figure 6. Vocabulary growth for Adam, Sarah (adapted from Marcus et al., 1992) and the model.

Figure 7. Overregularization of individual words for Abe (adapted from Marcus et al., 1992) and the model.

Figure 8. Proportion of times the model chooses an irregular or a regular past tense for a certain word. Words on the x-axis are sorted by frequency as in figure 1.

Figure 9. Proportion of times the no-feedback model produces an irregular past tense for an irregular verb as opposed to producing a regular past tense.

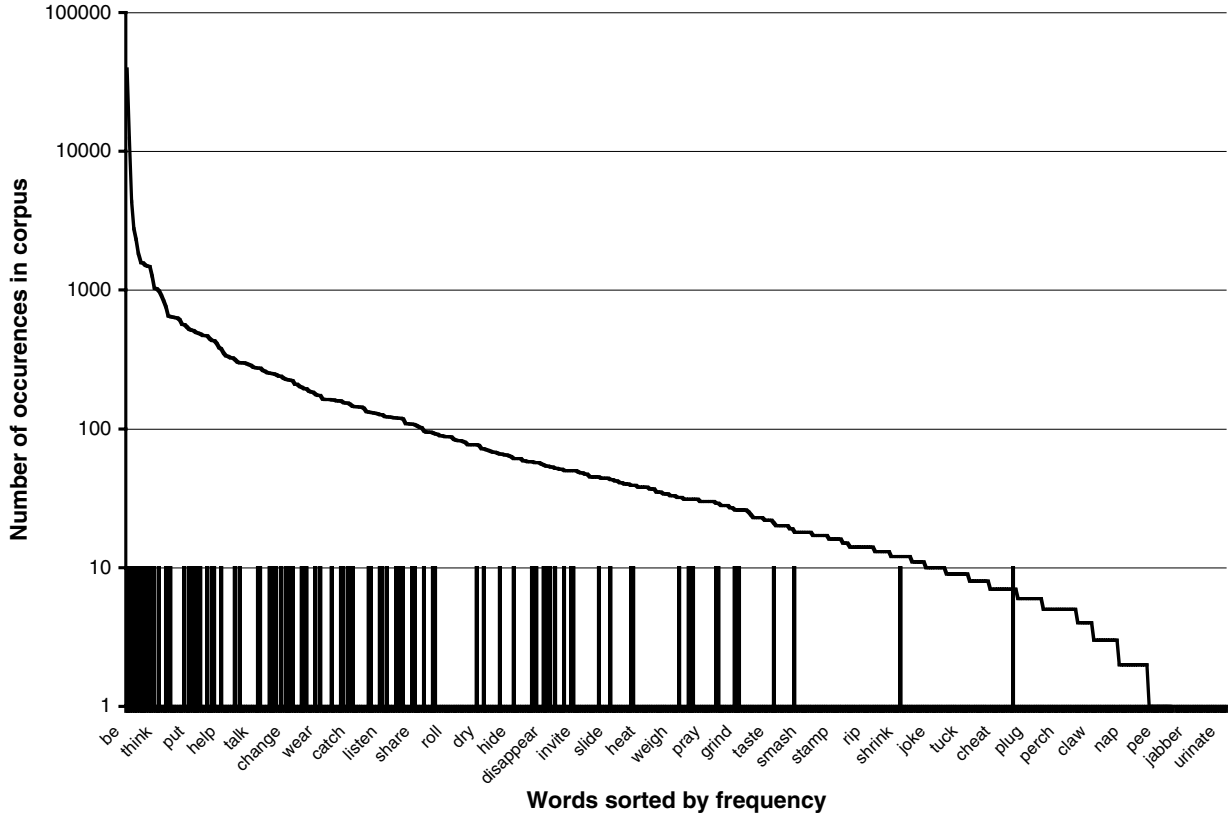


Figure 1. Frequencies of 478 words used by children or their parents from Marcus et al. (1992) according to the Francis and Kucera (1982) corpus. The curve denotes the number of occurrences in the corpus while the bars indicate irregular verbs.

