Cognitive Psychology **41,** 313–360 (2000)
doi:10.1006/cogp.2000.0737, available online at http://www.idealibrary.com on IDEAL®

German Inflection: Single Route or Dual Route?

Ulrike Hahn

University of Wales, Cardiff, United Kingdom

and

Ramin Charles Nakisa

Oxford University, Oxford, United Kingdom

The German plural system has recently become a focal point for conflicting theories of language, both linguistic and cognitive. Marcus et al. (1995) highlight the German plural as support for the dual-route account of inflectional morphology first proposed by Pinker and colleagues (Pinker & Prince, 1988). On the dual-route account, inflectional morphology is universally subserved by a symbolic rule route which deals with regular inflection and an associative memory component which deals with irregular inflection. This contrasts with single-route connectionist systems. We seek to counter supposed evidence for the dual-route account through large-scale simulations as well as through experimental data. We argue that, in its current form, the dual-route account is incapable of generating experimental data

This research was primarily conducted while both authors were at the Department of Experimental Psychology, University of Oxford. Ramin Nakisa is now at the Knowledge Lab, NCR Financial Systems Ltd. Earlier versions of the six basic simulations, presented in The Data Set to Back-Propagation Network, which investigated predictive accuracy on the extant lexicon, were presented in Nakisa, R. C., and Hahn, U. (1996), ''Where Defaults Don't Help: The Case of the German Plural System,'' Proceedings of the 18th Annual Meeting of the Cognitive Science Society, and briefly summarized in Nakisa, R. C., Plunkett, K., and Hahn, U. (2000), ''A Crosslinguistic Comparison of Single- and Dual-Route Models of Inflectional Morphology,'' in P. Broeder and J. Murre (eds.), *Cognitive Models of Language Acquisition,* MIT Press, Cambridge, Massachusetts. The order of authors is arbitrary. Thanks are due to John Greenhow, Nick Chater, Glyn Collis, Gordon Brown, Elizabeth Maylor, Todd Bailey, and Kim Plunkett. Special thanks are due to Linda Bautz from the Institut for Molecular Genetics, University of Heidelberg, Germany for her help in collecting our behavioral data. Finally, we thank Steven Pinker for very helpful comments as a reviewer. Ramin Nakisa was supported by an MRC Postdoctoral Training Fellowship.

Address correspondence concerning this article to Ulrike Hahn, School of Psychology, Cardiff University, P.O. Box 901, Cardiff CF1 3YG, United Kingdom E-mail: hahnu@cardiff.ac.uk or to Ramin Nakisa at The Knowledge Lab, NCR Financial Systems Ltd., 206 Marylebone Rd., London, NW1 6LY, United Kingdom.

1053-8100/00 $35.00

provided by Marcus et al. (1995) as support. Finally, we provide direct quantitative comparisons between single-route and dual-route models of German plural inflection and find single-route performance superior on these tests.

INTRODUCTION

Rumelhart and McClelland's (1986) connectionist model of the English past tense challenged long-held assumptions about the nature of linguistic knowledge, simultaneously addressing both knowledge representation and learning. In this model, declarative knowledge (e.g, symbolic rules) was replaced with weighted connections in a connectionist pattern associator; learning was construed as a data-driven adjustment of these connection weights. These aspects were specifically highlighted as potentially underlying not just inflectional morphology but linguistic knowledge in general. The model's impact went beyond research on language and it became one of the cornerstones in the more general debate on symbolic vs connectionist architectures (Fodor & Pylyshyn, 1988; Smolensky, 1988) as well as an important test case for rule- vs similarity-based categorization (Nosofsky, 1988a; Nosofsky, Clark, & Shin, 1989; Bybee, 1995; but see also Hahn & Chater, 1998).

As a result, the model prompted fierce debate (Pinker & Prince, 1988; MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1993). Many of the early criticisms, in particular of representation and training schemes, were addressed and rectified in subsequent research (MacWhinney & Leinbach, 1991; Plunkett & Marchman, 1991, but see also Marcus, 1998). However, inflectional morphology continues to provide a central testing ground for conflicting views of language. In particular, insistence on symbolic rules is maintained by proponents of the dual-route account put forth by Pinker (1991). According to this view irregular and regular forms (e.g., sing/sang vs walk/walked) are subserved by different architectural components: irregular forms are based on lexical memory while regulars are produced by a symbolic rule.

Recently, Marcus et al. (1995) supplied evidence for the dual-route account in the form of two important and interlinked arguments backed by new linguistic and psychological data. For the first and main argument, they compiled a wide variety of circumstances in which the regular form is produced, arguing that the assumption of an underlying symbolic rule provides both a unifying and parsimonious account of these otherwise heterogeneous circumstances. In a second argument, they drew attention to languages which differ markedly from English in their relative proportions of regular and irregular forms and which, as a result, ought to pose severe difficulties to connectionist models. Connectionist models, so the claim, are statistical models which require very specific statistical regularities in the input in order to achieve the required behavior, but these requisite distributions are not always present. Crucially, languages where the regular or ''default'' form

is not also the most frequent, so-called ''minority default inflections,'' are assumed to lie outside the scope of connectionist models and thus provide further evidence for a symbolic rule.

In the first instance, this article investigates the extent to which both lines of argument actually support the dual-route account.

We examine the extent to which the existence of minority default inflections supports the symbolic rule claim through computer simulations with Marcus et al.'s (1995) chosen minority default system—the German plural system. These simulations directly compare single-route and corresponding dual-route models. These simulations reveal that dual-route models, and not just connectionist models, are *distribution dependent*. That minority default inflections are necessarily better handled by dual-route systems turns out to be false; this finding undermines the ''minority default argument'' that the mere existence of such inflectional systems is evidence for the dual-route account. The simulations involve large language samples and use a range of models. These models are chosen such that a range of potential confounds in the debate—specifically, connectionist/symbolic, single-route/dual-route, and similarity-based/frequency-based—can be separated. Among these models is Nosofsky's well-known Generalized Context Model of categorization, which has previously never been tested on large, naturalistic data sets, nor been considered by linguists and psycholinguists. Consequently, these simulations are informative beyond the single- versus dual-route debate. The three models—a nearest neighbor classifier, the Generalized Context Model, and a standard backpropagation network, all in both single- and dual-route versions—are tested for their ability to correctly predict German plural forms for previously unseen words. In all cases, the dual-route models fail to achieve better performance than the single-route models; we illustrate the reasons for this failure by means of a simple artificial language and provide an example of where a dual-route model would be superior.

We then draw on behavioral data to examine the other argument for a symbolic rule—its parsimony and unifying power. We examine whether a single symbolic rule really can underlie the range of circumstances in which the regular form is preferred and, thus, whether the proposed unifying account is feasible. We provide both theoretical analysis and novel experimental data, which, we argue, are incompatible with default inflection through the kind of symbolic rule envisioned by Marcus et al. (1995). Crucially, we demonstrate that the various circumstances in which the rule applies display *differing levels of productivity* of the irregular forms, a phenomenon which is inconsistent with a system built around a single unifying symbolic rule.

Finally, leaving behind Marcus et al.'s two central arguments, the remainder of the article provides novel evidence against the dual-route account in the form of direct quantitative comparisons of matched single-route and dual-route models on experimental data. Again, single-route performance is found to be superior.

To prepare the ground for these three main parts, we begin with a more detailed exposition of the dual-route account.

THE DUAL-ROUTE ACCOUNT

The dual-route account (Pinker & Prince, 1988; Pinker, 1991; Prasada & Pinker, 1993; Marcus et al., 1995) modifies the traditional account of morphology developed within generative linguistics. The traditional account assumed a rule for regular forms in addition to rote memory storage of irregular forms; alternatively, in an attempt to do justice to common structure present among irregulars, irregular forms were also thought of as governed by a set of ''minor rules'' (Chomsky & Halle, 1968). Both variants of the traditional story failed to capture important facts about irregular forms. Irregular English past tense forms such as *ring*/*rang* and *sing*/*sang* can form the basis for analogous irregularization of a phonologically similar nonce word (e.g., *splang*) (Bybee & Moder, 1983; Kim, Pinker, Prince, & Prasada, 1991); such generalization is incompatible with rote memory storage — a fixed set of irregular lexical items — alone. The productivity of irregular forms necessitates some additional mechanism. ''Minor rules'' could provide this productivity, but they fail to explain the *gradedness* of irregular productivity: production of irregular past tenses for English nonce words seems to vary continuously as a function of similarity to prototypical patterns; the less phonologically similar a nonce word is to a prototypical pattern, the less participants are likely to irregularize it (Bybee & Moder, 1983).

Pinker's (1991) dual-route account maintains a rule as the basis for regular forms, while trying to provide a more adequate basis for irregulars. Lexical memory, which stores irregular forms, is supplemented with an *associative component* that allows generalization on novel words. This associative component links the phonological form of the various lexical entries in appropriate ways. Connectionist pattern associators, such as the Rumelhart and McClelland (1986) model, are viewed as ''systematic implementations'' (Marcus et al., 1995, p. 194) of such an associative memory.

The dual-route account also maintains the general relationship between memory and rule as specified in the traditional ''rote memory + rule'' account: the rule is used as the ''default'' whenever memory ''fails.'' The only difference is that memory retrieval need not yield exact matches in order for rule application to be blocked; the associative memory can generate irregular responses that are sufficiently strong to suppress the rule on novel words as well.

The rule is assumed to be *symbolic* because the diverse set of circumstances in which it can be applied have only their membership in a particular syntactic category, e.g., ''Noun'' or ''Verb,'' in common. Thus the rule is assumed to *operate over a symbol representing syntactic category;* for example, something like $\mathbf{V} \rightarrow$ *-ed,* i.e., ''the suffix *-ed* can be concatenated to

any verbstem **V**'' in the case of the English past tense. This rule is a *mental representation* actively (but unconsciously) invoked by the speaker when producing a regular form. Because regular forms are based on such a general, symbolic rule, the regular suffix can be productive regardless of similarity or dissimilarity to known regular words. The default rule can apply not only to unusual sounding, novel words, but also in a variety of circumstances which, in principle, are not lexicalized and for which memory consequently must fail such as quotations (e.g., ''there are 10 'the's in the text'').

The dual-route account contrasts with single-route connectionist models such as those of Rumelhart and McClelland's (1986) or Plunkett and Marchmen (1991, 1993) which produce regular and irregular inflection in one uniform architecture. The central idea in these models can be generalized into what Marcus et al. (1995) call the ''Pattern Associator Hypothesis,'' the claim that ''the memory mechanisms uncontroversially needed to capture irregular patterns serve for regular patterns as well'' (p. 195). The dual-route account differs from the Pattern Associator Hypothesis both in the insistence on separate routes for regulars and irregulars and the claim that these embody different styles of computation—symbolic and associative.[1]

The strongest kind of evidence for the dual-route account does not merely make the case for two different routes as does experimental evidence that regular forms lack standard properties of lexical items such as priming or frequency effects (Clahsen, in press; Seidenberg & Bruck, 1990, though for contrary findings see Stanners, Neiser, Hernon, & Hall, 1979; Rueckl, Mikolinski, Raveh, Miner, & Mars, 1997; Taft, 1979; Sereno & Jongman, 1997; Alegre & Gordon, 1999). Rather, the strongest evidence for the dual-route account focuses directly on the symbolic nature of the default rule, supporting the existence of two routes as a corollary. Marcus et al. (1995) provide two interrelated arguments for the symbolic nature of the regular or ''default'' route.

