Modeling the learning of the English past tense with memory-based learning

Rik van Noord RIKVANNOORD@GMAIL.COM
Jennifer K. Spenader J.SPENADER@GMAIL.COM

Institute of Artificial Intelligence, University of Groningen, The Netherlands

Abstract

Modeling the acquisition and final state of English past tense inflection has been an ongoing challenge since the mid-eighties. A number of rule-based and connectionist models have been proposed over the years, but the former usually have no explanation of how the rules are learned and the latter often rely on implausible vocabulary growth and feedback assumptions. We investigate an approach that is able to address these criticisms, a memory-based learning model that is based on analogy. We were able to model the learning of the English past tense well compared to previous systems. However, a more detailed analysis showed a number of results inconsistent with child language data, including the generation of incorrect irregular forms (e.g. *think-thank* instead of *thought*) and the inability of the system to produce regularized forms with irregular roots (e.g. *broked*). We discuss how the model could be modified to accommodate this additional child production data, though note that these modifications would also detract from the simplicity, and thus plausibility, of the approach.

1. Introduction

In this paper, we will use memory-based learning (MBL) to model the production of English past tense and we will focus on the actual learning phase instead of the final model. MBL is a purely analogical approach in which learning is defined as the storing of all instances. Only these instances are then used to guide the classification of a new instance. There is no feedback or rule system involved and thus no implausible division between rules and facts in memory. Purely analogy-based approaches have been created before, but they usually focused on modeling individual experiments on adults and thus ignored the learning phase (e.g. Eddington, 2000; Albright and Hayes, 2003; Daelemans, 2002; Keuleers, 2008). We show that MBL is able to learn the past tense well in that it models most of the findings by Marcus et al. (1992) in child language data, and shows comparative performance as one of the better models, Taatgen and Anderson (2002). However, in a more detailed analysis we show that the model's production differs from child language results in a number of ways. We discuss these shortcomings, propose ways to deal with them, and then discuss the consequences and then the remaining challenges.

2. Previous research

We first discuss the most successful systems that have modeled the learning of the English past tense. We follow this with a brief overview of two systems that have instead focused on the final state in the adult language, but have used analogical learning as their basis.

In the learning of the English past tense,[1] three stages are typically distinguished. In the first stage, when children start inflecting verbs to mark past tense, they form the past tense correctly most of

[1] These learning stages have also been suggested to be more general, applying to all inflectional systems.

the time, and there is no distinction between regular and irregular forms. In stage two there is a decrease in performance, due to *overregularizing* of the verbs (e.g. *go- goed, break-breaked*). This is presumably due to children spotting the pattern that most verbs form their past tense by adding the morpheme [-d] to their root. Children seem to do this even for verbs that were previously used correctly. This leads to a drop in performance for irregular forms. In the third stage, the overregularization disappears gradually until perfect performance is reached. Because accuracy with irregular forms begins high, gradually drops quite low, and then slowly rises again, the process is usually called *U-shaped learning.* Marcus et al. (1992) did an extensive investigation of the past-tense acquisition of a large number of children in the CHILDES databse. They found, among other things:

- **No discontinuity of vocabulary** Overregularization does not correlate with a sudden increase in the proportion of regular verbs in parental speech, child speech or child vocabularies.
- **Frequency protection effects** The more often a parent uses an irregular verb, the less often a child incorrectly overregularizes it.
- **No multiple irregular forms** Wrong forms of irregular verbs were almost exclusively the result of overgeneralization, and not the creation of an incorrect irregular (e.g. *thinked* instead of *thought*, but never *think-thank.*)
- **Similarity effects** Verbs are protected from overregularization by similar sounding irregulars, but they are not attracted to overregularization by similar sounding regulars.
- **Double marking** children also sometimes add the regular inflectional morpheme to an irregular form, e.g. *thoughted* instead of *thought.*

The challenge then is how to model U-shaped learning as well as the above facts with realistic assumptions. Two types of computational model dominate: connectionist approaches and rule-based approaches, the topics of the next two sections.

2.1 Connectionist models of learning

The first computational attempt at modelling English past-tense acquisition was a connectionist approach by Rumelhart and McClelland (1986) using a two layer feed-forward network. It did produce U-shaped learning, but was heavily criticized by Pinker and Prince (1988), mainly because the model required a large increase in vocabulary at the onset of U-shaped learning. Such an increase in vocabulary isn't attested in children's productions or their exposure. MacWhinney and Leihbach (1991) also created a connectionist model that addressed many criticisms of the Rumelhart and McClelland (1986) model, including using more realistic frequency of occurrence information. However, this model was also unable to replicate the correct initial production of irregulars, (i.e. no stage 1 behavior).

Later Plunkett and Marchman (1993) proposed a connectionist model that was able to produce U-shaped learning. However, they created Stage 1 behavior artificially by training the model on 20 verbs until perfection, before expanding the vocabulary. Also, they still needed a discontinuity in the growth in vocabulary. Research on actual child productions has not found such a discontinuity (Marcus et al. 1992). However, the update of their model, Plunkett and Marchman (1996), fixed the latter issue and later Plunkett and Juola (1999) were able to produce U-shaped learning for verbs as well as for nouns in their connectionist approach, but had similar problems as Plunkett and Marchman (1993) in that they were unable to produce stage 1 without using implausible input for the model.

In summary, connectionist models have the advantage of relying on a simple and plausible learning mechanism, but struggle to create U-shaped learning without unrealistic assumptions about children's vocabulary. For this reason many researchers have rejected these models in favor of a rule-based explanation.

2.2 Rule-based approaches to learning

Inspired by the results of their detailed study of children's acquisition patterns, Marcus et al. (1992) propose a two-part systems that relies on lexical memory and a simple rule. Past tense forms for irregular verbs are retrieved from memory. Regular verbs instead are formed by a rule that is able to generate a regular form for any verb. In the final adult state retrieval is believed to be faster than rule-application, so an irregular form will always be produced if one exists. But for children, if the memory trace for an irregular verb is not strong enough, the regular rule can be used, and this accounts for the overregularization seen in Stage 2.

