A connectionist account of English inflectional morphology: Evidence from language change

Mary Hare

Department of Psychology, Birkbeck College, University of London
hare@crl.ucsd.edu

Jeffrey L. Elman

Department of Cognitive Science, University of California, San Diego
elman@crl.ucsd.edu

Abstract

One example of linguistic productivity that has been much discussed in the developmental literature is the verb inflection system of English. Opinion is divided on the issue of whether regular/irregular distinctions in surface behavior must be attributed to an underlying distinction in the mechanisms of production. Looking at the course of historical development in English, the current paper evaluates potential shortcomings of two competing approaches. Two sets of simulations are presented. The first argues that a single-mechanism model offers a natural account of historical facts that would be problematic for a dual mechanism approach. The second addresses a potential problem for a single-mechanism account, the question of default behavior, and demonstrates that even in the absence of superior type frequency a network is capable of developing a "default" category. We conclude that the single network account offers a more promising mechanism for explaining English verb inflection.[1]

Introduction

A crucial fact about natural language is its productivity, and any successful model of language must offer an explanation for how this is accomplished. One example that has been much discussed in the developmental literature is the verb inflection system in English. Although the great majority of English verbs form the past tense through the regular and productive process of adding the suffix -d to the verb stem, English includes approximately 160 verbs which mark the past in some other way. Irregular verbal inflection has interesting properties that suggest it may be qualitatively different from the regular system.

There are two current proposals on how to account for these facts. One takes the difference in behavior between the regular and irregular systems to reflect an underlying difference in the mechanisms by which they are produced. On this account the regular verbs are the product of a symbolic system utilizing rules in the traditional sense, while irregularly inflected verbs result from an analogical network that identifies particular verbs as taking irregular past tense inflection, blocking the application of the regular rule (Pinker and Prince 1988). This view contrasts with the position taken by Rumelhart and McClelland (1986) and more recently by Plunkett and Marchman (1990, 1991). According to this work, both regular and irregular inflection result from the same mechanism, with differences in behavior attributed to other factors. The mechanism involved is a single connectionist network operating on largely analogical principles. The goal of the current paper is to consider these proposals in the light of other linguistically relevant data, having to do with historical change.

Issues

Both approaches deal only with the contemporary language facts. However, languages change over time, and it is important to ask how each account would explain historical change. The view of language as carried out by two competing subsystems raises interesting questions about the mechanisms of change. One reasonable assumption, given this approach, is that change involves the loss of a specific "blocking" effect: an exceptional inflection may be lost, allowing the regular rule to apply to a formerly irregular form. Hence irregulars may get regularized, but no mechanism is in place that would account for change in the opposite direction, or for analogical change among the regular verbs themselves. We believe historical change to be more complex than such a simple scenario suggests, and will report on the interim progress of a long-term project in this area. As we hope to demonstrate, a single-process model offers a natural account of facts that might be problematic for a dual mechanism approach and so appears to provide a more promising mechanism for historical change.

We will also address a potential problem for the network account, the need to learn a "default" inflection. It is important not to confound the phenomenon of a default with category size (cf Plunkett & Marchman 1991). Ear-

[1] This work was supported in part by a grant from the ESRC, a grant from the SERC Computational Science Initiative, and by a contract from U.S. Army Avionics (Ft. Monmouth, NJ).

ly work (Rumelhart and McClelland 1986) might be interpreted as suggesting that a network can learn to treat a category as an "elsewhere" case only if it is numerically superior to other categories. This interpretation arises from the belief that a network is a simple analogical system, capable of generalizing only on the basis of frequency and surface similarity. If this were true, a novel form would be treated as a member of the default category only if it resembles another form already learned in that category. Consequently, a true "default" could arise only if the category were well-populated and spanned the entire phonological space. We believe this to be an overly simplistic view of network dynamics, and will demonstrate that default behavior can be learned even in the absence of superior type frequency.