For the first, Marcus et al. establish that there is a wide range of exceedingly diverse circumstances in which the default applies, including unusual-sounding words, borrowings, names, quotations, acronyms, or derivations from other grammatical categories (e.g., *spit the pig*/*spitted the pig* as opposed to *spat the pig*). These heterogeneous circumstances—of which Marcus et al. (1995) list 21 in total—are given a single, unifying, and, hence, extremely parsimonious explanation through the assumption of a symbolic

[1] The claim that the rule route is symbolic, as outlined, boils down to the idea that *+ed* is associated with an equivalence class *verb*. All verbs are equally eligible; no further criteria are required. This concept could be implemented in a connectionist architecture (see Marcus et al., 1995); however, this would be a case of ''mere implementation'' where the (faithful) implementation adds nothing to the symbolic description, unlike standard connectionist implementations of morphology or the dual-routes' pattern-associator component, which have no corresponding simple, high-level symbolic description.

default rule as the basis of their production. In the following, we refer to this as the ''*unification and parsimony argument*.''

The second argument in favor of the symbolic nature of the rule route builds on this default behavior and argues that central properties of pattern associators rule them out as suitable alternative mechanisms for default inflection. The two critical properties are *generalization based on similarity* and *sensitivity to the statistics of the input* (i.e., the ''learning set''). In standard connectionist networks such as the two- or three-layer feedforward networks used by Rumelhart and McClelland (1986) or Plunkett and Marchman (1991, 1993) similar inputs tend to elicit similar outputs. Consequently generalization to novel forms is strongly determined by representational overlap—and, in this sense, similarity—between the novel input and the training items. This also forms the point of contact between connectionist models of inflectional morphology and the variety of broadly similarity-based schema accounts which have been put forth in the linguistic literature (Köpcke, 1988, 1993; Bybee, 1995). Empirical evidence for the symbolic account is obtained if it is shown that novel forms are regularized independently of similarity to known regulars. This has been the focus of ongoing experimental research (Prasada & Pinker, 1993; Lee, 1996; Hahn, Nakisa, Bailey, Holmes, Kemp, & Palmer, 1998). Marcus et al. (1995) add to this research their list of 21 circumstances, which they argue do not form a cohesive similarity space.

One way in which single-route pattern associators might nevertheless achieve the wide-ranging productivity of regular forms is through the statistical dominance of regular forms. As statistical devices, connectionist networks are highly sensitive to the statistics of the training set. Marcus et al. (1995) evaluate earlier connectionist models (e.g., Rumelhart & McClelland, 1986; Plunkett & Marchman, 1991, 1993) to argue that suitable levels of regular productivity in such models require that regulars constitute a majority in terms of *type frequency*. ''Type frequency'' refers to the proportion of words of a given class or type within the entire set of relevant items, for instance, the proportion of English verbs taking a regular past tense. Because defaultlike productivity requires a dominant type frequency of the regular form, standard connectionist networks should not be able to produce the desired ''default behavior'' for languages in which the regular form is not actually the most frequent. Because the regular or ''default'' type is defined by Marcus et al. (1995) as the type which is ''freely generalizable'' (p. 216), i.e., applicable wherever lexical memory fails, there is no need for this regular type to also be the most frequent. In English, the default *+ed* is the most frequency type by far, with 95% of all verbs, but dominant type frequency and default need not coincide. A dissociation is present in so-called ''minority default'' inflections. Marcus et al. (1995) provide two minority default systems as evidence for the dual-route account: the German participle and

the German plural. The existence of these systems, so Marcus et al. (1995), is compatible with a symbolic rule, but not with the Pattern Associator Hypothesis:

> . . . under the rule theory, majority status and default status are psychologically independent. If they correlate it could be because there were historical events that put rule-generated forms in the majority. The crucial prediction, then, is that there should be languages in which the default inflection is not in the majority. The pattern associator alternative predicts that that should not be possible. (p. 216)

We refer to this second argument as the ''*minority default argument*.'' It rests not only on the conjectured behavior of single-route pattern associators, but also on the assumption that the dual-route account can cope with minority default inflections. But neither the dual-route account, which has never been implemented, nor a single-route pattern associator have actually been tested and compared on a sufficient sample of actual German. Predicting model performance is a notoriously tricky affair. Thus, tests of these assumptions through actual simulations seem vital.

In the following section we test the soundness of the minority default argument with a range of simulations. The necessary implementation decisions for the dual-route models then forms the basis for a subsequent examination of the unification and parsimony argument.

TESTING THE MINORITY DEFAULT ARGUMENT

How Can We Compare Single- and Dual-Route Models?

The first issue we must address is choice of model(s). The Pattern Associator Hypothesis provides only a minimal constraint. Even committing to ''connectionism'' the potential set of models is vast, as connectionism implies little more than that processing should proceed via massively interconnected simple units, possibly in parallel. This problem of choice also affects the dual-route accounts pattern associator component.

In this situation, it seems desirable to test *several* simple models. A suitable selection will also enable us to untangle the effects of similarity and type frequency, which are confounded in the conjectured behavior of connectionist models on this task. All models will be implemented in a single-route and a dual-route version, which are exactly matched in all respects except the explicit tenets of the dual-route account: the existence of two routes, one of which is a symbolic rule. Such matched comparisons will allow direct identification of the dual-route contribution to model performance.

The next issue is evaluating model performance. A natural measure of success is the models' capacity to produce correct forms. However, single- and dual-route accounts do not inherently predict different outcomes on all tasks. For *familiar* words, i.e., those already acquired by the model, the dual-

route model specifies the following process: lexical memory is searched for an irregular entry; if this look-up procedure fails, the regular form generated via the rule route is used. Barring noise in memory look-up, this procedure is guaranteed to produce the correct plural for any correctly acquired form. This level of competence can trivially be matched in a single-route system; the single-route commitment is that irregular and regular forms are produced in the same way and simple look-up in lexical memory can be used to produce both irregular and regular forms. Again, all correctly acquired forms are (trivially) correctly reproduced. In other words, there is no general prediction of level of performance difference between single-route and dual-route accounts with respect to ''familiar'' words.

Dual-route and single-route accounts do, however, give rise to different predictions when it comes to *generalization,* i.e., behavior on novel, unknown words. According to the dual-route account, ''regular generalization'' (production of novel regular forms) is based on the rule and proceeds independently of known regulars, whereas single-route accounts base regular generalization on known regular forms. This difference in procedure should lead to different outcomes for at least some novel words, more or less regardless of how single-route generalization is fleshed out.

Consequently, generalization accuracy seems a desirable criterion for the model tests. This requires sufficient amounts of representative generalization data. One possible source would be experimental data of human nonword generalization. However, there are no large nonword generalization data sets. Extant data (Russ, 1989; Köpcke, 1988; Marcus et al., 1995), presented with sufficient resolution to be used for modeling (this excludes e.g., the data of Russ, 1989, and Köpcke, 1988), currently takes the form of small data sets, which are highly specific in the possibilities they sample. It is also unclear to what extent the experimental procedures with which these data are derived successfully tap into the normal workings of the cognitive system. For instance, do rating tasks, which are far removed from everyday language use, provide sufficiently veridical measures? Though we later describe modeling of experimental data, the simulations aimed at the minority default argument seek to avoid these difficulties by using the *extant lexicon* itself, i.e., real German words.

To test our models, we withhold some proportion of the extant lexicon from the model's training phase and use these novel, previously unseen, items to test model generalization. The ''correct'' generalization response against which the models are assessed is assumed to be the form the word actually has in the language. It is worth noting that this task has a natural counterpart in language acquisition, which involves a steady build-up of the lexicon; in this sense, generalization from known to newly encountered real lexical items of a language is an actually occurring, natural task.

With these general motivations in place, we proceed to the details of our simulations.

The Task

The data set. To provide a realistic test it is desirable to conduct large-scale simulations with real language. Our data set was drawn from the 30,100 German nouns in the CELEX database.[2]

Since the CELEX classification is fraught with error, we automatically classified nouns according to the nature of the transformation from singular to plural phonology. In other words, plural classes were defined by the transformation required to generate the plural form[3] from the singular[4] or ''stem.'' Four general types of transformation occur:

1. Identity mappings. Here, there is no change between singular and plural: for example, *Wasser* → *Wasser* or *hit* → *hit* for a corresponding example from the English past tense.

2. Suffixation. A suffix is added to generate the plural: e.g., *Kind* → *Kinder* or *walk* → *walked* in English.

3. Vowel change. The plural is indicated through a word-internal vowel change, in the case of the plural a so-called ''umlaut'': e.g., *Mutter* → *Mütter* or *sing* → *sang* for a rough English analog.

4. Rewriting of the final phoneme(s). Here, a segment is deleted and a suffix attached: e.g., *Thema* → *Themen,* a transformation which applies to words of Greek and Latin origin (*datum* → *data,* in English).

Vowel change and suffixation can also occur together, giving rise to a number of combinations (e.g., *Hut* → *Hüte*).

This transformation-based classification yields approximately 60 categories (some of which contain only one member). We then discard categories with a type frequency of less than 0.1%, resulting in a database of 24,640 nouns with 15 different plural categories (see Table 1).[5] This step removes primarily Latin and Greek words and a small number of German words with arbitrary plurals (suppletion, or singly occurring transformations). In effect this brings our classification into accord with the plural types described in standard linguistic analysis (Köpcke, 1988). The only further amendment in this direction was that the umlauts (ä, ö, and ü) were treated as one, as is standard in the literature.

This data set still contains a large element of redundancy, due to the fact that German allows very liberal compounding (i.e., ''wood pigeon'' would be a single word in German). Because the plural of a compound noun is determined exclusively by the rightmost lexeme, compounds add nothing and, in fact, could artificially boost model performance: if, for instance,

[2] CELEX can be obtained by contacting celex@mpi.nl.

[3] More precisely, the nominative plural because German contains overt case marking.

[4] More precisely, the nominative singular.

[5] Forty-five words and two duplicates were manually removed because they were obviously incorrect (e.g., incorrectly pluralized proper names and entries with errors in phonological form).

TABLE 1
Frequencies of Different Plural Types in the Complete Set of Nouns in CELEX and for the Noncompound Nouns[a]

	All nouns		Noncompound nouns	
Plural type	Frequency	% of total	Frequency	% of total
+ən	7012	28.109	2646	30.775
+n	4477	17.947	1555	18.086
+ə	4460	17.879	1178	13.701
Identity	4201	16.840	1992	23.168
Umlaut + ə	2017	8.085	239	2.780
+s	978	3.920	571	6.641
Umlaut + ər	692	2.774	54	0.628
+ər	289	1.159	36	0.419
Umlaut	255	1.022	35	0.407
ʊm → ən	135	0.541	95	1.105
a → ən	121	0.485	81	0.942
ʊs → ən	88	0.353	69	0.803
ʊm → a	45	0.180	40	0.465
+tən	27	0.108	1	0.012
+iən	25	0.100	6	0.070

[a] Suffixation is indicated by +suffix; rewrites are indicated as ''phonemes'' → ''phonemes.''

''table'' was one of the few words a model got right, actual performance would be masked by the additional, unremarkable, success on ''kitchen table,'' ''dining-room table,'' and so on. Thus, the set of 24,640 was further reduced to a set of 8,598 ''noncompound'' nouns. A ''noncompound'' noun was defined as a noun that did not contain another noun from the database as its rightmost lexeme. This leaves complex nouns which are not noun-compounds and noun-compounds for which the rightmost lexeme is not listed individually. In fact, this reduction seemed to have little impact on performance either way, as the performance of the nearest neighbor classifier (see below) on the entire data set and the noncompound subset were virtually identical with 72 and 71% respectively (Nakisa & Hahn, 1996).

The plural types and their respective type frequency are shown in Table 1. For the purposes of this article we assume that the s-plural is, in fact, the ''regular'' form as Marcus et al. (1995) claim, though this is by no means universally agreed upon. As can be seen from the table, the s-plural, the putative default, constitutes an extreme minority, applying to only 6.6% of the words in the data set.