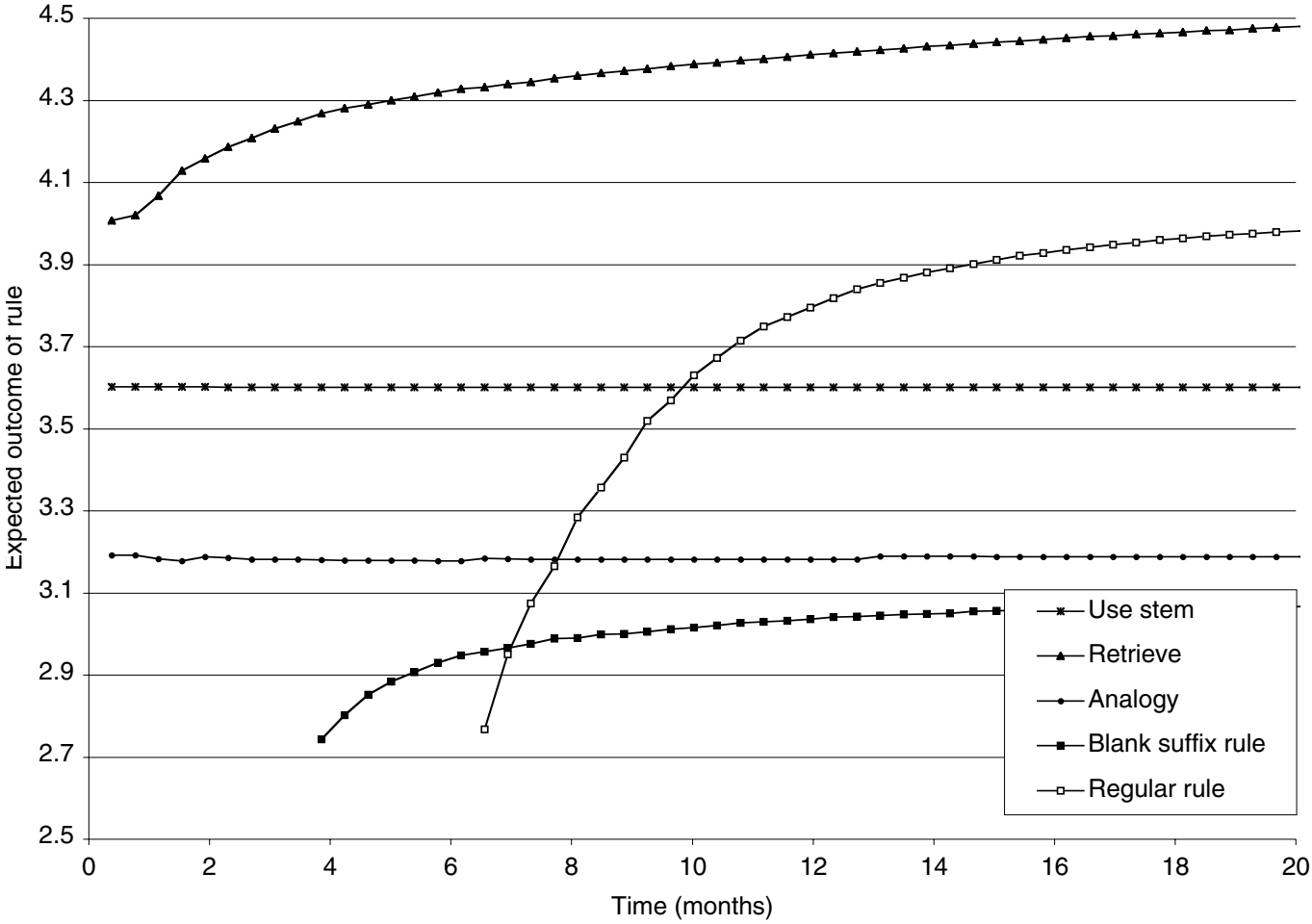


Figure 2. Expected outcomes of the different strategies for the model

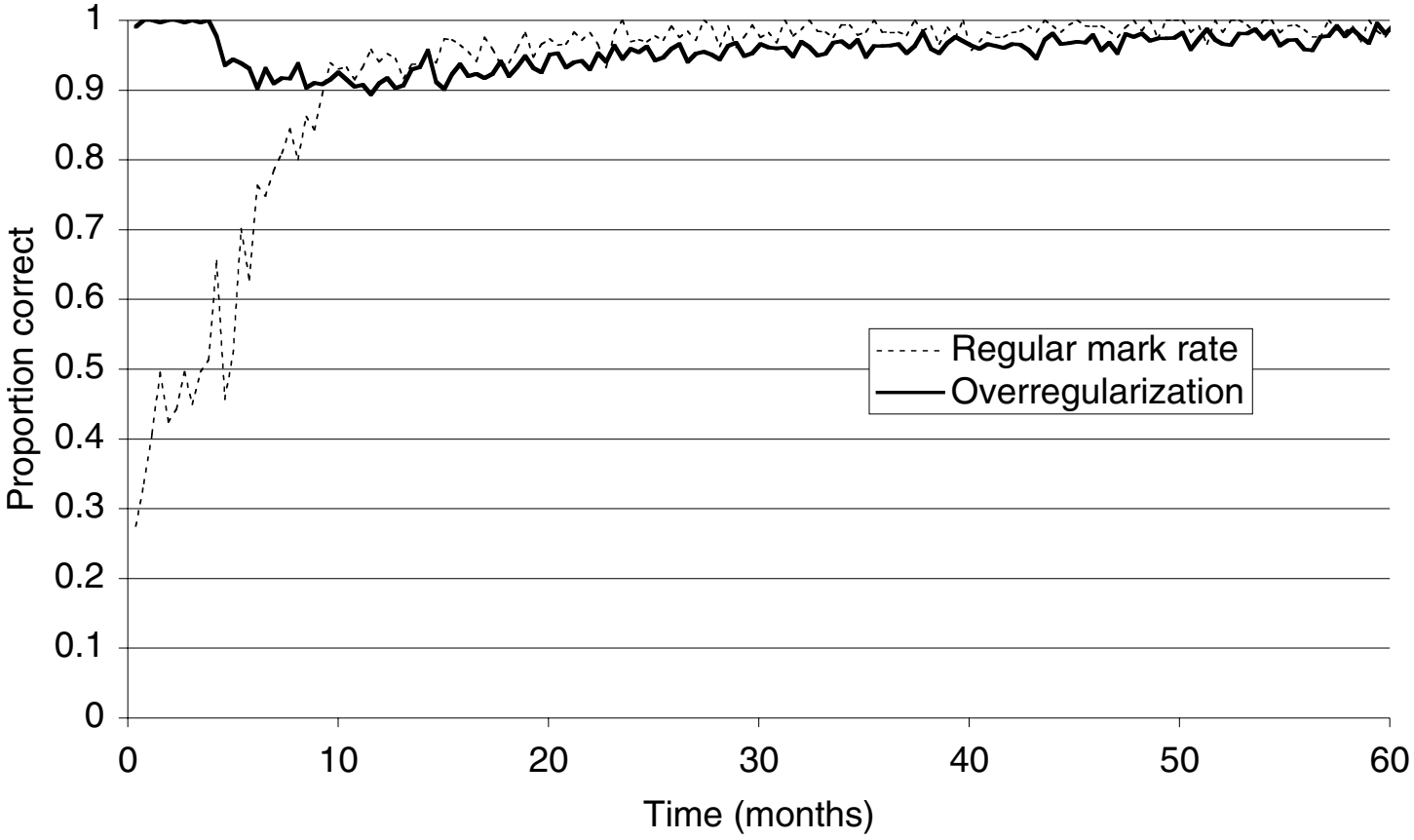


Figure 3. Overregularization and proportion of regular past tenses marked by the indirect-feedback model