The first computational rule-based model was done by Ling and Marinov (1993), the symbolic pattern associator. They did obtain U-shaped learning, but did not offer an explanation of how the different rules are learned. Also, their model is not generalizable to other inflectional tasks in language.

Taatgen and Anderson (2002) proposed a rule-based model using an implementation of an already existing memory system (ACT-R, for a recent overview see Anderson and Lebiere (2014)) that was able to offer an explanation for how the different rules are learned. The model was able to capture the blocking mechanism of the regular rule in rule-based systems rather than assuming it and did not rely on a sudden large growth of the vocabulary or implausible amounts of feedback. However, they implausibly represent rules and facts separately in memory, since the procedural and declarative memory are two completely separate units in ACT-R. Also, their model never learned to exploit quasi-regularities in irregular verbs (i.e. groups of similar verbs that are inflected in the same way, such as *lay-laid* and *pay-paid*), while it is shown that adults do exploit these patterns (Bybee and Moder, 1983). But rule-based approaches could be modified to face this criticism. McClelland and Patterson (2002) mention that by augmenting rule-based approaches with features such as graded rule activations or probabilistic outcomes, e.g. quasi-regularities could be treated empirically correctly. But McCelland and Patterson (2002) also point out that these changes in essence reduce rule-based models to a variation of connectionist models.

In summary, rule-based models struggle to model a number of production facts in the child and adult system. The previous studies all had in common that they focused on the learning phase, with a key aim of explaining children's overregularization. However, more recent work has shifted attention towards the final state.

2.3 Analogy-based learning of the adult state

Modeling results of experiments with adults is another approach to understanding the past-tense system. Most experiments with adults have been done using nonce (fake) verbs and humans intuitions on these tasks is compared to the output of a trained model. The computational models are trained until perfect performance is reached, but the actual (plausibility of the) learning phase is ignored. This approach has as an advantage that a specific goodness of fit can be calculated on adult data, instead of the more common visual inspection of the results used to compare models with the sparse child data available as was done in the earlier approaches. Among these models we find a number that applied analogy-based learning.

Albright and Hayes (2003) tested two models, an analogical model and an inductive rule-learning model. They performed two experiments. In one participants were asked for the past tense of a nonce verb (for example the verb *rife*). In the second experiment participants were asked to rate the different past tense forms of a nonce verb (*rofe/rifed*). Their results showed that the ratings depended on the phonological stem of the verb for both regulars and irregulars, and regulars get lower ratings when they had to compete with plausible irregulars. Albright and Hayes (2003) used this as evidence against rule-based models, since those models only have a single regular rule for creating regular past tenses, meaning that the inflection of regular verbs by definition would never be influenced by irregular verbs. Their two computational models were tested on the same data as the participants, and they calculated goodness of fit scores to compare results. For their analogical model, they calculated confidence scores per verb for each class. This approach basically calculates the probability of a verb belonging to a certain class based on the similarity of the verb to the whole class. Ultimately their inductive rule-learning model outperformed their analogical model, leading them to dismiss the analogical approach as a plausible model. They accept that inflectional processes do exploit phonological similarity, but they believe that the relevant similarity could only be captured with symbolic rules.

Keuleers (2008) created an analogical model based on memory-based learning and criticized the approach by Albright and Hayes (2003). For one, he found a scaling error in their results, which makes it questionable if their inductive rule-learning model outperformed their analogical model at all. Also, his memory-based learning model outperformed both of the models created by Albright and Hayes (2003), leading him to conclude that analogy-based learning can realistically model English past-tense inflection. However, he also did not focus on the learning phase, which allowed him to take a number of shortcuts such as only using type information of the verbs and thus ignoring token (frequency) information.

Our approach is similar to Keuleers (2008), but differs in three key ways. First, we do not ignore frequency information of the verbs, because this is an important factor in determining the order in which the verbs are learned. Second, we implement a different K-nearest neighbour algorithm. Third, our focus is on applying an analogy-based method to the learning phase, aiming to recreate U-shaped learning with realistic assumptions.

3. Method

3.1 Memory-based learning

In MBL, learning is defined as the storage of instances in memory, in the form of multi-dimensional features. These instances, characterized by their feature profile, are then used to determine the classification of new instances. A new instance, also represented by the same features, is compared to the stored instances (via their features) and is then classified according to the instance(s) to which it is most similar in the multi-dimensional array. A key aspect of MBL used with linguistic problems is that there is no attempt to simplify the model by eliminating low frequency events or even exceptions because these can still contain valuable information for classification (see Daelemans et al. (1999) for empirical evaluations of MBL with language). For learning the past tense, we represent each verb as a number of (predefined) phonological features. The classification of a new instance is then usually done by using the K-nearest neighbour algorithm (K-NN). The classifier searches in memory for the K most similar instances based on the predefined features. It then classifies the new instance as the majority vote of the classifications of the instances found. The similarity between two different instances is calculated by using the modified value difference metric (Cost and Salzberg, 1993). This method calculates a distance matrix between the values of each feature based on the co-occurrence of classes and feature values. Two different feature values are considered more similar if they have a similar distribution over target classes. This ensures that we have different similarity scores for all feature-value pairs, instead of just a binary conclusion that they are either different or equal.

The features may have different weights as well. Their weights are calculated based on Gain Ratio (Quinlan, 1993), which is an entropy based method to calculate how much each feature contributes to our knowledge of the class label. These weights thus represent the importance of the different features in determining the past tense inflection and are used in calculating the difference between instances. The number of neighbours that are taken into account can be varied to achieve optimal performance. The described mechanism is used to inflect both regular and irregular verbs and no feedback regarding the output is involved. Even though we know whether a classification was right or wrong, that information is never used to influence the algorithm in any way. Note that no generative system is involved, meaning that MBL should be able to adequately deal with low frequency events, even when expressing sub-regularities (Daelemans et al., 1999).