Input representation. In all simulations, the input to the models is a representation of the phonology of the singular form or ''stem.'' The actual phonological representation is impoverished, though standard for this type of

```
VCVCVCVCVCVCVCVC VCVCVCVCVCVCVCVC
________&p_kOm@_ ________&p_kOm@n
______&lot_ri_a_ ______&lot_ri_a_
______A_8s_l|z@r ______A_8s_l|z@r
__________g@bYS _________g@bYS@_
_______k_v&s_t@_ _______k_v&s_t@n
_____________rat __________r)t@_
```

FIG. 1. Example of input representation before conversion of phonemes into feature bundles.

modeling, in that it is limited to binary phonetic features. Phonemes are represented as a bundle of 15 phonetic features taken from the linguistics literature (Wurzel, 1981).

Specifically, 16 vowel/consonant slots were used. Since words vary in length, their representations had to be zero padded. Words were right justified since word endings are most salient for determining the plural type of German nouns.[6] For an example see Fig. 1.

Phonemes were then converted into feature bundles, yielding a 240-element vector for each word.

No other information was included in the input. This leaves out syllable or stress information. It also excludes several potentially relevant nonphonological factors such as the gender[7] of a word, its token frequency (the frequency with which it is encountered), or its semantics. However, the dual-route account contains no commitment on these issues for its own pattern associator component. Thus opting for the most simple approach seems neither an undue nor an unfair restriction and, crucially, because our single- and dual-route models will be matched, both are subject to the same limitations.

The generalization task. The task for all models was to produce the appropriate plural class for previously unseen singular forms. An alternative would be to have the model produce a word's actual plural form. We chose not to produce output forms directly because two of the models we used are classifiers. Furthermore, for a connectionist network, producing an actual form introduces an additional difficulty in the form of the ''alignment problem.'' This is a general technical problem with reproducing sequences and is not specific to morphology (for discussion and solutions see Bullinaria, 1997). As the problem affects both single- *and* dual-route models because of the

[6] This was determined by comparison of performance of left-justified, center-justified, and right-justified words using ID3 (Quinlan, 1992).

[7] All German nouns have one of three possible genders, feminine, masculine, or neuter. Gender is made apparent through the accompanying article and is also marked on adjectives, e.g., ''das grosse Haus'' (the big house) and ''der grosse Mann'' (the big man).

dual-route account's pattern associator component, it is fair to leave it out. Most importantly, because of the way our plural classes are defined in terms of surface transformations, all the *information* required to generate actual forms is present in the ''class'' (and, in fact, the scripts which automatically generate the plural class from a singular/plural pair could readily be adapted to run ''in reverse'' to generate a plural from a singular and its plural category). Given that we have no idea whether the cognitive system ultimately maps directly, via a class decision or proceeds in a different matter altogether, this seems sufficient.

To ensure a robust estimate of generalization performance we used cross-validation (Breiman, Friedman, Olshen, & Stone, 1984). The noncompound set was randomly divided into 10 subsets. Classification accuracy was iteratively tested for each subset, given the remaining 9 subsets as the model's lexical knowledge. In other words, for each model, we ran 10 simulations in which we tested generalization accuracy on a subset of roughly 860 randomly selected nouns, having trained the model on the remaining 7740. In this way, every noun was treated as a novel exemplar once; for each test item the predicted plural class was compared with the novel exemplar's actual plural class. The model's overall generalization performance was taken to be the mean proportion of items correct across all 10 test sets.

Single-Route Models

We implemented three different models. All three models—the nearest neighbor algorithm, Nosofsky's Generalized Context Model, and a three-layer back-propagation network—are well-known, standard models for classification, not custom-built approaches to inflectional morphology, and their behavior is well understood. All models solve the generalization task ''on the basis of known words,'' but differ in their ability to exploit similarity, frequency, and potential higher order features and regularities in the training set. Hence they allow us to disambiguate these various factors confounded in standard connectionist pattern associators and, consequently, in the minority default argument.

The nearest neighbor algorithm. The nearest neighbor algorithm is probably the simplest classifier imaginable. A new item is simply assigned the same class as the known item to which it is most similar. In other words, the nearest neighbor algorithm defines an extremely basic exemplar model. Encountered items and their class label are individually stored in memory. To classify a novel item, the most similar item in memory is determined and its classification adopted.

In our application, the known items are the words of the training set in a phonological representation; these constitute the model's lexical memory. According to their phonology, these words can be thought of as each occupying a distinct point in ''phonological space.'' A new word is classified the same way as its nearest known neighbor in phonological space, where

distance between items is determined according to a standard Euclidean distance metric.[8]

Consequently, the performance of the nearest neighbor algorithm is determined *entirely* by the similarity structure in the data set. It does not abstract summary information from its input, whether this be central tendencies, frequency information, or critical features. The nearest neighbor algorithm can be thought of as a simple ''structure mirror'' which reflects regularities implicit in the similarity structure without extracting these. Note also that because the algorithm adopts the classification of the nearest neighbor without any limit on how near that neighbor must be, it *always* provides a response no matter how distant (i.e., unusual sounding) an item is from known items. Ties, if they occurred, were broken randomly in our simulations.

Despite its simplicity, the nearest neighbor algorithm achieved a predictive accuracy of 70.8% (SE = 0.58%, $n = 10$) for the 8598 noncompound words.

The generalized context model. Nosofsky's Generalized Context Model (GCM) (Nosofsky, 1986, 1988b) is a slightly more sophisticated exemplar model. Though it is possibly the most successful model of categorization when it comes to fitting human performance data, it has never been applied in a linguistic context. It differs from the nearest neighbor algorithm both in its similarity assessment and its probabilistic response rule.

Roughly the GCM can be thought of as basing its response not on the single nearest neighbor but on the known items within a sphere surrounding the novel item (strictly speaking this sphere has no sharp boundaries, but the contributions of distant exemplars rapidly falls off). Within this sphere the model takes a ''majority vote.'' The strength of an individual ''vote'' depends on the degree of similarity to the novel item. In other words, the classification decision is not exclusively based on similarity, but is a joint function of similarity and type frequency. This can be seen by considering the response rule of the model (in the text below) in the two limit cases in which either all items are equally similar or where all type frequencies (class sizes) are equal. In the case of equal similarities, the response rule would ensure that decision is based on type frequency, whereas in the case of equal type frequencies the decision is based on similarity. Between these extremes, model behavior is a product of both.

Unlike the nearest neighbor, the GCM is also not restricted to a single response. It determines a probability for each possible class or category. This allows the model to be given a probabilistic or a deterministic interpretation. In these simulations, we took the model's response to be the plural class with the highest response probability. As with the nearest neighbor, the model's lexicon is built up by simply storing the relevant training set in memory.

To describe the GCM and its similarity metric formally: the strength of making a category J response (R_J) given presentation of stimulus i (S_i) is

[8] The square root of the summed dimensional differences.

found by summing the (weighted) similarity of stimulus i to all presented exemplars of category J (C_J) then multiplying by the response bias for category J. The denominator normalizes by summing the strengths over all categories:

$$P(R_J|S_i) = \frac{b_J \sum_{ij} c_J L(j, J)\eta_{ij}}{\sum_k b_K \sum_{k\varepsilon} c_K L(k, K)\eta_{ik}} \quad (1)$$

In Eq. (1) η_{ij} ($\eta_{ij} = \eta_{ji}$, $\eta_{ii} = 1$) gives the similarity between exemplars i and j, and b_J ($0 \leq b_J \leq 1$, $\sum b_K = 1$) is the bias associated with category J, though bias terms were omitted in our simulations to limit the number of free parameters. $L(j, J)$ is the relative frequency (likelihood) with which exemplar j is presented during training in conjunction with category J, which, again, was not manipulated in these simulations. The (Euclidean) distance, d_{ij}, is converted to a similarity measure using the transformation $\eta_{ij} = e^{-sd_{ij}^p}$, where the free parameter s scales the rate of decay; $p = 1$ yields an exponential decay similarity function and $p = 2$ gives a Gaussian similarity function.

We used a Gaussian similarity function in the following, as the first simulations with this model, described in Nakisa and Hahn (1996), had yielded slightly superior performance of the Gaussian.[9]

In the simulations, the GCM model performed better than the simple nearest neighbor. The simulations with the parameter s set to the best fitting value of $s = 1.42$ yielded classification accuracies of 74.3% ($SE = 0.40\%$, $n = 10$) on the noncompound data set.

Back-propagation network. Our final model is a connectionist network. This is Marcus et al.'s (1995) own suggestion for the pattern-associator component of the dual-route account. We chose a standard, three-layer back-propagation network. The network's output layer consists of 15 units, each a localist representation of one plural class. Thus, like the GCM, the network gives a graded category response to the input pattern. Analogously, we deemed the plural class with the highest activation to be networks response. While the nearest neighbor and the GCM are restricted to surface similarity between the phonological forms of the data set, the network can utilize its hidden layer to rerepresent the input space as desired. Directly adjacent points in input space may be mapped onto widely different regions of internal representation, while — conversely — widely separated regions of input space may map onto adjacent regions of internal representation. Consequently, the network's judgment need not be based on surface similarity; criterial features can be extracted from the input (e.g., ''all words starting with the phoneme /b/'' are class x) and so can relevant higher order statistics.

[9] When the scaling parameter s, which governs the rate of the decay and, hence, sensitivity to more distant neighbors, was optimized for the Gaussian similarity function ($\eta_{ij} = e^{-d_{ij}^2}$) the performance was 75.0% ($s = 1.46$). When optimized for the exponential ($\eta_{ij} = e^{-d_{ij}}$) accuracy was 74.4% ($s = 0.35$; Nakisa & Hahn, 1996).

To ensure robustness of results, simulations were conducted with nets of different sizes, different initial random seeds, and different amounts of training with the training set, with final performance always measured at the point of best performance.

We successively trained and tested networks with 10, 20, 30, 40, and 50 hidden units (though we experimented with values as extreme as 2 and 1000 for control). Weight update was incremental, and performance was tested in regular intervals between 10 epochs and 100 epochs, sampling in steps of 10. Performance on a given test item was scored as correct if the plural class with the highest level of activation matched the actual class of the test item. Each of the networks were subjected to the full cross-validation procedure, using different initial random seeds for all 10 cross-validation training sets. Final results reported are, again, the mean number of items correct across all 10 cross-validation splits.

The network's ability to rid itself of surface similarity yielded an increase in performance over both nearest neighbor and GCM. The best single-route network scored 82.7% ($SE = 0.40\%$, $n = 10$) on the noncompound data set with 40 hidden units and 50 epochs of training.

Dual-Route Models

Implementation of the corresponding dual-route models requires making the interaction between the two routes computationally explicit. Implicit in the dual-route account is a *threshold* governing route interaction. If the strength of the memory response rises above a certain threshold, the rule route is blocked. If the response is below threshold, the default rule is used. This threshold stems from the fact that memory responses are not all or nothing but a matter of degree, which in turn is a logical consequence of the fact that the memory component does not just involve rote look-up but allows generalization for novel forms. *Spling* might be inflected as *splung* because the pattern associator component produces a sufficiently strong irregular response on the basis of *spring* and *string*. But *how strong* is ''sufficiently strong''? There must be some threshold above which a memory response is strong enough to block the rule and, conversely, below which memory is taken to have ''failed'' and the default is used.

An implementation of the dual-route account must make this threshold computationally explicit. The most natural approach is based on the idea of a *certainty criterion*. The rule route is blocked *if and only if* the irregular route is sufficiently ''certain'' of its response, be it that lexical memory contains an exact entry or that the item—though novel—is sufficiently similar to a stored item(s) to elicit a strong irregular response. This idea of response certainty manifests itself in slightly different ways in each of the three models.

For the nearest neighbor, ''certainty'' can be based on the distance of the nearest known neighbor. The closer the nearest neighboring item in memory,

the more confidence can be attributed to the classification; as distance increases, certainty drops. The GCM's output is the probability that the item belongs to a certain category; ''certainty'' can be based on this probability. The more probable a category, the greater the certainty of the response. The network, finally, indicates ''certainty'' through the level of activation for a given output unit, i.e., class. The higher the activation of the output unit representing a given category, the greater the certainty.