Historical data

Early Old English (ca 870) had approximately three classes of weak verbs, and seven classes of strong verbs. The strong verbs were the descendents of the IndoEuropean vowel change or ablaut series of verbal tense/number inflection. These classes decreased continually in size over the Old and Middle English periods, although remnants of the system appear among the irregular verbs even today (Stark, 1982, Wright 1925, Flom 1930).

The weak verb classes were also in a state of transition during this period. In early Old English, weak Class I, once the most productive weak class, had splintered into three subclasses and various exceptions. Class III was no longer a class at all, but rather four exceptional and very common verbs (live, have, think, and say). Class II, the largest and most productive, was the only class to exhibit a consistent paradigm. Over the OE period Class II attracted new members not only from the strong verb classes, but also from the two smaller subclasses of weak Class I.

One question of interest in the current paper is the motivation for this transfer from one weak class to another. Our claim is that the process of analogical attraction, which is known to affect the strong verbs of English, motivated the movement among the weak verbs as well. If true, the fact that the same process affected both categories of verbs is clear evidence in favor of a single mechanism account. Before continuing, however, we must justify an assumption that is basic to this argument: All classes of OE weak verbs were regular in the sense outlined in the Introduction. Lacking this, one might propose that a dual-mechanism account could easily explain the analogical behavior of the weak verbs by assuming that the verbs of Class I were irregular.

The clearest argument for the regular status of the weak verbs comes from Pinker and Prince's description of the dual-mechanism approach (1988). There the authors explicitly state that denominal verbs cannot be irregular:

"...irregularity is a property of verb roots. Nouns and adjectives by their very nature do not classify as irregular (or regular) with respect to the past tense.... Such verbs [denomincal and de-adjectival] can receive no special treatment and are inflected in accord with the regular system, regardless of any phonetic resemblance to strong roots."

All OE weak verb classes were made up of derived verbs, and in particular the members of both Class I and Class II were predominantly denominal (Flom 1930, Stark 1982). Thus, by established criteria, the dual-mechanism account defines these verbs as necessarily regular.

Ths paper offers a network account of why instability should arise in the first place, and why the resulting change took the direction it did. Since this account relies on the formal (phonological) similarity between members of the various weak verb classes, we will begin with a sketch of the relevant data (data based on Stark 1982).

Change in the Weak verb system

Early Old English

Proto-germanic weak verbs are classified according to the derivative suffix they took between the stem and the preterit suffix. Class I verbs were those which originally took the suffix *-j-*. This segment triggered various phonological changes in the verb stem, eventually resulting in three distinctive subclasses. In the indicative voice, typical EOE West Saxon Class I paradigms are as follows:

	Ia	Ib	Ic
	de:man	fremman	nerjan
	'judge'	'do'	'save'
present:			
1st sg	*de:me*	*fremme*	*nerje*
2nd sg	*de:mst*	*fremest*	*nerest*
plural	*de:ma?*	*fremmad*	*nerja?*
past:			
1st sg	*de:mde*	*fremede*	*nerede*
plural	*de:mdon*	*fremedon*	*neredon*
pst. part.	*de:med*	*fremed*	*nered*

The subclasses of Class I can be distinguished by the form of their stems. Ia (*de:man*-type verbs) is made up of long-stem verbs, with either a long vowel or a final consonant cluster. Members of Ib (e.g., *fremman*) were originally short-stem verbs whose final consonant geminated under certain circumstances, resulting in a stem alternation between CVC and CVCC. Ic (e.g., *nerjan*) consisted of a small group of short-stem verbs ending in *r*, which did not geminate. Verbs of this third sub-class also had a high front glide (*j*) stem-finally in certain parts of the paradigm.