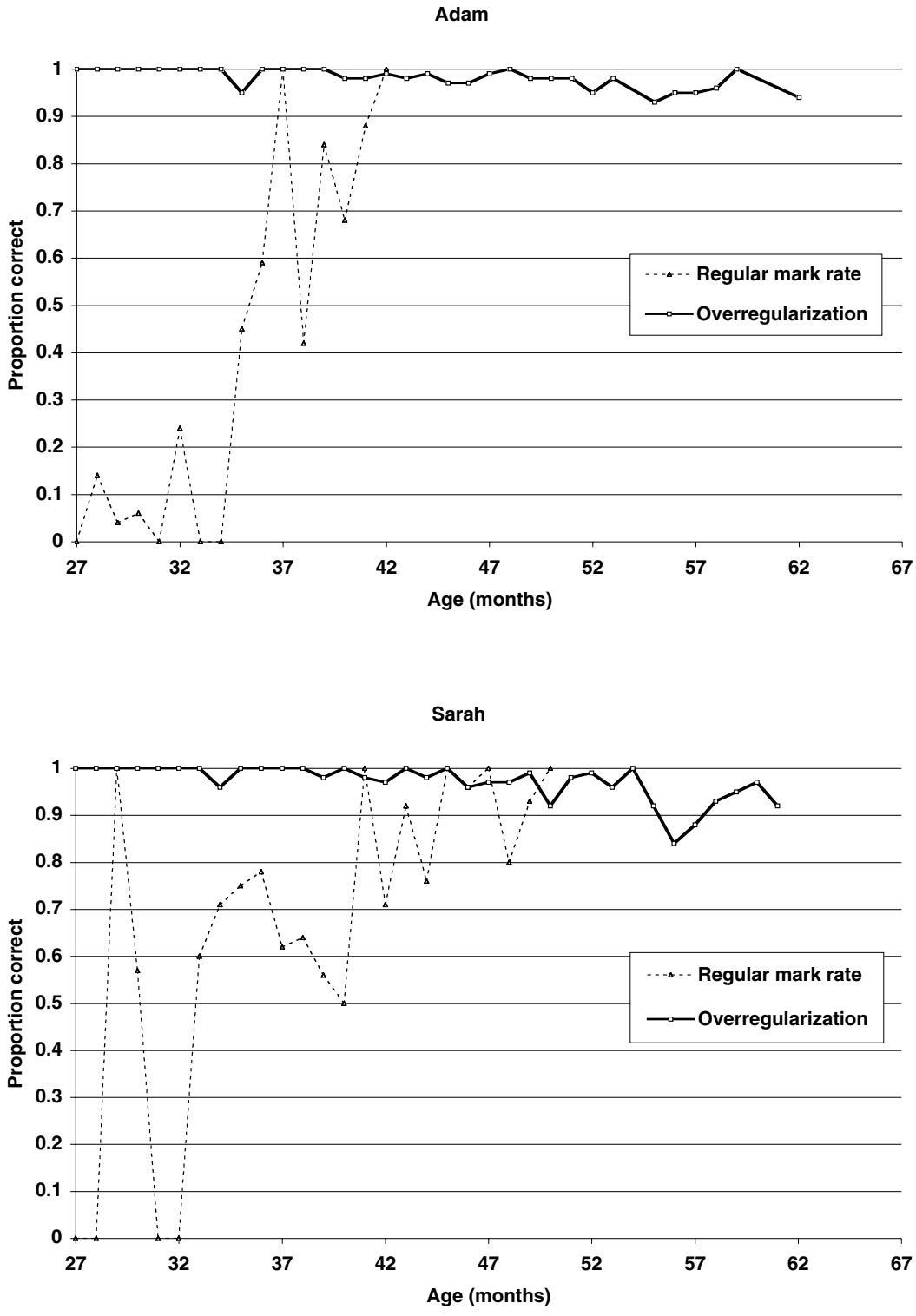


Figure 4. Overregularization and proportion of regular past tenses marked by Adam and Sarah. dapted from Marcus et al. (1992).

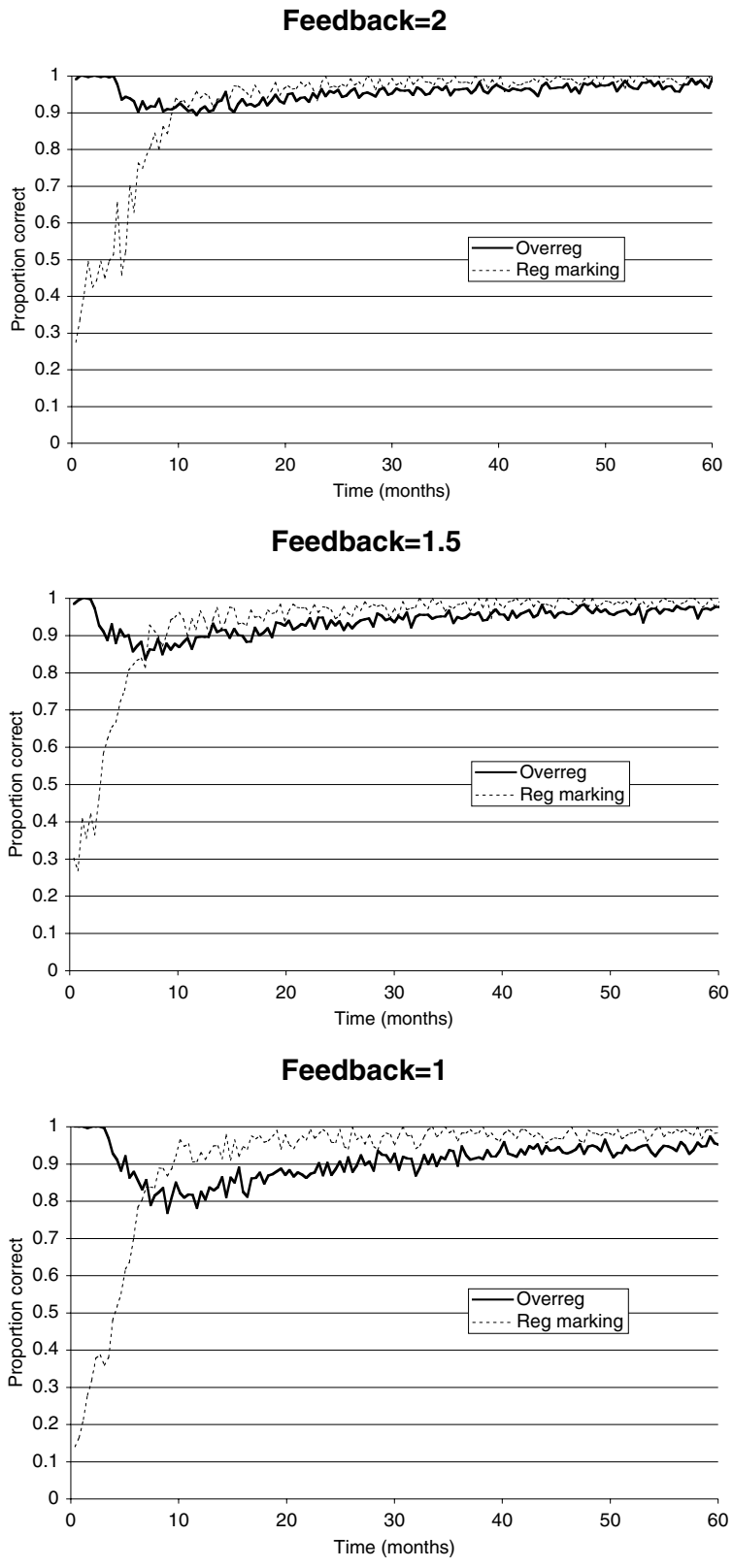


Figure 5. Overregularization for the model for different ratios of input from the environment and production.