One might argue that the memory system of MBL is not plausible, since it does not use activation or decay of the traces in memory. However, the storing of instances models activation implicitly, since a verb with more instances in memory is automatically easier to retrieve (more instances present in the range of K similar instances). This is exactly what activation is used for in other models, such as in Taatgen and Anderson (2002). The same logic applies for forgetting. There is no explicit decay mechanism, but due to the adding of other instances, it is harder for a particular instance to be retrieved (fewer instances present in the range of K similar instances). We then can consider this instance as less active (and eventually forgotten) on a higher level of abstraction.

3.2 Data

Table 1: A few examples of the possible phoneme features.

Verb	**Phoneme features**
be	*,*,*,b,ii,*
practise	pr,ae,k,tt,i,s
catch	*,*,*,k,e,tsj
stretch	*,*,*,str,e,tsj
match	*,*,*,m,ae,tsj

We use the same verb and frequency information as Taatgen and Anderson (2002). They used all the verbs that were mentioned in Marcus et al. (1992) and matched them with frequency of occurence information based on the work of Francis and Kucera (1982) who created a corpus of parental speech. Their list consisted of 466 verbs with their frequencies (382 regular, 84 irregular). Every instance consists of 8 predefined features; 6 features for possible phonemes (otherwise the feature has the value blank), one feature with the frequency and the final feature with the classification. Examples of the different phoneme features are shown in Table 1. The classification is not just regular or irregular. It would be impossible to form the past tense *was* from the verb *be* if the nearest neighbour was *have-had.* Very irregular verbs such as *be, have* and *tell* were all assigned their own category (i.e. *be* got the category BE). However, it would be possible to form *drank* from *drink* if the nearest neighbour was *shrink-shrank.* So, every group of regular irregulars was assigned to the same category, while every regular verb was assigned to the category REG. In this way, we can account for regular irregularities, but do not allow the model to make classifications that are too general. Specific numbers and categories can be found in Table 2.

Table 2: The categories and number of instances per category.

Category	**Number**	**Example**
REG	382	seem - seemed
IRREG-CONV	10	put - put
IRREG-I-V	6	win - won
IRREG-d-t	5	send - sent
IRREG-I-&	5	sit - sat
IRREG-i:-E+t	5	feel - felt
IRREG-@-U-u	3	grow - grew
IRREG-i:-E	3	feed - fed
IRREG-aI-I	3	fly - flew
IRREG-aI-@U	3	write - wrote
IRREG-E-O	2	get - got
IRREG-i:-@U	2	freeze - froze
IRREG-eI-U	2	take - took
IRREG-eI-@U	2	break - broke
IRREG-aI-aU	2	wind - wound
Single categories	31	be - was, have - had

3.3 Model

In the training phase, the training instances are drawn from the list with verbs based on a probability that is obtained directly through their frequency of occurrence in the parental speech corpus of Francis and Kucera (1982). The frequencies are linearly transformed to probabilities, i.e. a verb with twice the frequency in the corpus will also have twice the probability to be drawn from the list. This means that verbs with a very high frequency (e.g. *be, have, say*) occur way more often in the training set than infrequent verbs (e.g. *trick, faint, meow*). This is to ensure that the input of the model is similar to the input of real children. These training instances are then saved to memory. After a fixed number of training instances, the model is tested on all 466 verbs. This is to ensure that we observe exactly which verbs are overregularized and which verbs are not overregularized in all stages of development. This has no influence on the training of the model, since the tested verbs are not stored in memory and the weights of the model are not adjusted in any way. After every run, a number of accuracies of interest are calculated. They are shown in Table 3. The model is trained in total on 67500 training instances and tested 450 times (a test after every 150 training instances). This is, according to Taatgen and Anderson (2002), similar to the input of 3 years in child development. Every month of child development is thus simulated in our model by training on 1875 training instances. In our graphs, we simply display the number of months to accommodate the comparison with the results of Taatgen and Anderson (2002).

We mentioned before that the classification procedure is usually done by using the K-NN algorithm. We use a different implementation of this algorithm. Instead of searching for the K most similar instances in feature space, we look for all examples within a certain range from the instance to be classified. This range, denoted by variable σ (sigma), can be varied along the training phase to achieve optimal performance. This approach is usually called a Parzen Window (Parzen, 1962). We do this since, when children start to inflect for past tense, they do not have many examples in

Table 3: Formulas used to obtain the different results.

Results	**Formula**
Regular correct	Regular correct / Total regular
Irregular regularized	Regularized / Total irregular
Irregular no inflection	No inflection / Total irregular
Irregular correct	Irregular correct / Total irregular
Regular mark rate	Regular correct / Total regular
Overregularization	Irregular correct / (irregular regularized + irregular correct)

memory. This means that if they want to inflect for past tense, it is assumed that they have to take very dissimilar verbs into account to even produce an inflection. In other words, they have to increase their sigma to produce output. As they get older, their memory grows, and they are not dependent on dissimilar verbs anymore. They are now able to inflect for past tense by just using very similar verbs and thus decreasing their sigma again. We try to model this process by gradually changing the range in which the model looks for nearest neighbours. We would not get this specific gradual increase and decrease in that distance range by just varying the amount of neighbors we take into account. This phenomenon is also observed in another form by Albright and Hayes (2003), who note that when they performed the testing the traditional way (i.e. not on nonce verbs) their models classified new instances as regular virtually 100 percent of the time. This is a problem we would also encounter when testing on a large number of neighbours, which is necessary in the learning phase to even produce output at all. This is the reason for using the special implementation of the K-NN algorithm instead of the usual approach.

