A Single-mechanism Dual-route Model of German Verb Inflection

Nicolas Ruh (nruh@brookes.ac.uk)
Gert Westermann (gwestermann@brookes.ac.uk)

Oxford Brookes University, Department of Psychology
Gipsy Lane, Oxford OX3 0BP

Abstract

We present a constructivist neural network model of German past participle verb inflection. The model builds its architecture in response to the learning task in a way consistent with neurobiological and psychological evidence. In contrast to previous single-route models with a fixed, homogeneous architecture our model (1) reaches adult performance on an extensive corpus of German verbs, (2) shows U-shaped learning curves – even at the level of individual verbs – within a static learning environment, and (3) captures verb type specific dissociations with respect to neurological impairments. In contrast to dual-route symbolic theories, the model’s emergent notion of inflectional classes is based on distributional factors in the learning environment, thus obviating the need for in-built assumptions such as specific processing mechanisms based on grammatical class. By focusing on the German participle we demonstrate that the performance of the model does not depend on the existence of a dominant ‘default’ class. Taking seriously the constructivist, experience dependent nature of brain development, we suggest that such a single-mechanism dual-route model presents a step forward in the long standing, yet unresolved, past tense debate.

Keywords: past tense; neuroconstructivism; connectionist models; German past participle

Introduction

Models of past tense formation have in the past twenty years become representative of different theories of language acquisition and cognition in general. While connectionist approaches (see McClelland & Patterson, 2002) have maintained that both regular and irregular past tense forms can be produced in a homogeneous architecture by a single process, dual-route accounts (e.g., Pinker & Ullman, 2002) argue for two distinct mechanisms in different pathways, where regular forms are produced by a ‘default’ rule and irregular forms are stored in an associative memory.

Much of the discussion between the proponents of these opposite views has evolved around the phenomenon of U-shaped learning: the observation that children often go through a period of temporarily producing incorrect, usually overgeneralized past tense forms of verbs that were previously inflected correctly, until, eventually, the correct forms are produced again and an adult level of performance is reached (e.g., Marcus et al., 1992).

From a dual-mechanism point of view this phenomenon could be explained by the child’s discovery of the default rule which, initially, is applied too widely. Correct performance is then reached through a process of consolidation of the lexical entries for the exceptional (irregular) cases. This approach, however, struggles to explain the gradual onset of the overregularization phase and the observation that both correct and incorrect forms may be used intermittently for some period of time. Cases of irregularization, though infrequent, also appear to contradict this theory. From a more general point of view it can be argued that the lack of a computational implementation of the symbolic, dual-mechanism theory hinders a rigorous evaluation of this approach.

Conversely, single-route connectionist models were argued to exhibit U-shaped learning because initially, while the vocabulary is small and processing resources are ample, rote learning can lead to correct performance for early words (Plunkett & Juola, 1999). This is followed by a phase in which the expanding vocabulary leads to increased competition for limited resources until, eventually, a weight configuration is found that captures the underlying regularities and thus suits all verbs. These models have been criticized because their exhibiting a U-shaped learning profile crucially depends on the progressive extension of the training set during the learning process. This is problematic because, arguably, a growing vocabulary should be an attribute of the child rather than the child’s linguistic environment. Also, it remains unclear whether the observed overregularizations correspond to the unlearning of previously correct inflections or are due to new forms only.

Another area of controversy is concerned with evidence from brain imaging studies that appear to reveal differences in the localization of processes relating to regular and irregular inflections (e.g., Beretta et al., 2003). Similarly, verb type specific impairments have been observed in acquired (e.g., Patterson et al., 2001) and developmental (e.g., van der Lely & Ullman, 2001) neurological disorders. Although the causal attribution of such dissociations to a verb’s grammatical class is highly controversial (see, e.g., Seidenberg & Arnoldussen, 2003), these studies imply that the processing of different verbs can rely on different brain structures to a differential degree.