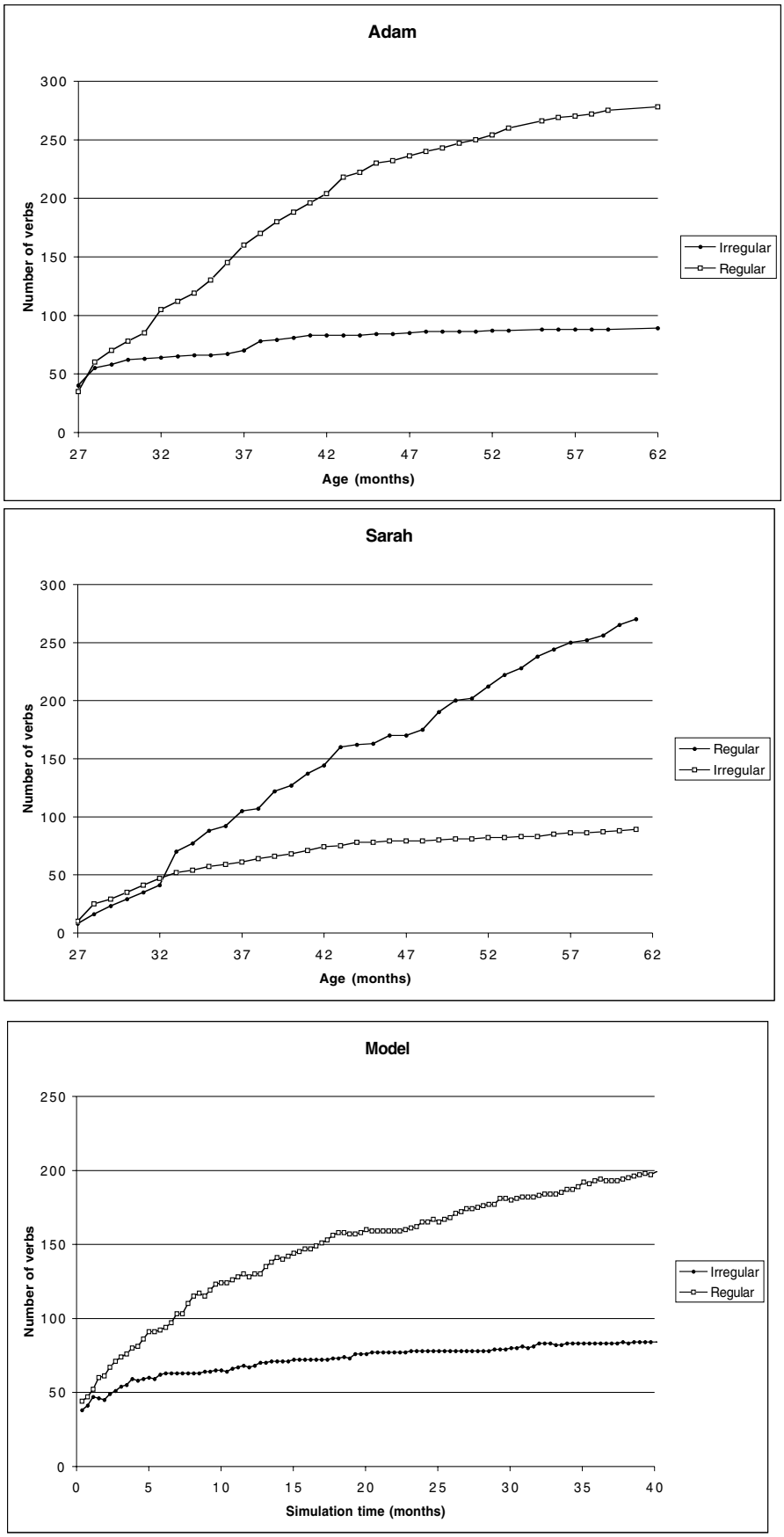
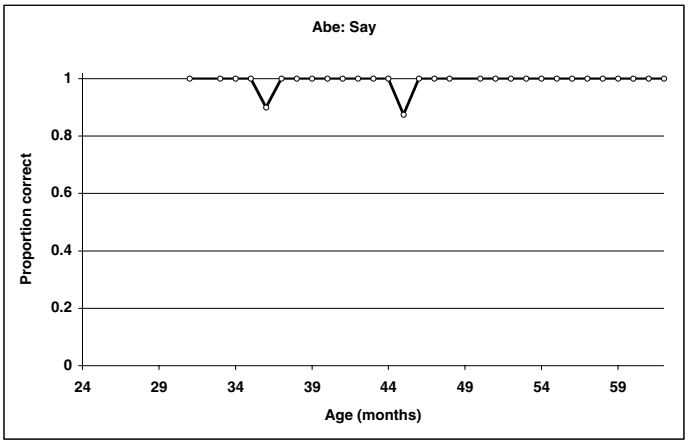
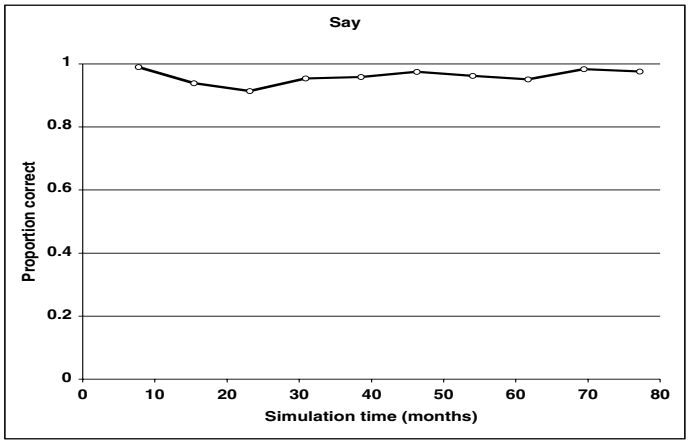