This approach means that we sometimes find only 1 (or even 0) nearest neighbours, but that it is also possible to find more than 1000 nearest neighbours in one of the later stages. If there are no nearest neighbours that fall within range, the algorithm outputs just the verb itself and failed to inflect for past tense. It is important to note that this does not count as a correct or incorrect classification. Remember that if a child is not able to mark for past tense, we do not know whether it was planning to use past tense at all. Therefore we cannot be certain that the model made a mistake and thus do not consider that example when calculating accuracy. This is also the case for verbs that have no inflection in the past tense (e.g. *hurt*) even though the model then accidentally produced the right inflection. We hereby ignore the fact that it is sometimes clear from the context that the children intended to use the past tense (e.g. by using words such as *yesterday* or *earlier*), following Taatgen and Anderson (2002). This artifact might also be one of the reasons why U-shaped learning occurs at all. When children just learn to inflect for past tense, they correctly inflect very common and easy verbs, while the failure to produce past tense for more uncommon verbs goes unnoticed since they produce no inflection at all. When their vocabulary grows, they still make a lot of the same mistakes as before, but now use the the wrong inflection instead of not producing an inflection at all. This results in more observed errors and thus in the observed but questionable U-shaped learning.

We try to model this phenomenon by varying the range in which neighbors are found and thus varying the sigma parameter. The threshold first *decreases* after every time the model is tested to account for the children being more liberal with their inflections and later *increases* to account for children not needing as many instances anymore to produce an inflection. This is modeled by the `divider` and `sigma change` parameter. `Divider` models how much of training set is used for increasing sigma, e.g. a value of 0.1 for `divider` means that `sigma` is increased in the first 10% of the simulation, while it is decreased in the other 90% of the simulation. `Sigma change` denotes by what value `sigma` is increased or decreased every 150 training instances. For example, with a

Table 4: Parameters with attented value ranges that were tested in the parameter optimization phase.

Parameter	**Explanation**	**Broad range**	**Refined range**
Divider	How much of the data is used for increasing the threshold?	{0.05, 0.1, 0.15, 0.2, 0.25, 0.3}	{0.17, 0.18, 0.19, 0.20, 0.21, 0.22}
Initial sigma	Starting value of sigma	{1.0, 1.5, 1.7, 1.9, 2.1, 2.3, 2.5, 3.0}	{1.75, 1.8, 1.85, 1.90, 1.95, 2.0, 2.05, 2.10}
Sigma change	Rate of change of sigma for every testset	{0.0001, 0.0005, 0.01, 0.015, 0.002, 0.003, 0.005}	{0.001, 0.0012, 0.0014, 0.0016, 0.0018 ,0.002}

`starting sigma` of 2.1, a `divider` of 10 and a `sigma change` of 0.002, the value of `sigma` after 6750 training instances (the increasing phase with 45 changes) will be 2.01. This threshold changing stops after the model reaches 100 percent correct performance at the point that is assumed to be the optimal value. To obtain optimal performance, we tested the model on a number of parameters, concerning the increasing range and thus sigma, the starting value of the threshold sigma and the value of by how much sigma is increased or decreased each time. The specifics are shown in Table 4. These parameters are now used to obtain the best general model, but can also be tuned in such a way to model the learning curve of individual children.

4. Results

Figure 1 shows the main results of the experiment. It shows the proportion of correct regular verbs, overregularized irregular verbs, irregular verbs without inflection and correct irregular verbs. In the final run we used the following parameters:

```
divider = 0.2
initial sigma = 2.1
sigma change = 0.0014
```

Figure 2 shows overregularization of the model as it is usually plotted together with the accuracy of regular verbs. The results are presented the same as in Taatgen and Anderson's study (2002), so that we are able to make a detailed comparison between the results. Their results are shown in Figure 3. We see in Figure 1 that the accuracy of irregular verbs quickly decreases after an initial short phase of mostly correct performance and then increases until ceiling performance is reached just below 100%. In the first few months we see an increase in overregularization, after which it gradually disappears again. We see the same trend in Figure 2. We also see in both figures that regular verbs are learned very quickly and that learning is almost perfected after only 4 months of input. In the first weeks of development we also see a lot of irregular verbs that were not inflected (not counted towards accuracy), but that also disappeared after 4 months. In general, the figures look very similar.

Figure 1: Proportions of regular correct, irregular regularized, irregular no inflection and irregular correct verbs plotted over time.

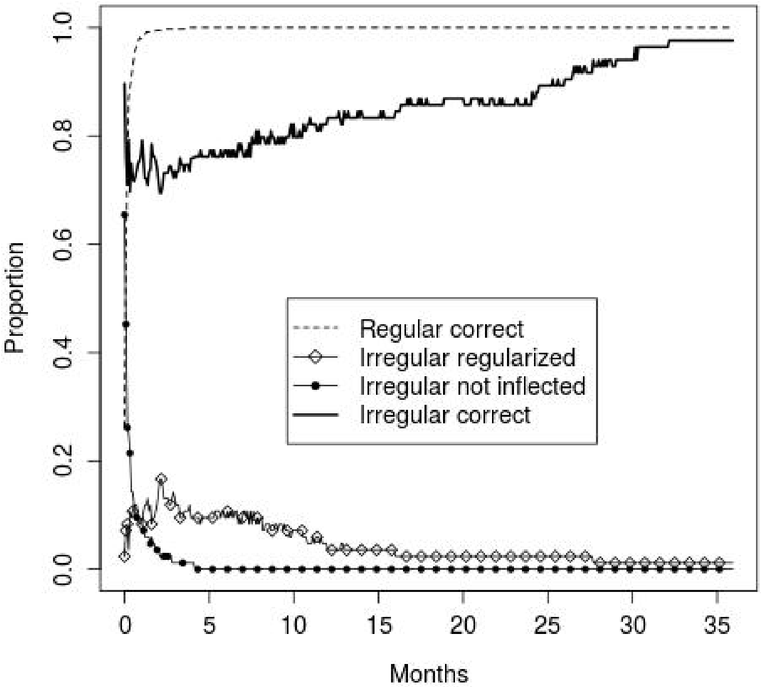


Figure 2: Regular mark rate and overregularization proportions plotted over time.