While homogeneous connectionist models have difficulties in addressing such findings, previous work on English past tense formation (Westermann, 1998, 2000) has demonstrated the potential of ‘constructivist’ neural network (CNN) models as an integral account of the above mentioned phenomena. CNN models aim to link brain and cognitive development by progressively adapting the network’s structure in response to the learning task. They are based on recent evidence that cortical structures develop...
partly in experience-dependent ways, and the argument that these (constructive and regressive) structural modifications lead to progressively more complex cognitive representations and enhanced processing of the types of stimuli that are frequently encountered. This view of cognitive development being closely intertwined with brain development has been termed ‘neural constructivism’ (Quartz & Sejnowski, 1997) or ‘neuroconstructivism’ (Mareschal et al., 2007; Westermann et al., 2007). From a learning theoretical point of view it has been shown that incorporating activity dependent structural modification into a learning system is not simply a matter of tuning performance, but that it changes the fundamental learning properties of the system (Quartz, 1993). Extending the notion of learning and development in neural network models to include structural modifications can thus overcome many problems that are associated with standard, fixed-architecture systems. Such constructivist models also allow for the investigation of how progressive modularization can occur in a single mechanism system where partial functional and structural dissociations emerge from distributional factors in the input-output mapping.

Critics of a single-mechanism approach to verb inflection have argued that the connectionist approach to modeling verb inflection crucially relies on a confound within the English inflection paradigm, namely the fact that the regular or ‘default’ case is also by far the most frequent one (Clahsen & Weyerts, 1994; Marcus, 1995). These authors suggest that German past participle formation presents a challenge to the connectionist approach because regular and irregular verbs are roughly equally frequent, and in contrast to English, both regular and irregular forms require a suffix to be attached to the stem. In parallel to English, verb type specific dissociations have also been found in a broad range of studies for German (Clahsen, 1999).

In this paper we respond to this challenge by showing that a CNN model can successfully account for the acquisition of German verb inflection. We demonstrate that the model captures a range of empirical phenomena (over-generalization, U-shaped learning, generalization to pseudo-verbs) during the process of reaching an adult level of performance. Further, it is demonstrated that the partial modularity which emerges as a result of the constructivist growth process results in a functional dissociation in which regular and irregular verbs can be selectively impaired when different pathways within the developed network are lesioned.

Past participle formation in German

German past participles are formed with a phonologically conditioned prefix ge-, a verb-stem and the endings -t or -n. Regular participles do not show stem changes in the participle and are suffixed with -t (1a-b). In contrast, irregular participles often show a modification of the stem vowel and take the ending -en (1c). Neither the stem vowel nor the phonological shape of the stem predict whether a verb is regular or irregular. Consider for example the verbs *blicken* (‘flash’) and *trinken* (‘drink’): whereas *trinken* (1c) has the irregular participle form *getrunken*, the verb *blicken* (1b) has the regular participle form *geblíkt*.

<table>
<thead>
<tr>
<th>Infinitive</th>
<th>Participle</th>
<th>Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1a) tanz-en</td>
<td>ge-tanz-t</td>
<td>‘dance’</td>
</tr>
<tr>
<td>(1b) blink-en</td>
<td>ge-blink-t</td>
<td>‘flash’</td>
</tr>
<tr>
<td>(1c) trink-en</td>
<td>ge-trunk-en</td>
<td>‘drink’</td>
</tr>
</tbody>
</table>

The Constructivist Neural Network Model

The model described here has previously been used to model acquisition and adult processing of the English past tense (Westermann, 1998, 2000). The model incorporates constructive and regressive events which depend on the learning task, implementing a neuroconstructivist approach to cognitive development.

![Figure 1: Initial architecture of the CNN. Arrows indicate full forward connectivity.](image)

The CNN model starts out with a minimal architecture (see Figure 1) with predominantly direct connections between input and output and a minimal hidden layer that grows during the learning process. Hidden layer units have a Gaussian rather than the more usual sigmoidal activation function, thus forming a receptive field (rf) for a region of the input space. All input vectors are positioned in this space according to their feature values and a hidden unit will become maximally activated if its position (= the center of its rf) coincides closely with that of the current input. For each input pattern, only the most active hidden unit contributes to the model’s output. Additionally, this unit’s connections from the input layer are adjusted slightly (learning rate = 0.001) so that the rf moves a fraction towards the position of the current input. All other connections in the model are adjusted via gradient descent (quickprop with a learning rate of 0.02 after each epoch).