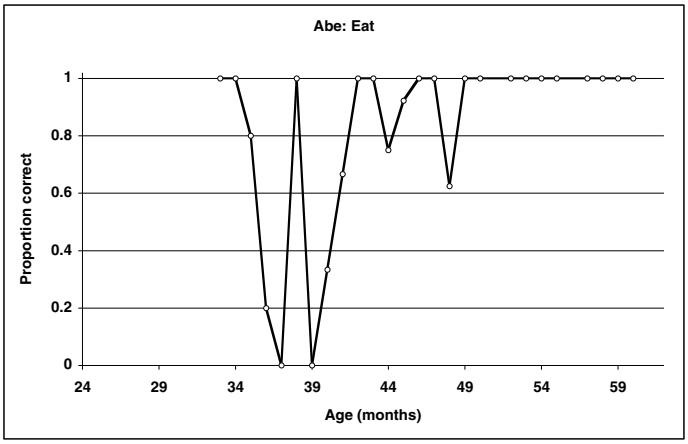
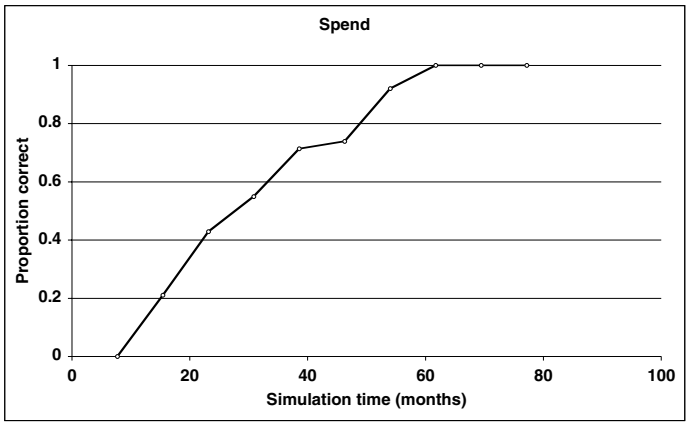