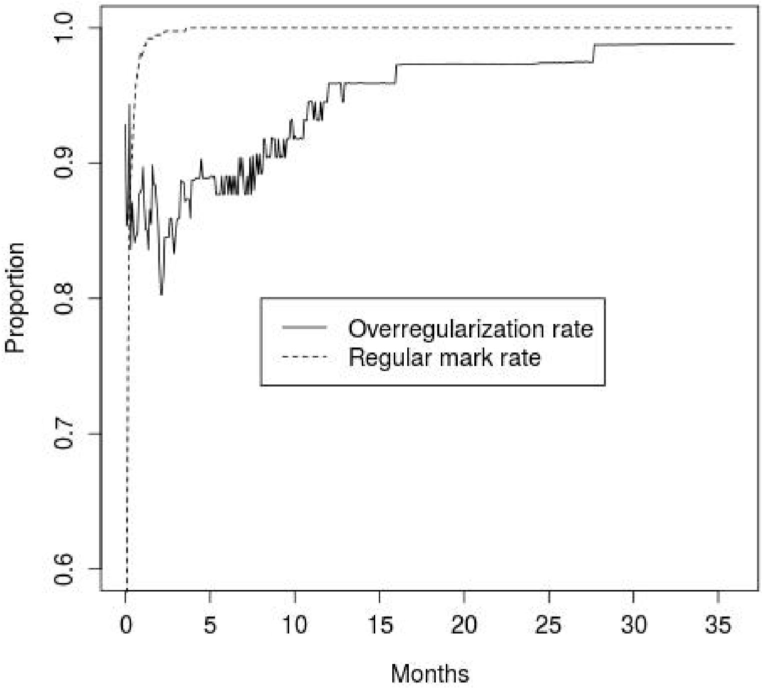
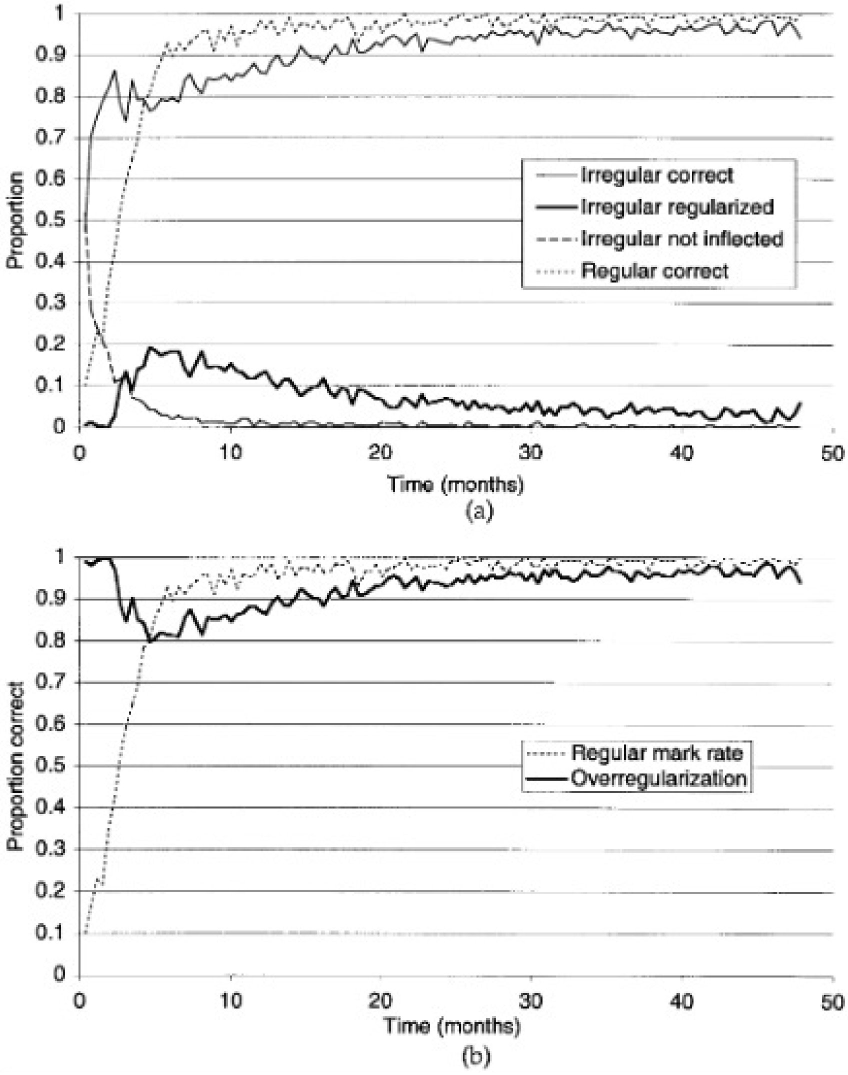


Figure 3: Results of Taatgen and Anderson (2002).



5. Discussion

In the following we compare our model to the ideal on several major points, including how closely it modelled U-shaped learning, whether or not it showed realistic frequency and similarity effects, and finally discuss its incorrect production of multiple irregular forms and failure to create double marked past tense forms. Throughout where possible we make direct comparisons with Taatgen and Anderson's (2002) model, which we followed quite closely in their assumptions and learning stages, though with a very different model.

5.1 U-shaped learning achieved

We see that our MBL model is able to replicate the learning of the English past tense by children in many respects. After an initial brief phase of almost correct performance in the inflection of irregular verbs, we observed a decrease in performance followed by a gradual increase to virtually perfect performance. MBL was able to do this without any predefined rules, explicit feedback or implausible discontinuity in vocabulary growth. It also does not show a decrease in performance for regular verbs during the decrease in performance for irregular verbs. It fits the observation made by Marcus et al. (1992) that the more often a parent uses an irregular verb, the less often the child will regularize it. This is because more instances in memory leads to MBL getting the right memory retrieval more often.