Class II verbs, unlike those of the Class I subgroups, had no formal criterion of membership. Stems were consistently CVC, with no alternation with CVCC forms, and no requirement that the stem end in a specific conso-

nant or contain a long vowel. A typical Class II paradigm is given below:

lufian 'love'

	present:	past:
1st sg	*lufige*	*lufode*
2nd sg	*lufast*	*lufodest*
plural	*lufiad*	*lofodon*
	p. part: *lufod*	

The distinctions of note are these: In Class I, the short-stem verbs (Ib and Ic) take the suffix vowel *e* both in the personal endings of the present singular and as the "medial vowel" between the stem and suffix in the past. The long-stem verbs of Ia take no medial vowel except in the past participle. Verbs of Class II take the suffix vowel *a* in the present, and the medial vowel *-o-*. Furthermore, the high front vowel *i* appears in the Class II paradigm in all the forms where the corresponding glide *j* appears in Ic. II and Ia are large and internally coherent classes of verbs. Although Ib is smaller, it is still a good-sized class. Ic, on the other hand, is quite small.

Developments during Old English

Two phonological changes affecting the language as a whole had interesting consequences for the OE verbal system. First, throughout this period English developed a strong tendency toward glide vocalization. Both the *j* of the Ic (*nerjan*) verbs and the *ig* of the Class II 1sg present went to *i*. As a result, these two groups closely resembled each other, differing only in their medial vowel. At the same time, and arguably as a result of this increased formal similarity, the two classes collapsed into one. Verbs of the small Ic subclass adopted the medial vowel *-o-* of Class II, becoming indistinguishable from members of that class.

It was also during this period that English began to simplify its geminate consonants. Recall that one major distinction of the Ib (*fremman*) subclass was its alternation between geminate and non-geminate stems. This distinction disappeared by late OE. Interestingly, most of the verbs of the *fremman* subclass then adopted the Class II paradigm as well. At the same time a very small number of these verbs drifted into Class Ia.

The long-stem verbs of Class Ia continue unaffected throughout old English. By late OE there remained essentially two weak verb classes, Class II and the long-stemmed (*de:man*-type) subclass Ia. Over the course of Middle English the picture simplified further. Vowel reduction in unstressed syllables (and eventual deletion of the unstressed medial vowel) eliminated most remaining distinctions between conjugation classes.

The result of these changes is the regular past inflection of modern English. Thus from a historical perspective the regular past can be said to result from the operation of an analogical system, although this is obscured when one looks only at the synchronic data. In the next section we will demonstrate that a single network model is capable of accounting for this analogical change.

Network account of weak-verb change

Plunkett and Marchman (1991) analyze the conditions under which competing stem to inflectional mappings are learned in a single network. This work suggests that any mapping with low type and token frequency will be difficult to learn and is likely to be lost. Learning is aided, however, if the network is able to exploit phonological regularities in the relationship between the stem and inflected form. This is consistent with the weak-verb facts. At one stage in the language the smaller weak classes exhibited a certain amount of formal coherence. This made the learning task easier, since it allowed the learner to exploit information about the phonological characteristics of each class. It was as general phonological change eroded these characteristics that membership shifted.

In the simulations that follow we will demonstrate that this combination of factors leads to the correct behavior in the network. Part of the training procedure involves training multiple 'generations' of networks; the targets for each new generation are the outputs (after learning) from the previous generation. Although this training regimen does not claim to exactly mimic the time-course of language change across generations of speakers, it does capture the gradual nature of such change, and the causal role played by inaccurate transmission. In this way we hope to model one of the mechanisms underlying historical change.

Architecture and stimuli

The problem was modeled with a feedforward network implementing the back-propagation learning algorithm. The input bank consisted of 480 units standing for individual verbs, and 6 units standing for individual tense/number inflections. At each training iteration, one "verb" unit and one "inflection" unit were activated simultaneously, and the task of the network was to produce a representation of the inflected verb over the output units. The output was designed to represent the formal features that distinguished the various classes of weak verbs. For each 21-element output there were 12 units dedicated to these features, followed by a 9-element random pattern intended both to mark each verb as unique and to allow the network to treat each set of six inflected forms as individual manifestations of the same verb. The 12 "inflection" units represented the following information: (1) presence and identity of medial (or inflectional) vowel; (2)presence and identity of stem-final high segment; (3) presence or absence of gemination; and (4) presence or absence of a long vowel.