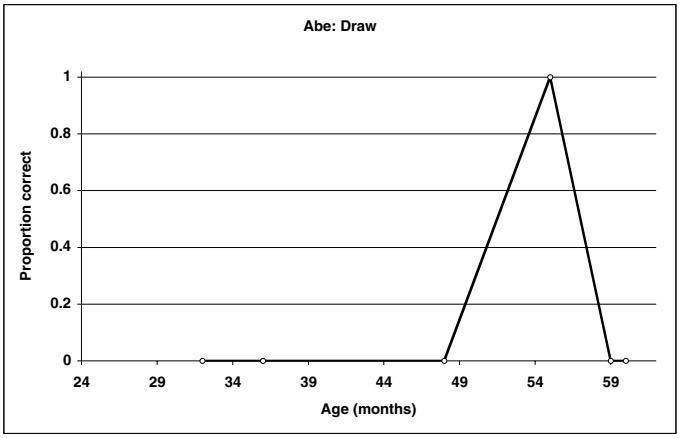
Figure 6. Vocabulary growth for Adam, Sarah (adapted from Marcus et al., 1992) and the model.

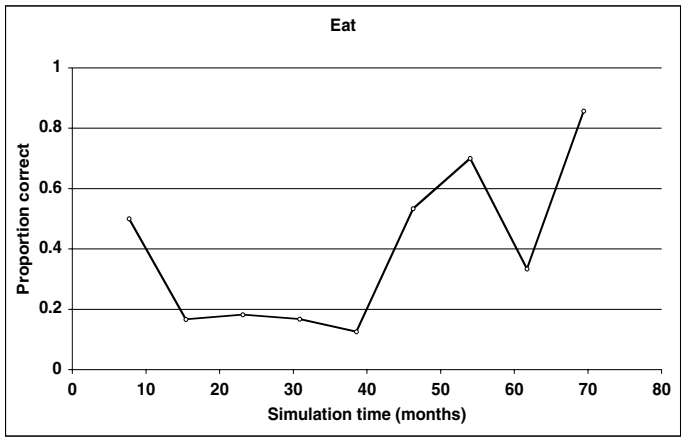


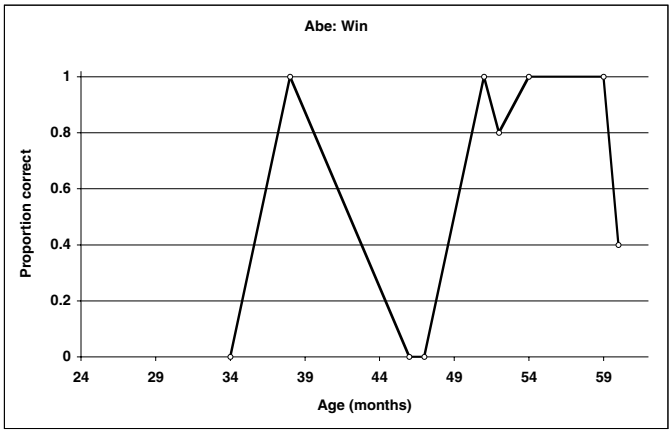












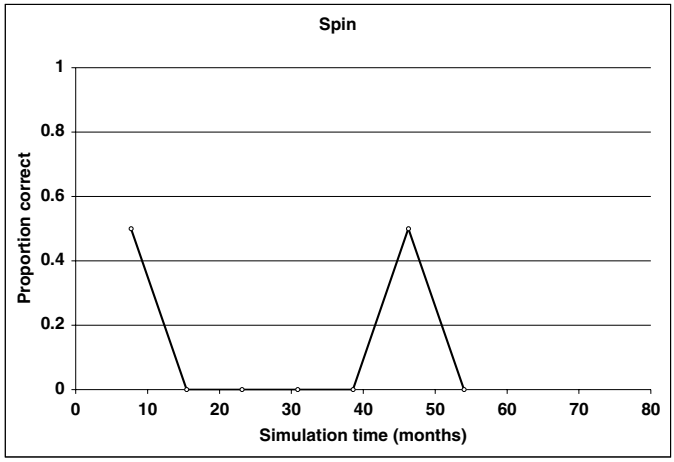


Figure 7. Overregularization of individual words for Abe (adapted from Marcus et al., 1992) and the model.