In comparison, Taatgen and Anderson lose some of the U-shape in the accuracy of irregular verbs. Their model starts out with an accuracy of about 50%, then rapidly increases to about 85% before we see the first characteristics of a U-shape. This is also the moment in development when overregularization starts to take place. Our model starts to overregularize at a similar time in development (about 2 months), but has a better initial phase of correct performance; it started out at about 90%, before showing a very clear U-shape (lowest point about 70%) at the exact point in time when overregularization is the most present.

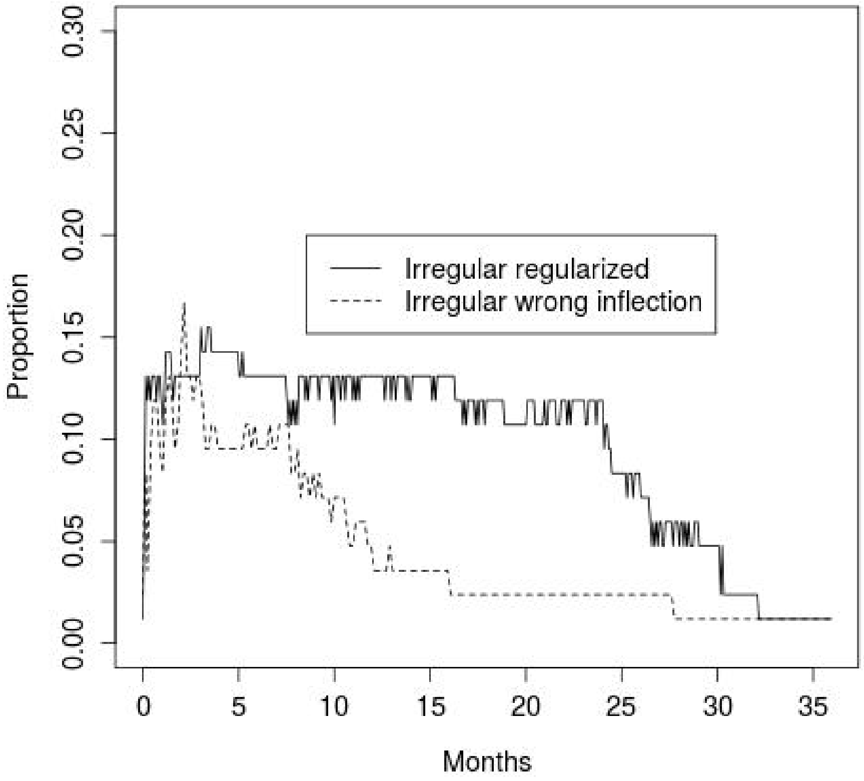
Our learning of the regular verbs is very similar to Taatgen and Anderson (2002) as well with accuracy increasing rapidly after initially starting out very low, although our model reached perfect performance somewhat more quickly. This same process can also be seen in the irregular verbs that received no inflection at all. The graphs are very similar, in that in the very early stages of development there are a lot of irregular verbs that do not get inflected, but both models quickly learn to produce inflections virtually 100% of the time. However, our model is again somewhat quicker in reaching 100% than the model of Taatgen and Anderson (2002). The overregularization rate shows a very similar trend in both models, an initial phase of overregularization, followed by the model learning the correct inflection and thus the disappearance of overregularizations.

5.2 Multiple irregular forms

However, on closer inspection we see our model has a number of problems. First, the overregularization period is about twice as long in the Taatgen and Anderson (2002) model than in our model. In our model, most of the overregularization disappeared from the output after about 12 months in development, while in their model the overregularization stabilizes only after about 24 months. Also, the proportion of verbs that is regularized in the initial phase is smaller in our model (average about 10%) than in their model (reaching as high as 20%, with a 12 month period higher than 10%). These two differences might not seem like major factors, but it effectively means that their model produces twice as many overregularizations as our model.

A possible explanation for this difference is that our model was able to inflect an irregular verb with the wrong *irregular* inflection, e.g. *draw-druw* or *ring-rong*, producing multiple irregular forms. It seems that when our model chose to inflect the irregular verb with the wrong irregular inflection, the Taatgen-Anderson model instead outputs an overregularization, and they explain that this is because rules that would produce such forms never become strong enough to be used. The overregularization and wrong irregular-inflection rates are shown in Figure 4. We see that initially, our model produces a wrong irregular inflection for an irregular verb about as often as it produces an overregularization of an irregular verb. However, inflecting verbs with the wrong irregular inflection disappears sooner from the output than overregularizations. It seems that if our model were to overregularize all irregular verbs that got the wrong irregular inflection, the proportions would be about the same as in the Taatgen and Anderson (2002) model. Since we know from Marcus et al. (1992), that children rarely produce the wrong irregular inflection for an irregular verb, this is an unfortunate result.

Figure 4: Proportion of the wrong irregular inflection for an irregular verb and the proportion of regularized verbs plotted over time.



5.3 Similarity effects

Recall as well that Marcus et al. (1992) also showed that verbs are protected from overregularization by similar sounding irregulars. In our model, irregular verbs are just as easily attracted to similar sounding irregulars as they are to similar sounding regulars. This results in the fact that it happens quite often that the wrong irregular inflection is used on an irregular verb, if that verb has a more frequent nearest neighbour of another irregular category (e.g. *buy* with neighbour *be*, *meet* with nearest neighbour *mean*). This means that verbs are indeed protected from overregularization by similar sounding irregulars in our model, but that it did not result in better performance for those verbs if the similar sounding verbs were from a different category. Because in Taatgen and Anderson (2002)'s model specific rules that lead to a verb being inflected with the wrong irregular form never grow strong enough to be used, their model is more realistic on this point. Modifying our model to deal with this problem will be hard if we want to retain the core principles of MBL that classification is done purely on the basis of analogy.