There were 480 "verbs", divided into four subclasses: 32 in class Ic, 64 in class Ib, 128 in class Ia, and 256 in class II. Each verb was learned in six inflected forms: 1st sg, 2nd sg, and plural present, 1st sg and plural pret-

erit, and past participle. These specific forms were chosen because in combination they illustrate the significant distinctions among the four subclasses of verbs.

Training and results

A training regimen was designed to test the hypothesis that low frequency mappings are difficult for the network to learn unless they are formally distinct. The first input-output mapping was based on the canonical OE weak verb system as described in Section III. A network was trained for 30 passes through the set of verbs in each of the six forms. After training the three largest classes (II, Ia, and Ib) were all produced correctly. The small subclass Ic showed the effect of attraction to Class II, producing a vowel rather than a glide for the stem-final high segment. The medial vowel in these verbs, however, is still that of Class I.

At this point a new network was set up and taught to produce the classes as formed after the application of glide vocalization and degemination. The second net was trained for 10 sweeps through the data set. Once more the two large and distinctive classes II and Ia are learned well. Error in class Ib is also low with one verb (out of 64) showing a tendency to adopt the high vowel of class II while a second shows an equally weak tendency to adopt the long vowel of Class Ib. The Ic verbs, as expected, show more interference. Three (out of 32) verbs tend toward Class II vowels. Still, the great majority of verbs of this class remain firmly Class I at this point.

A third network was then built and given as its teacher the output of the previous net. As a result, any errors in learning are propagated on to the next "generation", leading to increasing difficulty in learning those patterns. This training regimen was repeated for 5 subsequent networks, with each daughter network trained for 10 epochs to reproduce the output of its parent net. In each generation error on classes Ib and Ic increased as larger numbers of these verbs failed to learn the correct mapping and were produced on the model of Class II instead. Looking at the output of the sixth generation, we see that the Class Ic verbs are almost identical to those of Class II. No *nerjan*-type verb shows any interference from Class Ia, although at least two from the *fremman* class continue to be pulled very strongly in that direction. The majority of Ib verbs, however, have by this point merged with Class II.

Discussion

Consider the state of the weak verb system in early Old English. Class I had splintered into three groups, each of which had some strong formal criterion of membership. Class Ia, the long stem subclass, had the further advantage of high type frequency. Ic, on the other hand, had extremely low type frequency without any particular token frequency (Wright 1925). If a pattern of this kind is to be learned at all, a network account predicts that it must retain its formal characteristic. As predicted, it is when the distinctive form in the present tense is lost that Ic verbs begin to be modeled on Class II in the past tense as well.

A similar situation holds for the verbs of class Ib. The stem-final gemination was an identifying feature of these verbs, but once degemination had applied this information was no longer available. Class Ib, while not huge, did have a much higher type frequency than did Ic. This was an advantage, for although the loss of the geminate makes the class more difficult to learn, frequency of occurrence partially compensates. The class still eroded gradually, however. The more difficult learning task caused some members of this class to be mis-classified, and as this continued type frequency decreased, leading eventually to a situation not unlike that of Class Ic.

It might be asked, however, why the *fremman*-type verbs assimilated to Class II rather than class Ia. Part of the answer has to do with class size, for Class II was by far the largest class at this point. However, part of the answer also has to do with phonological form: Class II is the least restricted in terms of the phonological structure of its members. A small number of verbs, both in the network and in the real-language data, did in fact assimilate to Class Ia. Those that did so appear to have been drawn by surface similarity, suggesting that only the restricted set of Class Ib verbs sufficiently similar to those of Class Ia were able to merge with that class. No such constraint operated on assimilation to Class II.

While the two short-stem subclasses drifted into Class II, Ia remained unchanged. This is also the outcome observed in the network model. Ia not only had strong type frequency, but was also the one Class I subtype to maintain its formal characteristic. The combination of these two factors results in successful learning of the patterns involved.