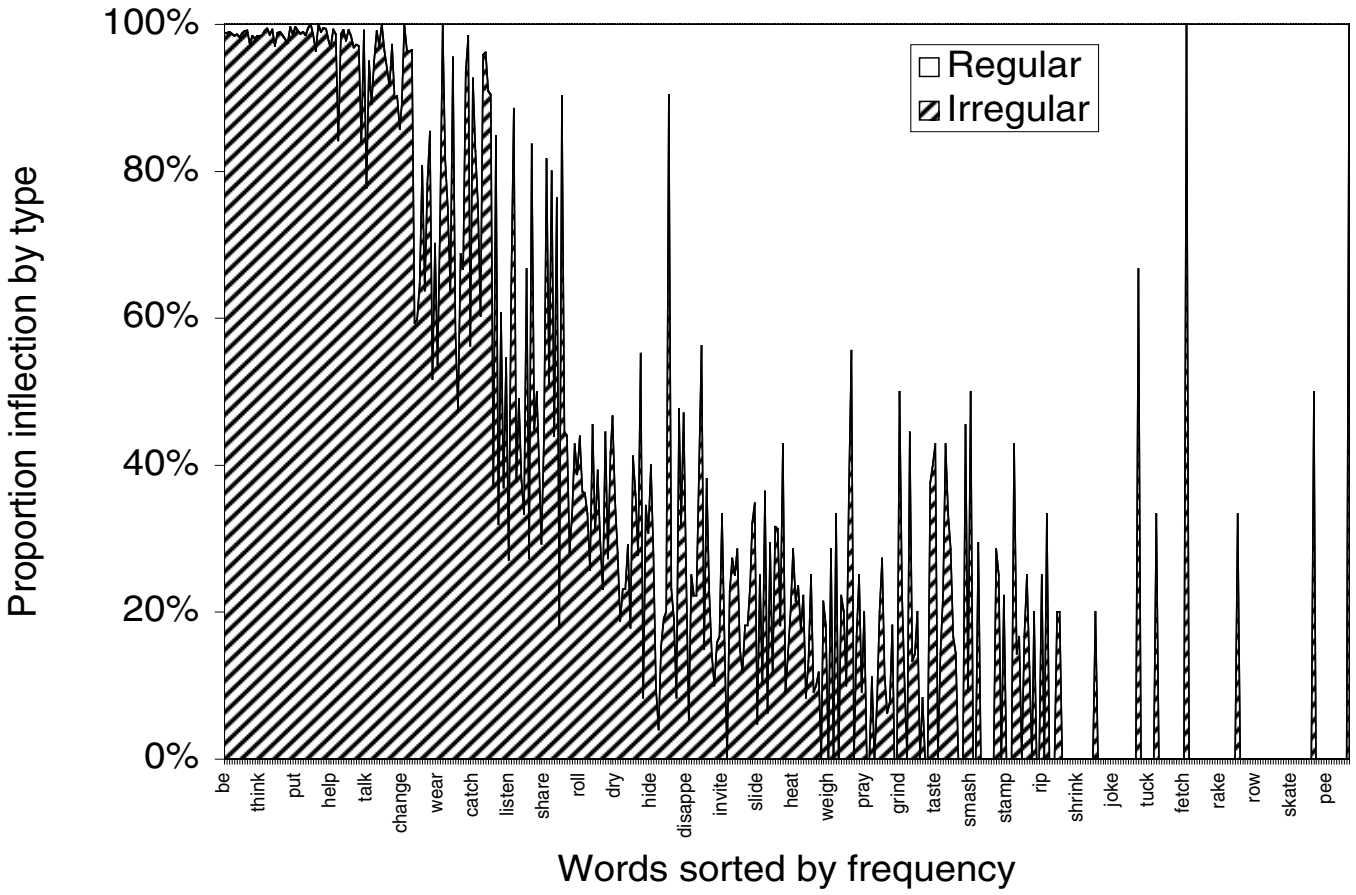


Figure 8. Proportion of times the model chooses an irregular or a regular past tense for a certain word. Words on the x-axis are sorted by frequency as in figure 1.