5.4 Double-marking

Another observation in real children that our model cannot account for is the tendency to regularize the irregular past tense itself (e.g. *break - broked, go - wented*) as we saw in Marcus et al. (1992) and Marchman and Bates (1994). Because our model always chooses the correct root of the verb, and then one of its possible inflections, it will never be able to produce such output. Most connectionist models and also the model from Taatgen and Anderson (2002) (after some modifications) did produce these double inflections.
A possible solution to this problem would be to allow the model to make these double inflections when the difference between the best and second best class in the range of neighbours is very small. However, this assumption is questionable, since there is no evidence in the literature that these double inflections correlate with multiple very similar verbs from different classes. If one would try this approach, it should be very fine-grained with possibly again a gradually changing threshold in determining whether the classes are similar enough.

Finally, one might question the variability of the regularization in MBL, due to the absence of probabilistic decision making in the algorithm. The calculation of the difference between the instances and the determining and application of the sigma threshold will always be performed without noise or variance. In practice, this means that verbs that are regularized at a certain point in time will continue to do so until they are learned the right way again. It is very rare that a verb is overregularized, learned and overregularized and learned again. Theoretically this might happen in MBL as well due to the random drawing of the instances, but in practice this very rarely occurs. In children, we see a lot more variability. Verbs that are used correctly can be incorrectly regularized a day later, and vice versa. This can be fixed by adding a noise component that possibly influences the calculation of difference between instances, the weight of the features and the estimation of sigma. Another possibility is adding noise to the saving of the instances in memory, since this is also never done incorrectly or flawed in our model. When implemented correctly, this could also solve the problem of double inflections and some verbs being regularized for virtually the whole development. However, it could also mean that we lose some of the plausible features in the process.

5.5 Comparison of individual verbs

An important aspect of the plausbility of any model is whether it overregularizes the same verbs as children usually overregularize. We compare our results to Marchman and Bates (1994), who did an analysis on the overregularization of 20 specific verbs in children. They showed the average of how often different verbs were regularized and how often they were produced correctly. It is unfair to compare the number of overregularizations per verb to our model, since we forced our model to be tested on every verb for every 150 instances in the training phase, while children often go months without producing a particular verb at all. To make a reliable comparison, we calculated the *overregularization probability* for every verb, which we defined as the number of overregularizations divided by the total number of forms produced for that verb. This is calculated for the Marchman and Bates (1994) data as well as for our data. Note that we cannot compare our model compares to the Taatgen and Anderson (2002) model in this aspect, since they don't report which verbs were prone to overregularization in their output. [2]

The results are shown in Table 5. We immediately see that the two most overregularized verbs in Marchman and Bates (1994), *drink* and *hold*, are never overregularized in our model. This is because they do not have a more frequent and very similar regular neighbour that attracts them to overregularization. For *blow, go, eat, make* and *lose*, the overregularization probability is more similar, but children still seem to overregularize more often than the model did. For the three most overregularized verbs in the model, Marchman and Bates (1994) unfortunately have no data

[2] Note also that we ignored many other important factors in the learning of verbs, such as transitiveness or different syntactic frames, see e.g. Tomasello, (1992) Naigles and Hoff-Ginsberg (1998).

Table 5: Comparison between overregularization probabilities for our model and the child data from Marchman and Bates (1994).

	Overregularization probability (%)	
Verb	**Child data Marchman and Bates**	**MBL Model**
Blow	48.3	24.9
Drink	39.0	0
Hold	38.9	0
Go	25.8	9.3
Eat	18.4	11.1
Make	15.9	7.4
Lose	12.4	9.4
Hurt	NA	81.1
Wear	NA	54.3
Catch	NA	51.9

available. It does not seem very plausible that *hurt* is barely learned over 2 years of development and is even overregularized 81.1% of the time. This is the case since *hurt* has the very frequent neighbour *hear* while being very infrequent itself. The range in which *hurt* looks for nearest neighbours is flooded with the regular *hear*, resulting in the fact that the correct examples of *hurt* do not have an influence on classification, except when sigma gets so low that the *hear* examples are out of range. The same logic applies (to a lesser extent) to *wear* (*share, scare*) and *catch* (*stretch*). This is not really plausible, since Marcus et al. (1992) showed that irregular verbs are not attracted to regularization by similar sounding regulars. In our analogy based model that is not the case by definition and indeed we found that irregular verbs with a very similar and more frequent regular neighbours are regularized the most often (e.g. *hear - hunt, catch - stretch, tell - spell, yell, smell*). Some test-runs with those regular examples removed from the train input showed that the irregular verb would indeed not be regularized so often (if at all). However, the fact that Marcus et al. (1992) failed to show a correlation does not necessarily mean that it is not there.

6. Conclusions and Future work

In this paper we presented a memory-based learning approach that was able to model the learning of the English past tense without using implausible feedback, implausible vocabulary growth or an implausible division between rules and facts in memory. Our results were very similar to Taatgen and Anderson (2002) and fit most of the findings by Marcus et al. (1992). However, a more detailed analysis showed a number of implausibilities and shortcomings of MBL that still need to be addressed in future work.

A very clear idea for future work is to test the final model on fake-verb data from other experiments, such as Albright and Hayes (2003). Recent approaches either focused on the learning phase or focused on the final model, but did not combine both into one model that could explain both phenomena. Our model also does not transfer well to tests on fake verbs, since it will never be able to inflect verbs that are very dissimilar to all other verbs in memory. If a very dissimilar verb such as *ploamph* is to be inflected, the algorithm will not find any instances in the range denoted by sigma. This is the case since sigma is very small after the development (to ensure it only took the right examples into account), but if a verb never occurred in the training phase, the range of possible neighbours will be empty and the output just the verb itself. This might be unsolvable if we want to model both the development and adult stage with the same model. However, it does not seem like a stretch that after full development, humans developed other mechanisms that deal with verbs that they never heard before, instead of relying on the same mechanisms that were used in their initial learning phase.