We thus formally specify the interaction between the two routes as follows:

Nearest Neighbor Algorithm

Memory fails if the nearest neighbor in phonological space is at a distance greater than the certainty threshold, t. In this case the default inflection is used.

if distance($\mathbf{e} - \mathbf{n}$) $< t$ inflect **e** as neighbor **n.**
Otherwise use default inflection.

Nosofsky's GCM

Memory fails if the largest class probability, P_j, is less than the certainty threshold value.

if MAX($\mathbf{P}_J$) $> t$ inflect as most probable class.
Otherwise use default inflection.

Neural Network

Memory fails if the greatest output unit activity, MAX(o_i), is less than the value of the certainty threshold, t.

if MAX(o_i) $> t$ inflect as class of most active unit.
Otherwise use default inflection.

Our dual-route models, then, are simply the three single-route models with the additional default rule (inflect as $+s$) which is used when certainty drops below the threshold, t.

The correct value of t, which might vary between languages and also conceivably between speakers, we took to be an empirical question. Thus, we tested *all possible values* of t, i.e., we sampled throughout the entire interval ($0 < t < 1.0$ for GCM and network and $0 < t < \infty$ for the nearest neighbor).

The dual-route models were tested on exactly the same test sets as the single-route models. They were also ''trained'' on the same training sets as the corresponding single-route models *except* that regulars were removed from the training set. This removal is necessitated by the dual-route account's claim that ''use of the regular affix does not depend on stored forms'' (Marcus et al., 1995, p. 196, second paragraph). According to the account, regular inflected forms are *generated by* the rule rather than *retrieved from* lexical memory. The only way to ensure this in the dual-route models is to remove all regularly inflected forms from the associative lexical memory, which

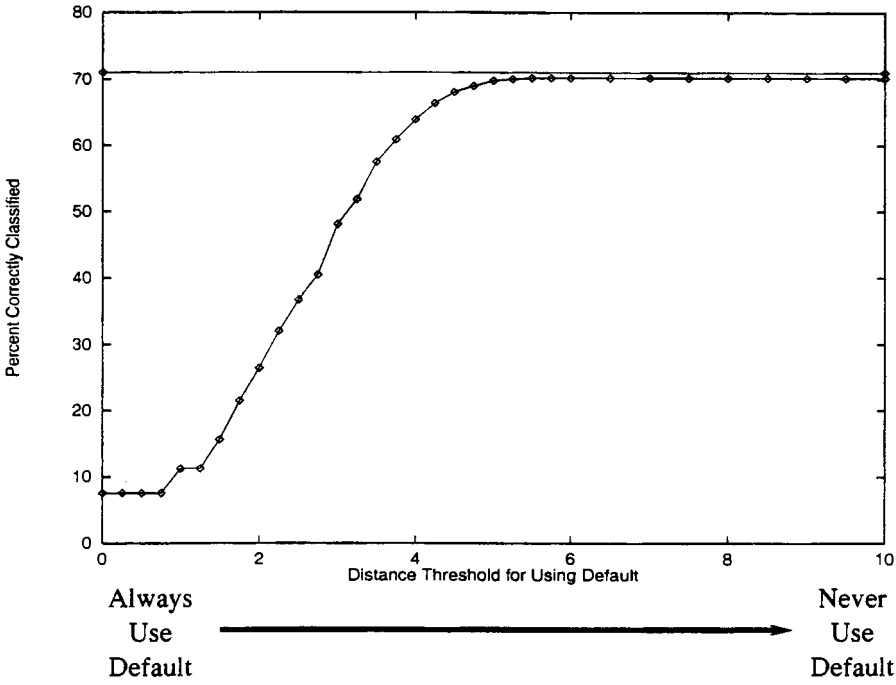

FIG. 2. Plot of the dual-route nearest neighbor's classification performance as a function of the value of the distance threshold, t. For comparison, single-route performance is indicated by the horizontal line at 70.8%. For $t = 0$ the model always uses the rule. Increasing the level of t increases the contribution of the associative component. The optimum value of t is ∞, i.e., the optimum is *never* to use default.

means removal from the training set for this associative memory component.[10]

Dual-route nearest neighbor. Rather surprisingly, there is no value of threshold t at which dual-route performance exceeds single-route performance for dual- and single-route nearest neighbor. The comparisons are plotted in Fig. 2. Single-route performance is indicated by the horizontal line, dual-route performance is the curve indicating performance for a given value of t.

Varying t affects the model's behavior in the following way. If t is very low, which means that distance to the nearest known neighbor must be very

[10] That the removal of the *s*-plurals from the ''training set'' which constitutes the models' lexical memory is directly necessitated by Marcus et al.'s (1995) exposition of the dual-route account is overlooked by Clahsen (1999), who seems to feel that this puts dual-route models at a disadvantage, presumably because they have not had any chance to learn how to generate *s*-plurals appropriately. In fact, no such learning is necessary, as the symbolic route has this knowledge built in.

small for the irregular response to be ''certain'' enough to block rule application, then no novel irregular will ever be similar enough and no irregular generalization will arise; all previously unseen forms will be regularized. If t is very high, some irregular will always be similar enough and memory will always block rule application; thus no regulars are produced. Always using the default gives the percentage correct, which corresponds to the type frequency of the s-plural, roughly 6%; never using the default (i.e., $t = \infty$) corresponds to a single-route model without the ability to produce regular forms. Between these extremes, the dual-route model uses both routes and produces both regulars and irregulars, with the exact mixture depending on the exact value of t. The striking result of the simulations is that gradually increasing the efficacy of the rule route (by lowering t) first has no effect and then leads to a monotonic *decrease* in performance as the model generates more and more false positives (i.e., overgeneralizes the $+s$ form). Rule use fails to increase generalization accuracy.

Dual-route GCM. The picture is similar for the GCM: again the single-route model remains better than the dual-route model throughout, as plotted in Fig. 3. The dual-route GCM's performance at its optimal value of t is

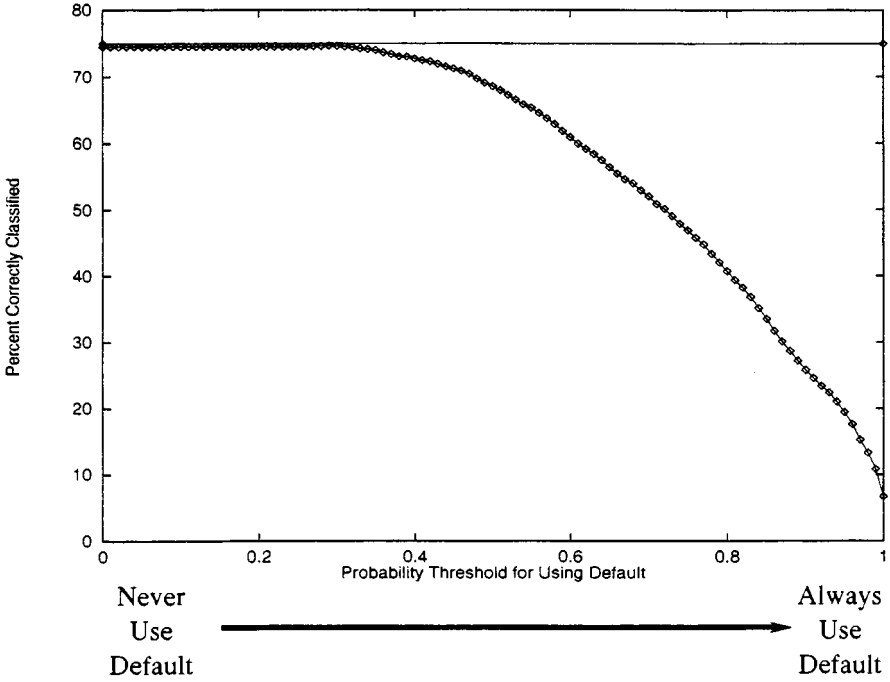

FIG. 3. Comparative plot of single-route GCM performance (the horizontal line) and dual-route GCM performance at the various levels of threshold t.

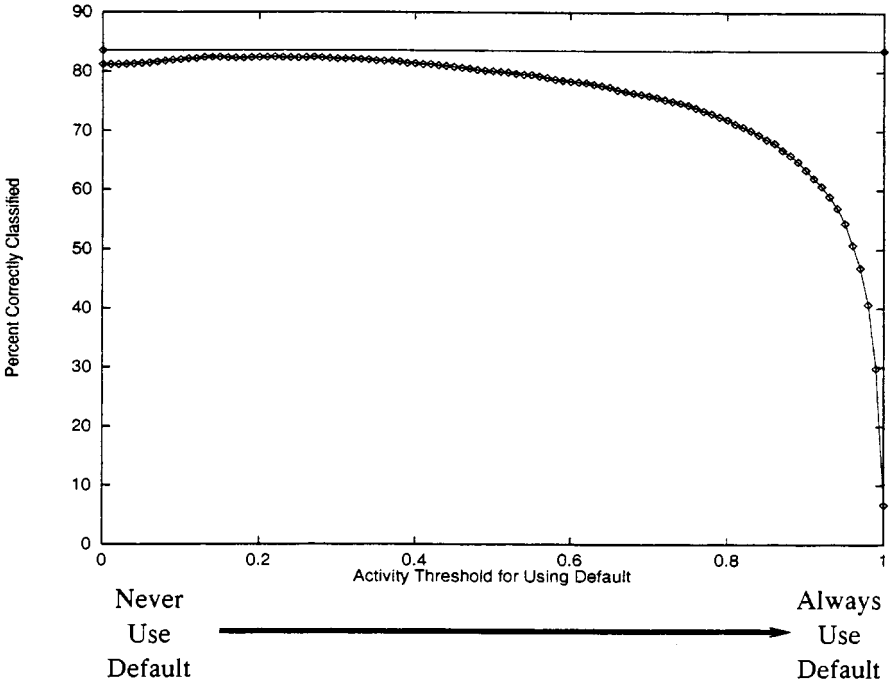

FIG. 4. Plot comparing generalization accuracy of the single-route network (horizontal line) and the dual-route network at various levels of the threshold *t*.

73.8% (*SE* = 0.34%, *n* = 10); that is, below the accuracy attained by the single-route GCM.

Again, increasing the ''certainty threshold,'' i.e., increasing the probability that an irregular response must attain for memory to be ''certain enough,'' gradually increases the efficacy of the rule route. This does not boost overall model performance, but merely gradually produces more and more false positives.

Dual-route neural network. The dual-route results for the network, finally, complete the picture in that, again, single-route performance actually marginally exceeds dual-route performance rather than being inferior as predicted by the minority default argument. The optimum score of the rule plus network model is 81.2% (*SE* = 0.24%, *n* = 10). The comparisons for all possible values of the threshold *t* are plotted in Fig. 4.

Error analysis. Model errors can shed further light on the models' behavior. Given the large scale of these simulations, a detailed analysis is beyond the scope of this article, but some general comments and exemplary analyses can be provided.

The pattern of errors in the single-route models is basically the same for all three: there is an interaction between type frequency and class structure.

Generally, performance declines with decreasing type frequency, but some low-frequency classes can nevertheless be classified quite accurately. This can be seen clearly in Fig. 5, which plots the error proportions for each plural for the single-route GCM. Plural class 12 (final /us/ to /en/), constitutes only 0.8% of all items, with 69 words total. Nevertheless, it is one of the best performing plural classes. The default +/s/ (plural class 6) does comparatively poorly, with only 14.5% of its 571 members predicted correctly. Also informative is the *nature* of the errors. As seen in Table 2, this too reflects the interaction between type frequency and similarity that determines GCM performance. The erroneously predicted plural classes are not simply the most frequent types; there is no simple linear relationship between type frequency and degree of overgeneralization, and all but the most infrequent plural class, which comprise only a single item, are overgeneralized at least once.

The network simulations deviate from this pattern insofar as the five lowest frequency classes are highly sensitive to initial random seeds with the consequence that performance on these items can vary drastically between networks.