The model attempts to learn the task with the given architecture until performance ceases to improve. When the global error has stagnated for at least five consecutive epochs, a new hidden unit is inserted. To comply with the idea of allocating additional processing resources where they are most needed, the novel unit is placed next to the rf whose activation previously caused the highest global error. Problematic areas in the input space thus become more densely covered with receptive fields, up to the (possible)
extreme of having hidden units that are exclusively used by one specific input pattern and thus perform purely localist, identity based processing of that input. Hidden units that never (within one epoch) contribute to processing are pruned.

Using this constructivist training regime, the CNN starts out with just two units in the hidden layer, each being responsible for roughly half of the input space. As training progresses, more structural resources are inserted until the coverage of crucial areas of the input space is sufficient to solve the task. Because the network always attempts to find the optimal solution with a given architecture, it will be forced to constantly reorganize itself as it develops over the course of acquiring the task.

Simulations

The training data for the model was based on all verbs within the German part of the CELEX database (Baayen, Piepenbrock, & van Rijn, 1993). From the initial set of 3,015 verbs we removed entries that were ambiguous with respect to their past participle form (including homophones), verbs whose stem had more than three syllables, and obvious CELEX errors. Further, prefixed forms of the same verb were combined into one simplex type, provided that such a simplex verb existed. For example, the verbs *einsehen*, *absehen*, *zusehen* were collapsed into the simplex verb *sehen* -> *gesehen* by summing all the corresponding past participle frequencies and assigning this accumulated frequency to the simplex type. Verbs with an accumulated frequency below 3 in 6 million were excluded from the corpus. This resulted in a basic corpus of 967 verb types and 47523 tokens.

In order to keep computation times within limits and to simulate individual differences with regard to the linguistic environment, the training corpus of an individual simulation run was constructed by randomly extracting 20000 tokens from this basic corpus. Table 1 shows the statistics of a typical training set.

Table 1: Distribution of regular and irregular verbs.

<table>
<thead>
<tr>
<th></th>
<th>types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regular</td>
<td>686 (81.5%)</td>
<td>11113 (55.6%)</td>
</tr>
<tr>
<td>Irregular</td>
<td>155 (18.5%)</td>
<td>8887 (44.4%)</td>
</tr>
<tr>
<td>total</td>
<td>841</td>
<td>20000</td>
</tr>
</tbody>
</table>

For the input to the model, each phoneme was translated into a binary phonetic feature vector, representing features such as *high*, *low*, *rounded*, and *frontal* for vowels (6 bits) and *coronal*, *nasal*, *voiced* for consonants (7 bits). This representation was entered into a three-syllabic template of the form XCCCVCCC for each syllable (X = stress (one bit); V = vowel; C = consonant), resulting in a 147 bit binary input vector. Distributed representations for phonemes were used to enable the model to represent degrees of phonological similarity between words; this property is relevant to the model’s ability to generalize.

For the model’s output, verbs were classified according to how their past participle is formed, resulting in 21

inflectional classes. One class represented the regular case (suffix *–t*, no stem change), the remaining 20 irregular classes accommodated all possible stem changes and suffixations (*–t* or *–n*). The main objective of this classification was to guarantee an unambiguous mapping of a stem to its past participle, given the inflectional class.

Training was non-incremental: the whole training set of 20,000 stem/past-participle-class pairs was presented in random order at each epoch. Hidden units were inserted depending on the learning progress (see previous section), and the network’s performance on all verb types was tested at intervals of 10 epochs. Training was terminated when classification performance was 100% correct over five consecutive tests. Classification was deemed to be correct when the activation of the output unit standing for the correct class, but no other output, was higher than 0.7.

Results

Results are based on ten simulation runs with randomly initialized weights (in the range of +/− 0.1).

![Figure 2: Mean performance of the 10 networks during training (small dotted lines indicate standard deviations).](image)

Learning: All ten networks reached the given performance criterion within 3000 epochs (mean = 2776.7, std = 103.9). By this time, the growing hidden layer contained on average 394.1 (std = 14.9) hidden units. As shown in Figure 2, learning was stable and regular verb types tended to be classified more correctly than irregular verbs. Note that neither of previous connectionist models (e.g., Rumelhart, 1986; Plunkett & Juola, 1999) of (usually English) past tense formation reached 100% correct performance.