The problem of the default

We now turn to a separate question, which is whether the default category can be learned even when it contains few members. Plunkett and Marchman (1991) raise this issue in a discussion of the Arabic plural system, in which the "sound" plural, the default, is of relatively low type and token frequency. They suggest that a crucial fact of this system is that "the numerous exceptions to the default mapping,... tend to be clustered around sets of relatively well-defined features."

The inflectional situation in earlier stages of English was parallel in many respects. As discussed in detail above, the weak preterit was treated as the default inflection even at a point when it enjoyed low type and token frequency. Furthermore, Vowel-change (strong) inflection applied to sets of verbs that exhibited internal phonological coherence. Each strong class had its own vowel series by definition, as well as other formal features by

which the classes were distinguished. The weak verb classes as a whole had no such criteria for membership. Although certain weak sub-classes could be formally characterized at different points in their development, this was not consistently true.

The hypothesis is that phonological information will be exploited by a network in following way. In learning to produce the exceptions, the net must learn to respond to the phonological characteristics that are typical of each class. To a network, this absence of phonological basis for default classification can be equally informative, allowing all patterns *not* meeting the criteria for membership in an exceptional class to be classified together, regardless of surface dissimilarity. The goal of the next section is to test this hypothesis and demonstrate that the learning requirement, far from being a stumbling block, leads to an insightful account of default behavior.

Simulation of default data

In the simulation that follows, a connectionist network is trained to perform a classification task based on OE data. The results of this simulation show that even when the type frequency of the default class is artificially constrained, a network is able to generalize appropriately to novel patterns.

Architecture and training

The network used a feed-forward architecture with 50 input, 18 hidden, and 6 output units. The net was trained using the back-propagation of error algorithm. Inputs were 50-element vectors, each representing a word in which a subpart is a particular VC or VCC pattern, defined over distinctive features. There were six output nodes, one for each of six categories. The task of the network was to learn to respond to each input pattern by activating the appropriate category label on the output.

Five of the six classes were made up of patterns based on the "characteristic" by which the OE strong verb classes were distinguished. In the network, these are the following:

1. *i* + any one consonant
2. *e* + one stop or fricative
3. *e* + a consonant cluster
4. *i* + a nasal+stop cluster
5. *a* + one consonant
6. any other VC or VCC

The goal was to show that the network could learn to treat the sixth class as an elsewhere case by reason of its phonological diversity. However, if this diversity allowed the class to sample the entire phonological space of the language, then generalization of new patterns need not result not from the special status of class six, but from similarity to some other member of the class. For this reason, the default class (6) contained a minority of the total forms to be learned. In training, the network was shown 32 randomly generated members of each class. Each was presented once per pass through the data set. The network was trained for 20 such passes, at which point error was extremely low.

Results

Generalization tests were then applied to assess the success of the network. An additional 32 members of each category were chosen from the initial random set, and given to the fully-trained network as input. All novel forms for classes 1-5 were categorized into the appropriate class. Of the 32 novel members of Class 6 (the default class), three were of the form "æ+C" and were most strongly classified as Class 5. A fourth which included the string "irn" was ambiguously categorized between classes 3 and 4. All others novel patterns were placed in Class 6. A second generalization test was then run, in which the network as tested on 63 patterns that did not match any subtype seen during the training phase. The vast majority were treated as members of the sixth (default) class. As in the first test, the net "mis-categorized" certain strings of the form "æ + C" into class 5, and often treated patterns with the vowel *i* followed by a CC cluster as being in either 3 or 4.

Discussion

These results are consistent with an account which claims that the network generalizes membership in classes 1-5 to novel patterns on the basis of surface similarity, but treats class 6 as the elsewhere category. This account explains what might have appeared to be errors in generalization. For example, Class 5 is made up of patterns including the string "a + C" and in the code used as input, *a* and *æ* differ in only one feature. Therefore the "æ + C" patterns are sufficiently similar to those of Class 5 to be attracted to that class. Note also that Class 4 patterns have *i* followed by a NC cluster, while Class 3 patterns have CC clusters after the vowel. Again, the similarity to previously learned patterns leads to predictable overgeneralization.