The dual-route patterns are generally very similar. As can be seen from the corresponding figure and table (Fig. 6 and Table 3), however, the dual-route model with the best overall performance *never* uses the rule, with the consequence that *no* +s-plurals are produced. The dual-route GCM is thus even further from producing the desired default regularization than the single-route GCM.

Summary. The simulation results are summarized in Table 4.

Our three simple models—nearest neighbor, GCM, and back-propagation network—show rather remarkable accuracy at predicting plural class on the basis of word phonology, tested on a large-scale data set. The models' predictive accuracy is surprising given that the seemingly highly irregular German plural system is commonly described as highly arbitrary. For example, Marcus et al. (1995) describe the pairing of plural form and individual word as ''to varying degrees arbitrary'' with ''some correlations between plural form and the gender and morphophonology of the root, though like English past tense forms, they defy simple summary'' (p. 226). Koepcke (1993) states that all known descriptive generalizations such as Augst (1979) or Mugdan (1977) only summarize tendencies which are subject to long lists of exceptions.

A genuinely arbitrary system, however, would not allow successful prediction. The fact that approximately 80% of all German plural forms *can* be predicted by our models reveals the system to be highly structured.

The performance of the two simple exemplar models, nearest neighbor and GCM, is particularly interesting, as this type of model has not previously been considered in the linguistics literature at all. Both constitute what Koepcke (1993) has called ''weak analogy models.'' Even for the proponents of

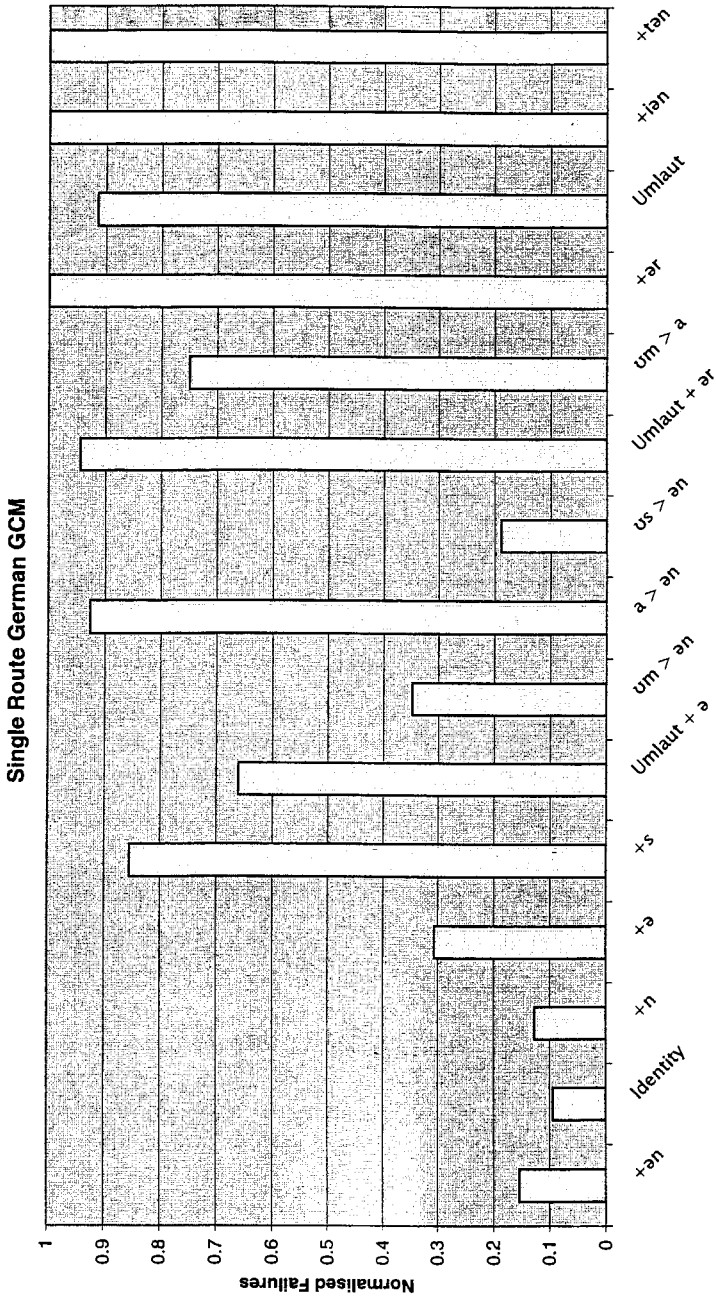

FIG. 5. A plot of the proportion of items incorrectly predicted by the single-route GCM for each plural class. A maximum value of 1 indicates that all words in that plural class were misclassified.

TABLE 2
The Single-Route GCM's Errors Indicating the Way Words in Each Plural Class Were Misclassified[a]

	Falsely predicted values															
Class	+ən	Identity	+n	+ə	+s	Umlaut + ə	ʊm → ən	a → ən	ʊs → ən	Umlaut + ər	ʊm → a	+ər	Umlaut	+iən	+tən	Total for class
+ən	—	107	87	185	10	8	5	1	1	1	1	—	—	—	—	406 of 2646
Identity	60	—	70	34	14	2	—	1	1	1	—	—	1	—	—	187 of 1992
+n	20	156	—	2	9	—	2	6	—	—	—	1	1	—	—	197 of 1555
+ə	143	118	37	—	36	18	2	—	2	—	—	4	—	1	—	361 of 1178
+s	60	57	234	124	—	5	3	4	—	1	—	—	—	—	—	488 of 571
Umlaut + ə	13	4	3	127	9	—	—	—	—	2	—	—	—	—	—	158 of 239
ʊm → ən	22	3	2	1	4	—	—	—	—	—	1	—	—	—	—	33 of 95
a → ən	—	1	66	—	7	—	—	—	—	—	—	—	1	—	—	75 of 81
ʊs → ən	2	5	—	5	—	—	1	—	—	—	—	—	—	—	—	13 of 69
Umlaut + ər	7	2	—	34	—	8	—	—	—	—	—	—	—	—	—	51 of 54
ʊm → a	22	2	—	—	—	—	6	—	—	—	—	—	—	—	—	30 of 40
+ər	8	—	2	23	3	—	—	—	—	—	—	—	—	—	—	36 of 36
Umlaut	—	29	2	—	—	—	—	1	—	—	—	—	—	—	—	32 of 35
+iən	1	2	1	2	—	—	—	—	—	—	—	—	—	—	—	6 of 6
+tən	—	—	—	1	—	—	—	—	—	—	—	—	—	—	—	1 of 1
sum	358	486	504	538	92	41	19	13	4	5	2	5	3	1	0	2074 of 8598

[a] Each row has the errors of a particular plural class with the distribution of false predictions over the other plural classes. For example, of the falsely predicted items of the most frequent plural class (row 1), 107 were misclassified as identity plurals, 87 were misclassified as +n plurals, and so on. The columns thus tabulate overgeneralization errors for the column's class. As can be seen from column 5, the +s plural is overgeneralized 92 times despite its low type frequency.

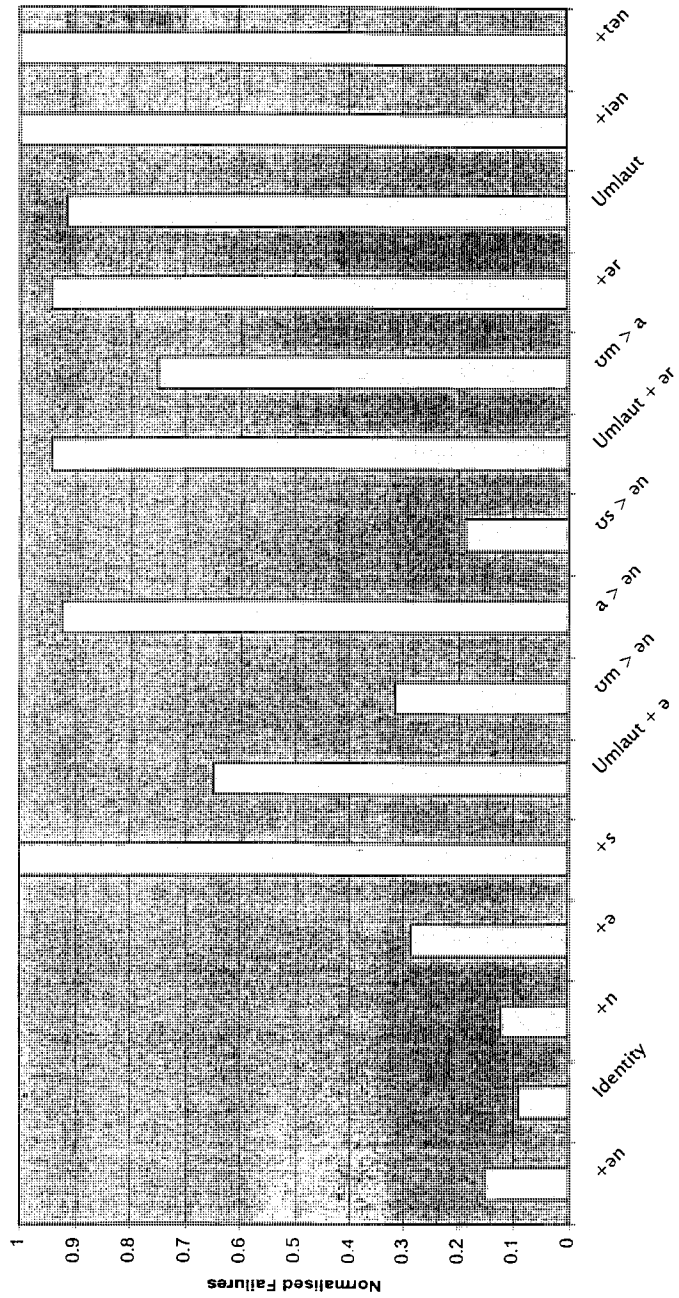

FIG. 6. A plot of the proportion of items incorrectly predicted by the dual-route GCM for each plural class. A maximum value of 1 indicates that all words in that plural class were misclassified. The value of 1 for the $+s$ plural reflects the fact that the best performing dual-route GCM never uses the default rule.

TABLE 3
The Dual-Route GCM's Errors Indicating the Way Words in Each Plural Class Were Misclassified[a]

	Falsely predicted values															
Class	+ən	Identity	+n	+ə	+s	Umlaut + ə	ʊm → ən	a → ən	ʊs → ən	Umlaut + ər	ʊm → a	+ər	Umlaut	+iən	+tən	Total for class
+ən	—	107	93	188	—	8	5	1	1	1	1	—	—	—	—	405 of 2646
Identity	61	—	75	36	—	3	3	1	1	1	—	—	1	—	—	182 of 1992
+n	20	157	—	2	—	—	2	6	—	—	—	1	1	—	—	189 of 1555
+ə	145	122	38	—	—	21	2	—	2	—	—	4	—	1	—	335 of 1178
+s	65	64	276	148	—	10	3	4	—	1	—	—	—	—	—	571 of 571
Umlaut + ə	14	4	4	131	—	—	—	—	—	2	—	—	—	—	—	155 of 239
ʊm → ən	22	3	2	1	—	—	—	—	—	—	2	—	—	—	—	30 of 95
a → ən	1	2	71	—	—	—	—	—	—	—	—	—	1	—	—	75 of 81
ʊs → ən	2	5	—	5	—	—	1	—	—	—	—	—	—	—	—	13 of 69
Umlaut + ər	7	2	—	34	—	8	—	—	—	—	—	—	—	—	—	51 of 54
ʊm → a	22	2	—	—	—	—	6	—	—	—	—	—	—	—	—	30 of 40
+ər	8	—	2	24	—	—	—	—	—	—	—	—	—	—	—	34 of 36
Umlaut	—	29	2	—	—	—	—	1	—	—	—	—	—	—	—	32 of 35
+iən	1	2	1	2	—	—	—	—	—	—	—	—	—	—	—	6 of 6
+tən	—	—	1	—	—	—	—	—	—	—	—	—	—	—	—	1 of 1
sum	368	499	565	571	—	50	22	13	4	5	3	5	3	1	—	2109 of 8598

[a] Each row has the errors of a particular plural class with the distribution of false predictions over the other plural classes. As can be seen from column 5, the +s plural is never overgeneralized because the best-performing dual-route model never uses the rule route.