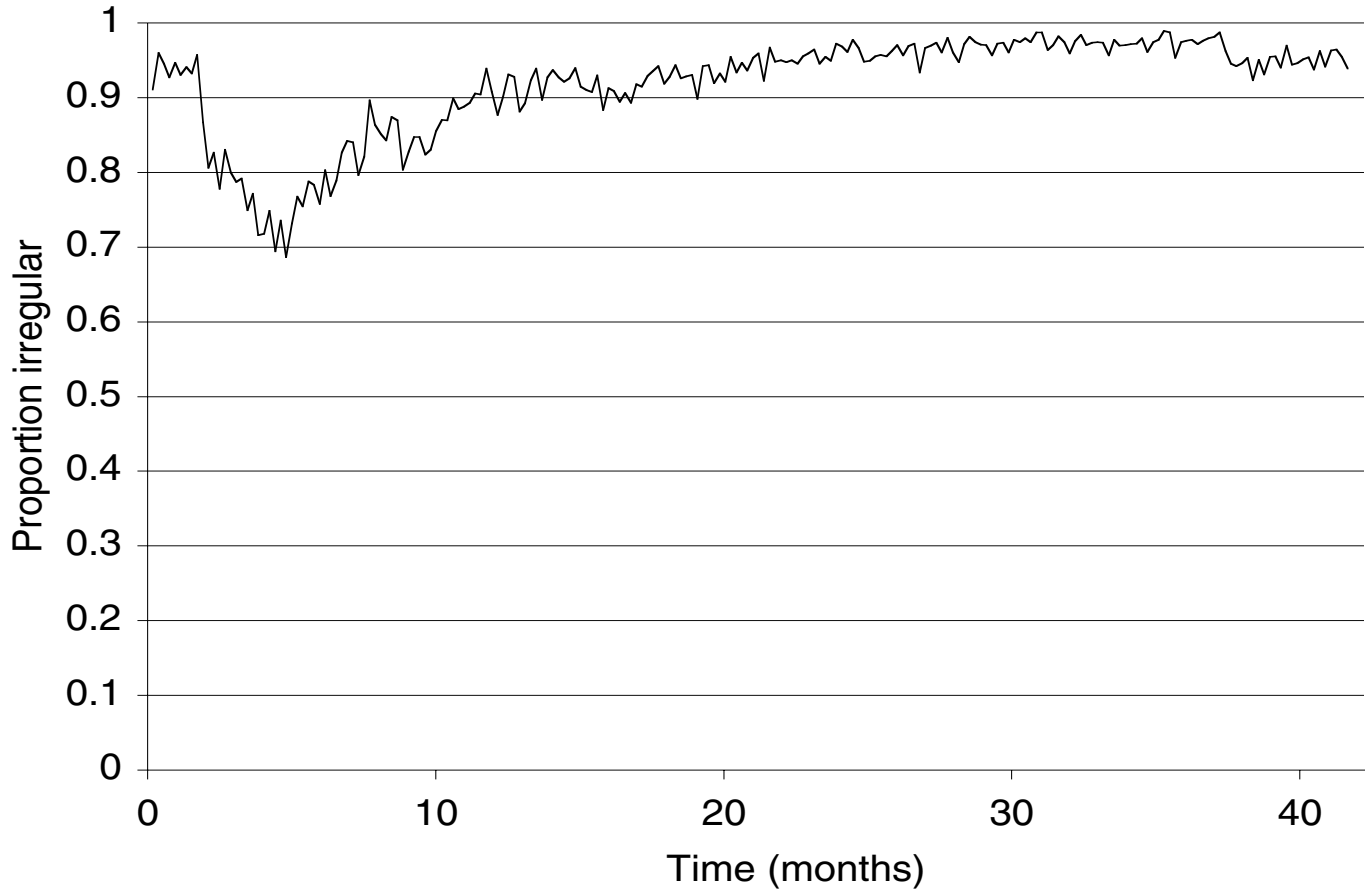


Figure 9. Proportion of times the no-feedback model produces an irregular past tense for an irregular verb as opposed to producing a regular past tense.