Overgeneralization: Figure 3 displays the proportion of verb types that were overgeneralized although they had been classified correctly at an earlier point in training. The model’s performance shows a good fit with respect to the empirical data: reported overgeneralization rates during the acquisition of German verb inflection (Clahsen & Rothweiler, 1993; Weyerts, 1997) are in the range of 5-10%, 5-15% of which are incorrectly inflected regular verbs (irregularizations).
In order to quantify this effect at a more detailed level we analyzed the network’s performance on a subset of verbs that have been used in a recent empirical study with 5-7 and 11-12 year old children (Clahsen, Hadler, & Weyerts, 2004). The 60 verbs used in this study were divided into 4 conditions by the factors ‘verb type’ (regular/irregular) and ‘frequency’ (high/low). Participants were presented with a verb stem in a sentential context and had to produce the corresponding participle form.

Table 2: Mean percentage of suffixation errors (standard deviations in parentheses) of two age groups in German verb inflection, adapted from Clahsen et al. (2004).

<table>
<thead>
<tr>
<th></th>
<th>5-7 year olds</th>
<th>11-12 year olds</th>
</tr>
</thead>
<tbody>
<tr>
<td>irregular high</td>
<td>6.3% (9.0)</td>
<td>0.3% (1.5)</td>
</tr>
<tr>
<td>irregular low</td>
<td>27.4% (15.8)</td>
<td>8.6% (9.4)</td>
</tr>
<tr>
<td>regular high</td>
<td>1.8% (3.8)</td>
<td>1.1% (2.6)</td>
</tr>
<tr>
<td>regular low</td>
<td>1.4% (2.8)</td>
<td>0.7% (2.1)</td>
</tr>
</tbody>
</table>

Comparison of children’s performance (see Table 2)1 with the behavior of the model on the same 60 words (see Figure 4) yields a good qualitative and quantitative match with the empirical data. Similar to the younger children, networks in the early stages of training have most problems with low frequency irregular verbs and, to a lesser extent, high frequency irregulars. Intermediate stages then see a drastic reduction in the error rates for irregular verbs, down to negligible with respect to high frequency items.

Generalization to novel verbs. The models were tested on a set of 13 pseudo-verbs (adapted from Clahsen & Weyerts, 1994). Seven of these verbs (gornen, tolmen, tepfen, reipsen, tisseln, kirs en and taapen), while complying with the phonotactic regularities of the German language, were not phonologically related to any existing verb; a further three verbs (pachen, mauchen, mellen) were constructed as rhyming with existing regular verbs, and the last three (kessen, tiken, teiben) presented rhymes with irregular verbs. Figure 5 gives the regularization ratio of the CNN model for each of these groups. While the overall ratio of regularizations is close to the rate observed in children (69-95%), the study by Clahsen & Weyerts (1994) did not find a reduced tendency to regularize for the irregulars. Such an effect, however, was reported in a more extensive study with English pseudo-verbs (Prasada & Pinker, 1993) where the regularization rate for pseudo-irregulars was close to 50%, as opposed to ~90% for pseudo-regular verbs. This study also yielded a frequency effect for pseudo-irregular, but not for pseudo-regular verbs. Although the CNN model appears to display a similar tendency, the empirical basis (three verbs per group) with regard to German pseudo-verbs is currently insufficient to reliably address this issue.

U-shaped learning. U-shaped learning curves relate to the unlearning of previously correct inflectional forms and their subsequent re-learning (e.g., gegangen - gegeht/gegangt - gegangen). The CNN model captures this well documented phenomenon (e.g., Marcus et al., 1992, for English; Clahsen & Rothweiler, 1993, for German) at the level of individual verbs and, what is more, it also shows transition periods in which correct and incorrect inflections of specific verbs co-exist for a length of time. Figure 6 demonstrates that, as with children, low frequency verbs are more prone to overgeneralization than highly frequent ones. Looking at the magnification window it is discernible that the model’s classification of individual verbs often fluctuates for some time before settling for the correct inflection.

A further result corresponding to child language data concerned the protection from overregularization by similar sounding irregulars: the four irregular verbs rennen, brennen, nen nen, and kennen, for example, were hardly ever...