TABLE 4
Summary of Comparative Single-Route and Dual-Route Performance

Pattern associator	Single-route	Dual-route
Nearest neighbor	70.8	70.1
Nosofsky GCM	74.3	73.8
Three-layer perceptron	82.7	81.2

similarity-based processes in morphology such models which incorporate no abstraction from their input have seemed obviously hopeless. Given that the GCM's performance is not far behind the network, further examination of this model seems profitable, not just from a linguistic perspective but from a psychological perspective as well. The lexically based tasks modeled here would seem to provide a particularly apt testing ground for this model which has been so successful in categorization research, due to both the fact that these tasks provide large quantities of ''real-world'' data which move beyond artificial laboratory stimuli and to the fact that one of the main criticisms of exemplar models in other areas, namely that extensive storage of exemplars is both costly and implausible, patently does not apply in the case of the lexicon.

Possibly even more surprisingly, supplementing each of these models with a symbolic rule route as posited by the dual-route account does not lead to an increase in predictive accuracy, despite the fact that the dual-route models possess an additional free parameter. However one views the level of accuracy achieved by the models, the results defeat the claim that the existence of low type-frequency defaults constitutes evidence which is damaging to single-route systems and favorable to dual-route systems: if one views single-route performance at about 80% as low, the crucial point is that dual-route performance is no better. If one finds, as we personally do, that predictive accuracy of 80% is remarkable, then single-route models do just as well as dual-route models.

Either way, the minority default argument is undermined. The dual-route account was meant to allow a dissociation between default and largest type, whereas the Pattern Associator Hypothesis was taken to predict that this is impossible. The critical assumptions were that pattern associators can generate default behavior only for a statistically predominant type, whereas the dual-route account suffers no such limitation. Our simulations reveal these assumptions to be oversimplified. The dual-route models' behavior is a product of the *interaction* between rule and associative component. While the rule route is entirely unaffected by frequency or similarity, this interaction means that a dual-route model *as a whole* cannot necessarily provide the right generalization behavior for both regulars *and* irregulars at the same time. Because the dual-route account subsumes a pattern-associator compo-

nent which through competition affects the use of the rule, it too—like single-route models—is *distribution dependent*.

Consequently, minority default inflections are not a priori compatible with the dual-route account. Whether a dual-route system can adequately generalize for a particular distribution is an empirical question just as much as it is for single-route systems. This has been obscured because the dual-route account was never actually implemented, and it necessarily means that the mere existence of minority default inflections says nothing about the psychological reality of a symbolic rule route.

Showing that the minority default argument is fallible in this particular way requires only proof of an instance where its assumptions break down and it is such proof that our simulations provide in the first instance. For this particular purpose, it is immaterial that these simulations, as any modeling effort, are limited in a number of ways. But can we make anything more of these results? Specifically, can they be interpreted as evidence not just against the minority default argument but as evidence which adjudicates between single- and dual-route models directly?

This depends on the representativeness of the test as a direct comparison: Are models and test fair? Do they genuinely capture the crucial aspects of either account? We turn to these question in the final section of the article, which presents direct quantitative comparisons. However, it also depends on the robustness of these results. The more likely this pattern of results is to reoccur given different implementational choices and different test sets, the less the specific choices in the present simulations matter and the more relevant they become as direct comparisons of the two accounts. Thus we next attempt to clarify the robustness of these results.

Robustness of Results

Different input representation. Because the nature of the input representation is a key determinant for the behavior of any computational process, we repeated our simulations with an alternative representation. This representation, a phonological feature set supplied by Brian MacWhinney, is more compact than Wurzel's (1981) scheme and employs only nine binary features to encode each phoneme. The results obtained show higher actual percentages correct, particularly for nearest neighbor and GCM, and correspond exactly in the overall picture of superior performance of single-route over dual-route models.

The nearest neighbor single-route performance was at 75.73% ($SE = 0.33$, $n = 10$), and the dual-route performance was at 74.57% ($SE = 0.29, n = 10$). The GCM too showed a marked benefit of the more compact representation scheme; single-route performance was at 78.59% ($SE = 0.29, n = 10$) while dual-route performance was at 77.01% ($SE = 0.31, n = 10$). For the network, finally, single- and dual-route results were 82.19% ($SE = 0.42, n = 10$) and 80.61% ($SE = 0.32, n = 10$) respectively.

These additional simulations bolster confidence in the generality of our results, but this input representation still excludes potentially relevant information such as phonological information about stress and syllabification, gender, token frequency, and semantics. How might their inclusion affect comparative model performance? Would we still see single-route performance equal or above dual-route performance?

There is obviously a huge space of possible models. The only way to address this question in general terms is to isolate the causes of this pattern of results. The pattern of slightly superior single-route performance indicates that both single- and dual-route models are distribution dependent, but what aspects of the distribution are crucial? We must ask for what distributions single- and dual-route model performance would be equal as well as where a dual-route model would be superior.

Where Defaults Would Help

The above error analysis indicates that all models' performance is determined not just by type frequencies but by the similarity structure in the language. Further confirmation of this comes from simulations which treated the *+en* plural (the most frequent type) as the default and which again produced the same general pattern of results. Hence the cause for the dual-route model's failure to exceed single-route performance does not lie in the small type frequency of the s-plural or in any other (directly) frequency-based consideration. It must be sought in the actual distribution of items in ''phonological space.''

Dual-route performance is never superior because even for the optimum value of *t* the rule produces *false positives*. Increasing performance on the regulars *decreases* the system's performance on the irregulars. This means the distances of the regular words to their irregular neighbors are not sufficiently different from the within-group distances of the irregulars. If they were, it would be possible to ''drive a wedge'' between regular and irregular distances, i.e., to select a value of *t* that correctly classifies regulars while leaving the irregulars untouched. In a language, in which interclass regular distances differed sufficiently from intraclass irregular distances, dual-route models would *match* single-route model performance.

These considerations further suggest that there should be distributions for which a default would help.

To demonstrate this, we generated two simple artificial languages (Nakisa & Hahn, 1996). Both languages consisted of five plural types distributed in a two-dimensional ''phonological'' space. Each noun class was generated around a centroid with a Gaussian distribution. For the first language, all five plural types had the same variance, whereas for the second, one group, the ''default,'' was exploded to occupy the entire space homogeneously without changing the type frequencies. Both distributions are depicted in Fig. 7.

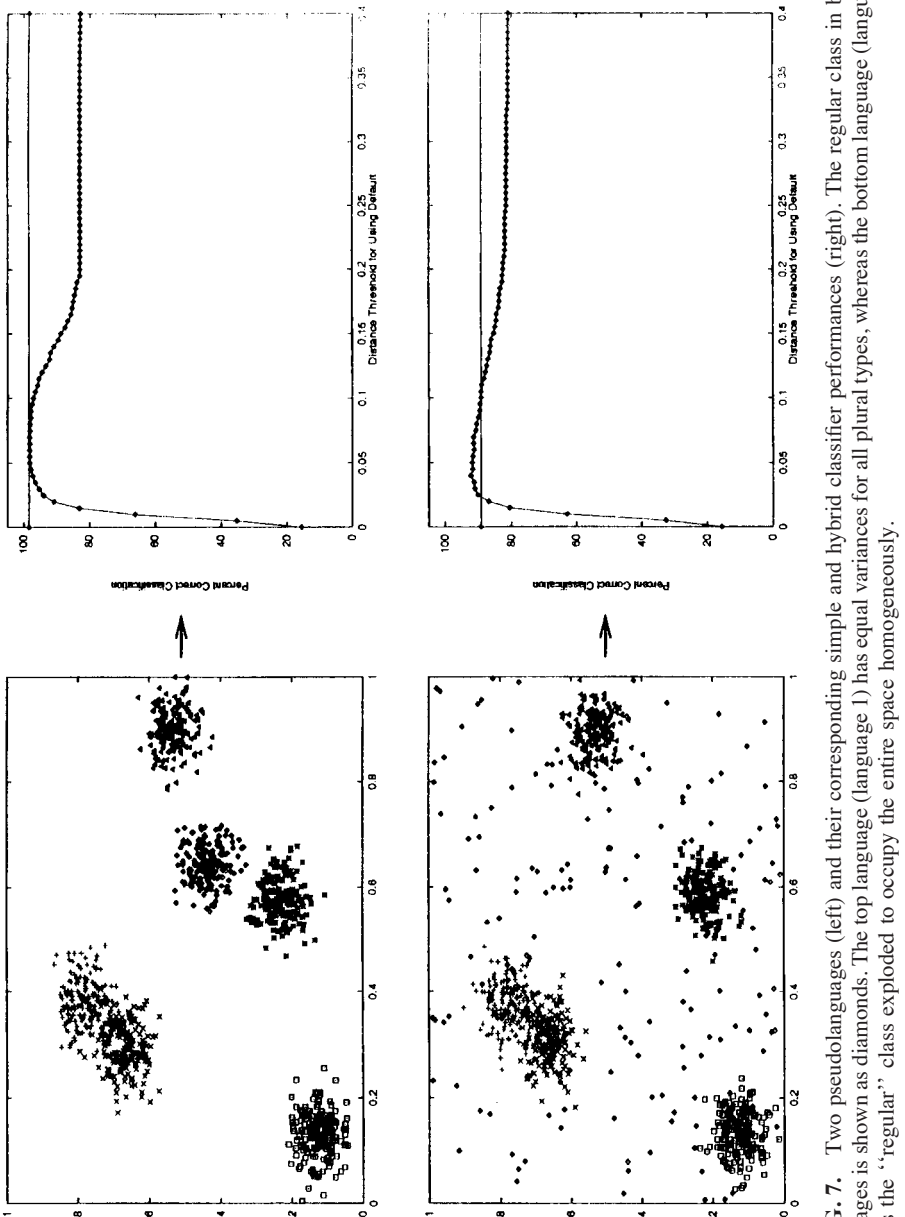

FIG. 7. Two pseudolanguages (left) and their corresponding simple and hybrid classifier performances (right). The regular class in both languages is shown as diamonds. The top language (language 1) has equal variances for all plural types, whereas the bottom language (language 2) has the ''regular'' class exploded to occupy the entire space homogeneously.

For the first language, where the ''default'' plural type had the same variance as the other types, the simple nearest neighbor classifier outperformed the hybrid classifier. On the second language, the dual-route nearest neighbor classifier is superior. For such a distribution where the irregulars are relatively compact and the regular is homogeneously distributed, adding a default *can* be beneficial for generalization.

Note that it is not the fact that regulars are homogeneously distributed per se that allows superior dual-route model performance. Regulars in isolated regions of the space, ''isolated regulars,'' are themselves equally well classified by dual-route and single-route models. As just outlined, differences in regular interclass and irregular intraclass distances are merely a precondition which allow the dual-route performance to equal single-route performance overall: isolated regulars alone allow a threshold t at which the dual-route's performance is not worse, but they do not enable it to do better.