Another possibility for future work is the addition of a noise component somewhere in the system. Many of the implausibilities that were described earlier might be due to the absence of probabilistic decision making in MBL and might not be due to the principle of using analogy-based systems. It is interesting to combine this approach with the addition of other semantic features (such as features derived from the context), since Ramscar (2002) showed that meaning influences past tense inflection. MBL is a model that is easily extendable to contain semantic features (although they have to be predefined), as is already the case in models concerning coreference resolution (Hoste and Daelemans, 2005) and dative alternation (Theijssen et al., 2013). This way, MBL might lose some of its implausible deterministic and non-probabilistic features and in return may provide us with a plausible model that is able to model both the development and adult stage in the production of the English past tense.

7. Acknowledgements

Special thanks go out to Antal van den Bosch for his helpful suggestions and comments throughout the project.

References

Albright, Adam and Bruce Hayes (2003), Rules vs. analogy in English past tenses: A computational/experimental study, *Cognition* **90** (2), pp. 119–161, Elsevier.

Anderson, John R and Christian J Lebiere (2014), *The atomic components of thought*, Psychology Press.

Bybee, Joan L. and Carol Lynn Moder (1983), Morphological classes as natural categories, *Language* pp. 251–270, JSTOR.

Cost, Scott and Steven Salzberg (1993), A weighted nearest neighbor algorithm for learning with symbolic features, *Machine learning* **10** (1), pp. 57–78, Springer.

Daelemans, Walter (2002), A comparison of analogical modeling to memory-based language processing, *Analogical modeling. Amsterdam, Benjamins* pp. 157–179.

Daelemans, Walter, Antal Van Den Bosch, and Jakub Zavrel (1999), Forgetting exceptions is harmful in language learning, *Machine learning* **34** (1-3), pp. 11–41, Springer.

Daelemans, Walter, Jakub Zavrel, Kurt van der Sloot, and Antal Van den Bosch (2004), Timbl: Tilburg memory-based learner, *Tilburg University.*

Eddington, David (2000), Analogy and the dual-route model of morphology, *Lingua* **110** (4), pp. 281–298, Elsevier.

Francis, W. and Henry Kucera (1982), Frequency analysis of English usage, Houghton Mifflin Company.

Hoste, Véronique and Walter Daelemans (2005), Learning Dutch coreference resolution, *LOT Occasional Series* **4**, pp. 133–148, LOT, Netherlands Graduate School of Linguistics.

Keuleers, Emmanuel (2008), *Memory-based learning of inflectional morphology*, Universiteit Antwerpen, Faculteit Letteren en Wijsbegeerte, Departement Taalkunde.

Ling, Charles X and Marin Marinov (1993), Answering the connectionist challenge: A symbolic model of learning the past tenses of English verbs, *Cognition* **49** (3), pp. 235–290, Elsevier.

MacWhinney, Brian and Jared Leinbach (1991), Implementations are not conceptualizations: Revising the verb learning model, *Cognition* **40** (1), pp. 121–157, Elsevier.

Marchman, Virginia A. and Elizabeth Bates (1994), Continuity in lexical and morphological development: A test of the critical mass hypothesis, *Journal of child language* **21** (02), pp. 339–366, Cambridge Univ Press.

Marcus, Gary F, Steven Pinker, Michael Ullman, Michelle Hollander, T John Rosen, Fei Xu, and Harald Clahsen (1992), Overregularization in language acquisition, *Monographs of the society for research in child development* pp. i–178, JSTOR.

McClelland, James L and Karalyn Patterson (2002), Rules or connections in past-tense inflections: what does the evidence rule out?, *Trends in cognitive sciences* **6** (11), pp. 465–472, Elsevier.

Naigles, Letitia R and Erika Hoff-Ginsberg (1998), Why are some verbs learned before other verbs? effects of input frequency and structure on children's early verb use, *Journal of Child Language* **25** (01), pp. 95–120, Cambridge Univ Press.

Parzen, Emanuel (1962), On estimation of a probability density function and mode, *The annals of mathematical statistics* pp. 1065–1076, JSTOR.

Pinker, Steven and Alan Prince (1988), On language and connectionism: Analysis of a parallel distributed processing model of language acquisition, *Cognition* **28** (1), pp. 73–193, Elsevier.

Plunkett, Kim and Patrick Juola (1999), A connectionist model of English past tense and plural morphology, *Cognitive Science* **23** (4), pp. 463–490, Elsevier.

Plunkett, Kim and Virginia A Marchman (1996), Learning from a connectionist model of the acquisition of the English past tense, *Cognition* **61** (3), pp. 299–308, Elsevier.

Plunkett, Kim and Virginia Marchman (1993), From rote learning to system building: Acquiring verb morphology in children and connectionist nets, *Cognition* **48** (1), pp. 21–69, Elsevier.

Quinlan, J (1993), R.(1993) c4. 5: Programs for machine learning.

Ramscar, Michael (2002), The role of meaning in inflection: Why the past tense does not require a rule, *Cognitive Psychology* **45** (1), pp. 45–94, Elsevier.

Rumelhart, David E. and JL McClelland (1986), On learning the past tenses of English verbs, parallel distributed processing: explorations in the microstructure of cognition, vol. 2: psychological and biological models.

Taatgen, Niels A and John R Anderson (2002), Why do children learn to say broke? a model of learning the past tense without feedback, *Cognition* **86** (2), pp. 123–155, Elsevier.

Theijssen, Daphne, Louis ten Bosch, Lou Boves, Bert Cranen, and Hans van Halteren (2013), Choosing alternatives: Using Bayesian Networks and memory-based learning to study the dative alternation.

Tomasello, Michael (1992), *First verbs: A case study of early grammatical development*, Cambridge University Press.