---

1 Note that the measure of incorrect suffixation does not correspond exactly to the overgeneralization error of the networks because of a small number of so-called ‘mixed verbs’ that require suffixation and a stem change.
overregularized (0.7%) whereas irregulars like dürfen (11.7%) or mögen (16.4%), despite comparable token frequencies, had a much higher chance of being transiently classified as regulars.

The CNN model was thus successful in capturing the phenomenon of U-shaped learning in verb inflection at a considerable level of detail. Note that this realistic developmental profile emerged in the absence of an external manipulation of vocabulary size (Rumelhart, 1986; Plunkett & Juola, 1999) or an essentially arbitrary parameter to control how often an exception verb has to be seen in order to be memorized (Ling & Marinov, 1993). Instead, it is caused by the ongoing reorganization of the network’s structural and functional set-up which predominantly affected those items the model found hard to process.

**Emergent modularity.** The functional dissociation between verb classes that has been observed in psycholinguistic experiments and in certain neurological disorders has led to the claim that the production of these forms is mediated by distinct neural pathways (Pinker & Ullman, 2002). According to this dual-mechanism view irregular forms are retrieved from an associative memory whereas regular forms rely on the application of a mental rule instantiated in procedural memory. This rule is assumed to be independent from environmental input, possibly having a genetic basis.

In the CNN, in contrast, a partial functional dissociation between regular and irregular verbs emerged as a direct outcome of the neuroconstructivist developmental process. This can be demonstrated by selectively lesioning the two pathways within the network (Figure 7): lesioning the hidden layer in the fully trained network left production of the regular verb class mostly intact but severely impaired the production of irregular classes (with about 90% of the errors consisting of overgeneralizations). By contrast, lesioning the direct connections had a larger impact on the regular verbs.

Apart from demonstrating that the irregular verbs, through the constructivist learning process, have come to rely more on the processing resources provided by the hidden layer, these findings are also reminiscent of similar double dissociations that have been observed in English patients (e.g., Tyler et al., 2002). The model’s account of German neurological data is reported elsewhere (Penke & Westermann, 2006) in more detail.

**Discussion**

The simulations described in this paper provide empirical evidence that constructivist neural networks can capture a host of phenomena related to the acquisition, processing and breakdown of German verb inflection. The ability of the CNN to develop its structure in response to the specifics of the learning task not only allowed it to allocate more structure to the difficult-to-learn irregular verbs, but also led to U-shaped learning curves and to emergent functional dissociations between regular and irregular verbs. Importantly, this (partial) functional modularity was not pre-specified but developed from the interaction of the constructivist approach and the distributional properties of the inflection paradigm.

Initially, the network had only two hidden units and the model therefore had to rely on the direct input-output connections for producing the correct participle classes. Given these restrictions the CNN initially performed poorly on many irregulars. During the training process, however, the model gradually extended its hidden layer, adding more receptive fields close to stimuli that produced poor performance. The ensuing reorganization of the internal representations thus affected mainly verbs that were at a disadvantage with regard to distributional factors (such as, e.g., neighborhood composition or frequency), and many of these particularly vulnerable items are irregular.

The fact that such verbs are particularly prone to be temporarily overgeneralized and, in general, to rely more on the additional processing resources provided by the hidden layer is thus a result of their distributional properties. It is this entanglement of processing and task structure which causes the model to closely follow the developmental profile observed in children and to reflect, in its final architecture, properties that can be found in normal and impaired adults. Together with the theoretical arguments for
constructivist learning these results offer compelling evidence for the usefulness of constructivist models in the study of cognitive development.

In view of the ongoing past tense debate our model advocates two main theoretical contributions. The first is to emphasize that the number of routes and the number of mechanisms need not necessarily coincide. Constructivist models are capable of developing distinct processing pathways on the basis of a single underlying mechanism, without the need of built-in domain specific knowledge, e.g., about grammatical class. The second point is to emphasize that the emergent functional and structural dissociations are only superficially aligned with the regular/irregular distinction. What really matters to the network – and possibly to the brain – is how easy or hard it is to process a specific stimulus. (see Seidenberg & Arnoldussen, 2003 for a similar perspective). Therefore, what we should be doing is to identify the distributional factors or combinations of factors that determine easiness.

Acknowledgements

This research was supported by ESRC grant Res-061-23-0129 to Gert Westermann.

References