Crucial for superior dual-route performance is the existence of a particular kind of regular item. It is the regulars forming a shell around each of the irregular clusters that are correctly classified by the hybrid model but not by the simple classifier, illustrated in Fig. 8. We call these regulars ''interfacial'' because they are distributed on the surface of the irregular clusters. Thus, increasing the ratio of ''interfacial'' to ''isolated'' regulars increases the benefit of the default. This can be achieved both by increasing the number of

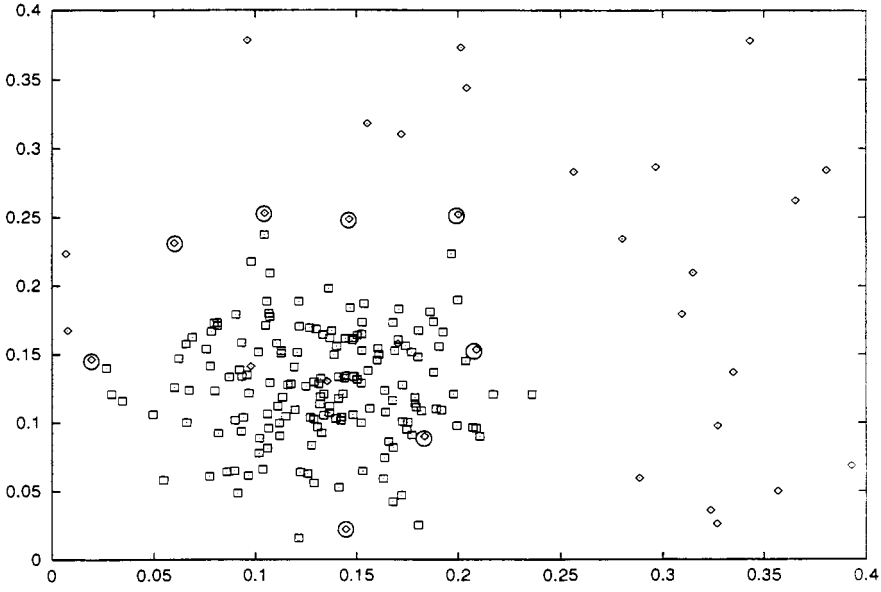

FIG. 8. The ''interfacial'' regulars for which adding a default can make a positive difference.

irregular plural types and (or) by increasing the surface area of irregular plural types.

The importance of interfacial regulars is foreshadowed with an idea expressed by Marcus et al. (1995). One of the 21 heterogeneous default circumstances is the use of the default for words with *competing similar entries in memory* (Marcus et al., 1995, p. 199):

> Because of the associative nature of memory, families of similar listed entries should pull a new entry toward its associated pattern. But the omnipotent regular past tense process can escape this attraction. . . . Indeed, in every irregular territory in phonological space, there are interloping regulars.

Our artificial language simulations refine this only slightly by making clear that it is not ''interloping'' items, i.e., items within an irregular cluster (which we will be misclassified by both single- and dual-route models), but items on the surface of irregular clusters which are crucial.

However, our real-language simulations also suggest that the German /s/ plural, at least with our current input representation, does not actually have this particular default characteristic: with enough interfacial regulars in the language, the dual-route model would have surpassed single-route performance.

Having identified the circumstances in which a default rule would help, we can return to the issue of how robust the pattern of comparative single-route/dual-route performance found in our real-language simulations might be. Would inclusion of semantics, token frequency, more complex phonological representations, and so on change the overall pattern of results? The artificial language simulations indicate that the single-route superiority would only be overturned if, as a consequence of changes to the representation scheme, the distribution of items in phonological space was altered in a very specific way: interclass distances for regulars would have to become sufficiently distinct from intraclass distances for irregulars so as to avoid dual-route performance decrements through false positives, and there would have to be interfacial regulars to make the dual-route not just equal, but superior.

It is possible that such an alteration might ensue from a richer input representation, but given the fact that there would seem to be many more possible distributions which do not meet these two constraints than ones that do, it seems rather unlikely. We would thus expect to find the overall pattern of results to be robust with respect to the inclusion of such factors in our models, though ultimately only actual implementation could tell.[11]

[11] Nakisa and Hahn's (1996) original work on comparing dual-route and single-route models in this way has subsequently been repeated for Arabic (Plunkett & Nakisa, 1997; Nakisa, Plunkett, & Hahn, 1998), where differences are more marked, and for English (Hahn, Nakisa, & Plunkett, 1997).

Summary

In summary, we have seen three very simple single-route models which achieve remarkably good performance given both the paucity of input and the fact that they were in no way custom-made for the task. Comparisons with matched dual-route models revealed that the minority default argument does not support the dual-route account because it cannot be assumed that an implemented dual-route model fares any better on the distributions in question. Crucially, making the dual-route account computationally explicit revealed that it too — like the models it is challenging — is *distribution dependent*.

The next section of the article further draws on these insights into the nature of a computationally explicit dual-route account to address the second, and key, evidence in favor of the dual-route account: its supposed parsimonious explanation of a variety of otherwise heterogeneous circumstances through the symbolic nature of the rule.

TWENTY-ONE DIFFERENT CIRCUMSTANCES — ONE LEVEL OF *t*

To recapitulate the ''unification and parsimony argument'': the symbolic default rule is employed wherever lexical memory fails and thus explains and unites the wide range of circumstances in which regular or default inflection is employed. In total, Marcus et al. (1995) list a set of 21 circumstances — categories — in which access to information in memory fails and the default is used. These heterogeneous circumstances, which are united only by the fact that they can bear the common symbol ''Verb'' (in the case of verb inflection) or ''Noun'' are summarized in Table 5, which we have adapted from Marcus et al. (1995).

Crucially, these encompass items which are not presumed to have ''normal'' lexical entries, i.e., items which are not so-called *canonical roots,* such as proper names, quotations, acronyms, or truncations. These are phenomena excluded by connectionist modeling to date as well as by our own models above. A symbolic rule seems to be a straightforward way of generating the desired behavior in a model and thus, at least in terms of scope, these phenomena seem to provide an important argument for the dual-route account.

We will show, however, that a closer look at the data and a more careful consideration of computational issues reveal that the dual-route account's single symbolic rule *cannot,* in fact, unite these circumstances.

To develop this argument we start with the example of ''Mickey Mouses'' and the quotation ''there are three 'man's in the first paragraph.'' The key aspect of these examples is that the same phonological form — man or mouse — can receive two different plurals as a function of specific grammatical circumstance. According to the dual-route account, one of these, the irregular ending (men or mice) is allocated through lexical memory and the

TABLE 5
The 21 Phenomena for Which the Default Rule Is Meant to Provide a Parsimonious Account[a]

Circumstance	Kind of word	Example
1. No root entry	Novel words	*wug*
2. Weak entry	Low-frequency words	*stinted*
3. No similar entries	Unusual-sounding words	*ploamph*
4. Competing root entry	Homophones	*lied, lay*
5. Competing similar root entry	Rhymes	*blinked*
6. Rendering of sound	Onomatopoeia	*peeped*
7. Mention vs use	Quotation	*''man''s*
8. Opaque name	Surname	*the Childs*
9. Foreign language	Unassimilated borrowings	*latkes, cappucinos*
10. Distortion of root	Truncations	*synched*
11. Artificial	Acronyms	*PACs, OXes*
12. Derivation from different category	Denominal Verbs	*high-sticked, spitted*
13. Derivation via different category	Denominal nominalized verbs	*flied out, costed*
14. Derivation via name	Eponyms	*Mickey Mouses*
	Products	*Renault Elfs, Top Shelfs*
	Teams	*Toronto Maple Leafs*
15. Referent different from root	Bahuvrihi Compounds	*sabre-tooths, low lifes*
	Pseudo-English	*walkmans*
16. Lexicalization of phrase	Nominalized VPs	*bag-a-leafs, shear-a-sheeps*
17. Children	Overregularization	*comed, breaked*
18. Normal speech errors	Overregularization	*comed, breaked*
19. Alzheimer's	Overregularization	*comed, breaked*
20. Williams Syndrome	Overregularization	*comed, breaked*
21. Anomic Aphasia	Overregularization	*comed, breaked*

[a] After Marcus et al. (1995). These are claimed to apply equally to German participles and plurals. For ease of understanding we show English only. The reader is referred to Marcus et al. (1995) for further details.

other arises from the rule route. The rule route can apply even though memory already contains an irregular entry because ''the grammatical mechanisms that allow information in memory entries to be passed to the word are systematically disabled'' (Marcus et al., 1995, p. 196).

For the account to be both explanatory and predictive necessitates an explanation of *why* memory is disabled; furthermore it is necessary for these circumstances to have *general descriptions* if the explanation is to avoid being ad hoc. Marcus et al. (1995) attempt to meet these constraints with the broad categories of ''memory failure'' of Table 5.

''Disablement of memory'' is necessarily all or none: an item can either access lexical memory or not—it cannot access memory just a little. Coupled with the need for category-level explanations (i.e., general descriptions of

circumstances of memory failure) this gives the dual-route account a very binary, all-or-none flavor: *all items of a given category of ''memory failure'' should behave the same way;* namely they should take the default inflection. Furthermore, there should be *no difference between different categories of memory failure*. In all cases, memory access is impossible or suppressed, yielding uniform production of regular forms as an outcome.

Of course, there can be some ''noise'' without threatening the account, but the accounts' predictive power at the level of individual words is directly linked to the level of exceptions for a given category. Furthermore, the account is challenged by any deviations from its basic all-or-none predictions which are *systematic* in a way that is incompatible with the nature of the account. We argue that behavioral data contain levels and kinds of systematic variability which are incompatible with the dual-route accounts' binary nature. Not only do the different categories of Marcus et al. (1995) allow exceptions, but there exist systematic, statistically significant differences *between categories* in the levels of regular inflection produced.

Human production data which indicates systematic differences between categories of memory failure is, in fact, provided by Marcus et al.'s own experiments (Marcus et al., 1995). They report an experimental study of German plurals to support the dual-route argument, in particular to demonstrate the use of the default +/s/ in circumstances in which access to memory is systematically disabled because the item in question is not a lexical root. Participants were presented with 24 German nonce words embedded in a sentential context. For all 24 items, subjects were asked to rate the ''naturalness'' of different possible pluralizations. Items were presented in three different between-subjects conditions: a sentential context that presented them as a lexical root (i.e., an ordinary noun), a sentential context that presented them as a foreign word, and a sentential context that presented them as a name.

The dual-route account's predictions are:

1. In the lexical root condition items should access lexical memory, and only if memory fails, because there is no sufficiently similar stored item should the rule route be used; thus, a mixture of irregular and regular inflection, depending on the phonology of the item, is expected.

2. Foreign words (''unassimilated borrowings''): according to Marcus et al. (1995) these, like onomatopoeia (''oink'') or quotations, are treated as ''stretches of sound,'' not as lexical roots (p. 200); memory access is disabled and uniformly regular inflection is predicted.

3. Names, too, are ''opaque stretches of sound'' (p. 200), not roots, and thus cannot evoke similar roots; i.e., memory access is systematically disabled with uniform regular inflection as a consequence (i.e., ''Mickey Mouses'' not ''Mickey Mice'').

These predictions are only partially born out by Marcus et al.'s data. As predicted, the lexical root condition yields a mix of irregular and regular

preferences. The ''borrowing'' and ''name'' condition also yield significantly higher ratings for the default +/s/, but the data do not display the predicted uniformity in these conditions. For 10 of the 24 items in the root condition, the rating for the regular +/s/ exceeded that of the highest irregular. In the foreign (or borrowing) condition, +/s/ received the highest rating 20 of 24 times and for the names this went up to 22 of 24. Thus while the naturalness of +/s/ increases as predicted, it does not do so as uniformly as predicted. This is even more apparent when the mean ratings of regular inflection and highest irregular inflection are contrasted. On a scale of 1–5, with higher numbers indicating greater ''naturalness,'' the ratings for irregulars and regulars respectively were 4.3 vs 3.6 for roots, 3.8 and 3.8 for borrowings, and 2.9 vs 4.2 for names. The predicted means, if memory access is systematically disabled, however, should be (or at least approach) 1 vs 5 for both borrowings and names.