However, the results as stated do not rule out a second possibility: the network may be generalizing *all* novel patterns on the basis of perceived similarity to known patterns. To argue convincingly that the network developed a true default, we must eliminate this possibility. Two further tests were devised to make certain that membership in Class 6 was not based on similarity to known forms.

Similarity tests

In the first of these tests, we carried out a hierarchical clustering analysis on the input vectors of the training and initial test items. We found that for each of the five well-defined classes, the training data were clustered into the five appropriate groups. Members of class 6 (the default) were scattered randomly through the cluster tree. This is as expected, since there is no similarity basis for class 6 membership.

The initial test items for classes 1-5 were also clustered with the appropriate groups. Many of the class 6 test items also fell near the class 6 training items, since in the first test many of the patterns were indeed similar to the training data. Others, however, were clustered in a major branch by themselves, and dissimilar to any of the other forms (including the previously learned class 6 members). These truly novel forms were also classified by the network as class 6.

Finally, we constructed a third generalization test set which contained patterns that were not only dissimilar from any that had been seen during training, but deviated in various ways from the well-formedness criteria by which the training data were generated. This dissimilarity was apparent in a clustering analysis. When these new items were presented to the network, all were classified as members of class 6, the default.

General discussion

Certain synchronic facts about English are compatible with both the dual- and the single-mechanism approaches to inflectional morphology. First, in the current state of the language there is a single "weak" or *d*-suffixed conjugation. Second, this suffixed past is by far the most common form of English past tense inflection. The weak past tense is the default in that it applies productively as the elsewhere case to any verb not previously learned as irregular: Borrowings from other languages, verbs derived from nouns, adjectives, and other verbs, and neologisms all routinely take the *-d* suffix.

Old English and its immediate predecessors differed from the modern language in two crucial respects. First, there was not a unique *d*-affixed conjugation in early Old English. In proto-Germanic there were four such classes, and while these were distinct from each other, all were productive, applying to the many borrowings and denominal verbs that are characteristic of the language. By early Old English these four classes had been effectively reduced to two, and over the course of the OE period many verbs of the first class began to take on the characteristics of the second. This shift presents a difficulty for an account which disallows phonological information in the application of the regular rule, for as discussed above the migration from one weak class to another was largely governed by phonologically-based analogy.

Second, the weak preterit was an innovation in proto-Germanic. At that stage in the language it had no statistical edge over the more established ablaut series of past tense formation, yet even at an early point it appears to have been treated as the default, used to inflect borrowings and derived verbs. This points to the need to separate out the properties of the default from the effects of frequency and class size.

In this work we have tried to demonstrate the importance of considering historical change when trying to understand the language mechanism which underlies morphological processes. We believe historical facts may usefully constrain hypotheses about the nature of the language mechanism in cases where synchronic facts alone admit multiple hypotheses. The current work, although preliminary, suggests that a single network account of English verbal morphology may have advantages over dual-mechanism accounts. The current work also acknowledges that default categories in language need not require large type frequency, and suggests that this property can be successfully modeled in a connectionist network.

References

Flom, G. 1930. *Introductory Old English Grammar and Reader*. Boston: Heath and Co.

Pinker, S., and Prince, A. 1988. On language and connectionism: Analysis of a Parallel Distributed Processing model of language acquisition. *Cognition* 28:73-193.

Plunkett, K., and Marchman, V. 1991. U-shaped learning and frequency effects in a multi-layered perceptron: Implications for child language acquisition. *Cognition* 38: 3-102.

Rumelhart, D., and McClelland, J. 1986. On Learning the Past Tense of English Verbs. In Rumelhart, D., and McClelland, J., eds. *Parallel Distributed Processing, Vol. II*. Cambridge, MA: MIT Press.

Stark, D. 1982. *Old English Weak Verbs*. Turbingen: Niemeyer.

Wright, J., and Wright, E. 1925. *Old English Grammar*. Oxford: Oxford University Press.