This discrepancy is indicative in and of itself. Crucial, however, are the *between-category differences* in the data. The first of these is the difference between the borrowing and the name conditions. This difference is baffling if memory access is disabled in both cases. Marcus et al.'s own explanation of this difference is that it ''may be due to subjects' ability to treat some of the borrowings as fitting the native sound pattern and hence to rate them as being like roots'' (p. 238). This explanation seems unsatisfactory for two reasons.

First, all nonce words were either rhymes of extant lexical items or created using a list of permissible and nonpermissible onsets and codas from which only nonexisting but possible combinations were selected. Consequently all items arguably fit ''the native sound pattern'' by design. Second, if borrowings *were* treated like roots whenever they are perceived to fit the native sound pattern, then the ''foreign words'' category would seem to be entirely redundant. The appropriate predictions for foreign words should then simply match those of the lexical root condition, where some words are preferentially irregularized and some preferentially regularized, depending on their similarity to known items. In other words, foreign words are either treated as opaque stretches of sound, in which case uniform regularization is predicted, or their phonological form is uniformly taken into account, in which case we would expect a replication of the ratings observed in the lexical root condition; the actually observed pattern of results, however, falls between these two.

Even more striking is a further between-category difference in this data set. If there is anything like a strict rule in German inflectional morphology, it is that surnames are pluralized with +/s/. If Thomas Mann's family were coming over to tea, this would only ever be expressed as ''die Manns kommen zum Tee'' even though ''Mann'' is a noun with irregular plural (analogous to ''the Childs are coming for tea'' vs ''the Children are coming over for tea''). In this one case, the application of the default really is strict and

native-speaker intuitions concur exactly with the binary predictions of the dual-route account. However, this is already *not* the case where Christian names are concerned; female firstnames such as Ulrike or Beate could conceivably be inflected as ''Ulriken'' or ''Beaten'' (presumably in analogy to a strong pattern for feminina ending in schwa). Native-speaker intuitions thus suggest that the production of +/s/ is universal for surnames, but only prevalent, not universal, for other names.

Marcus et al.'s (1995) data set provides an ideal test set for this claim because only half the items in the name condition were presented as a surname, whereas the other half were presented as first names or as names of films, books, or rivers. In fact, an analysis of the items in the name condition (Marcus et al., 1995, Appendix 3) reveals that ratings for +/s/ are significantly higher among the 12 items presented as surnames (family names) than among the 12 items presented as other names (first names, rivers, or books). The surname mean rating is 4.65 (where 5 corresponds to a judgement of the form as ''perfectly natural'') with a standard deviation of .64; but the mean rating for other names was only 3.8 with a standard deviation of .27. This difference is statistically significant, $t(22) = -4.27$, $p < .01$.

This difference is not attributable to the items' phonology because there is no significant difference in the means of both groups when they are presented in the sentential context treating them as lexical roots rather than names: mean = 3.64, $SD = .49$ vs mean = 3.67, $SD = .52$; $t(22) = .12$.

Analogously, the *irregular* ratings were significantly lower for the items presented as surnames than for those presented as other names: mean = 1.99, $SD = .44$ (surnames) and mean = 2.97, $SD = .47$ (other names); $t(22) = 5.29$, $p < .01$.

Thus, the statistically significant differences between *types* of names provide another seemingly systematic between-category difference in irregular productivity. It is worth emphasizing again that the Marcus et al. experiments provide the ideal test data for this point because they demonstrate these between-category differences in maximally controlled conditions. *All* words are new, thus ruling out strange, unevenly distributed quirks in the language (stemming from diachronic factors, historical accidents, influences of particular dialects, and so on) as the source of these between-category differences, and due to the way that stimuli were designed it can be ruled out that the effects are based on exceptional items of unusual phonological form.

Because the point is so important, however, we sought to complement these data with our own observations to cover an even wider range of the 21 circumstances. We present these next.

Materials and Methods

The stimulus items fall into several different categories. We included surnames, Christian names, truncations, acronyms, product names, low-frequency real words, and nonwords. Among the nonwords is the Marcus et al. (1995) set of 24 monosyllabic nonce words, presented

as lexical roots. To these, we added a number of polysyllabic nonwords. Truncations, acronyms, and product names were gathered by looking through two daily papers (*Süddeutsche Zeitung* and *FAZ*) and two weekly journals (*Stern* and *Spiegel*). The items were presented in a written questionnaire, which presented the word's singular in a sentential context, followed by a sentence requiring the plural form with a blank in the appropriate position. The entire set of words is presented in Appendix A.

Participants. Participants were 28 members, staff and students, of the Institut für Molekulargenetik (Molecular Genetics Institute) of the University of Heidelberg, Germany.

Procedure. Questionnaires were handed out in the institute and collected in the course of the day. Participants were instructed not to deliberate over their answers, as it was their gut feeling that mattered.

Results

We present our results only at a very general level, sufficiently detailed only to underscore the general point about graded levels of irregular productivity; for a more in-depth linguistic analysis the reader is referred to Hahn and Nakisa (in preparation). To this end, Fig. 9 shows the results in the following way. Separated by thick black lines are various categories of words; these are based on Marcus et al. (1995) (see Table 5 above), though some have been subdivided further; for instance, names have been split into first names and surnames in line with what we said above. Each row of boxes represents the 15 possible plurals for a particular word. The colored squares indicate which plurals are actually attested. The darker the square, the more participants produced this particular plural as a response. The category ''Lexical Pseudowords'' contains the 24 monosyllabic nonce words used by Marcus et al. (1995). These were supplemented with several polysyllabic nonce words as well as with low-frequency real words and a category ''Lexical Conflict'' comprised of real words with competing plurals. For all but these last four categories memory access should be systematically disabled according to the dual-route account (Marcus et al., 1995), with the consequence that the only attested forms — barring experimental noise — should be +/s/, visible as a single black column. This, however, is clearly not the case. Notice also the difference between surnames and first names, confirming further the differing levels of productivity.

Consequences of differing levels of irregular productivity. In summary, we have seen three sources of evidence for the claim that irregular productivity does not suddenly cease wherever the dual-route account predicts that memory access is disabled and for the claim that levels of irregular productivity vary systematically between different categories of memory-disabling circumstance.

These deviations from the predictions of the dual-route account have two consequences. First, the exceptions *within* a category reduce the accounts' ability to make predictions about the fate of individual words. Second, that there are systematic, statistically significant, variations in the degree of irregular productivity observed for different memory-disabling circumstances constitutes an aspect of inflection which needs accounting for. Within the

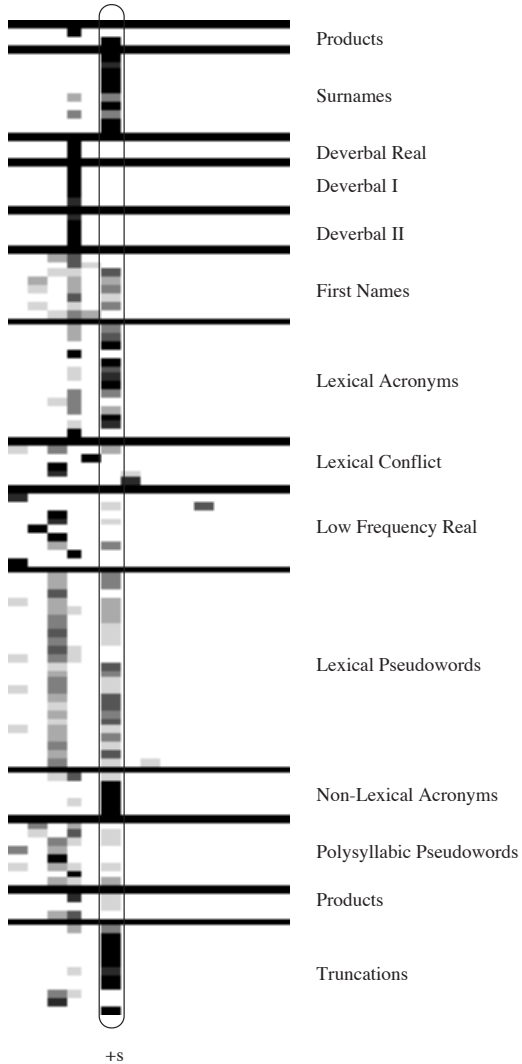

FIG. 9. The data presented by category. Each row consists of 15 squares, one for each plural type. The darker a square the more responses of this type were made. The +/s/ is highlighted. For all categories except the 24 ''Marcus Words'' and the low-frequency regulars the dual-route account predicts +/s/ exclusively. Note also (''Lexical Pseudowords'') how this prediction is fulfilled by the surnames, but not for other categories, indicating graded irregular productivity.

dual-route framework, however, the only way to achieve differential levels of irregular productivity (as opposed to a clean-cut binary transition from productivity to no productivity) is through changing the value of the threshold *t,* the parameter governing route interactions. As we saw in the simulations above, changing the value of *t* changes the efficacy of the rule route, thus determining the relative productivity of irregulars vs regulars. In other words, different levels of irregular productivity require *different levels of t*. They are logically incompatible with a single threshold *t*. If productivity varies as a function of category (surnames vs borrowings vs first names, etc.) then each of these categories requires its own level of *t*. This, however, means having a *different* pattern associator plus rule ensemble for *every* category. There is no longer a single symbolic default unifying all 21 categories, but rather 21 different, category-specific rules. As currently specified by Marcus et al. (1995), the dual-route account is incapable of producing graded levels of productivity with the consequence that the argument based on parsimonious uniting of 21 otherwise heterogeneous circumstances fails. Empirically, they *are* heterogeneous in an important way.

It should also be noted that these consequences do not depend on the particular way in way in which we have defined the threshold *t;* rather, the argument holds as long as route interaction is governed by a single parameter, however that parameter is elucidated. That it should be a single parameter, however, stems from the heart of the account whose core is the single, unifying symbolic rule.

Thus the second body of evidence in favor of the dual-route account, in fact, not only fails to support it, but on closer inspection seems to provide evidence *against* the account in its current form.

FITTING HUMAN BEHAVIORAL DATA: DIRECT QUANTITATIVE COMPARISONS BETWEEN SINGLE- AND DUAL-ROUTE MODELS

So far, we have provided evidence against two central arguments made to support the dual-route account, the minority default argument and the argument from the symbolic rules unifying explanatory power. The refutation of this supporting evidence does not directly bolster the case of single-route models given that both accounts by no means exhaust the set of possible models of inflection. Consequently, the remainder of the article addresses the *direct comparison* of single- and dual-route models.

We earlier raised the question of whether the simulations which were addressed at the minority default argument might also be considered as direct quantitative comparisons between single- and dual-route accounts. If yes, the superior performance of the single-route models would provide direct evidence in their favor; but is this additional interpretation sustainable?

We argued above that generalization is the most suitable way to contrast the two accounts; we also addressed limitations of the input representation

through our examination of the robustness of these results. This leaves the question of whether the data set is sufficiently representative when viewed in this context.

The test data is the entire German lexicon as stored in CELEX, but this excludes nonlexical items such as acronyms, quotations, truncations, or names—data which might be thought to benefit the dual-route account. As the previous section has shown, however, the dual-route account cannot actually accurately capture these default circumstances. Furthermore, there is no in principle reason why a single-route model might not be able produce default inflection in these circumstances. If, for example, proper names exclusively take /+s/, then whether an item is a proper name must be incorporated into the input representation. A network can extract this as a critical feature; for the similarity-based nearest neighbor and the GCM, the ''proper name feature'' would have to have sufficient weight in the input representation to ensure that proper names formed their own subcluster in phonological space. A different matter, of course, is whether these models could accurately reproduce the patterns described in the previous section; crucially, however, the restriction to lexical roots cannot be seen to bias the comparison in favor of single-route accounts a priori.

A second possible concern is that generalization on the extant lexicon links only indirectly to actual human performance. It might be seen as an essentially normative task, i.e., capturing what people *ought* to do rather than what people would *actually* do when faced with lexical items whose inflection they do not know. One might counter that the normative and descriptive can diverge only so far without undermining the stability of a system: if people's intuitions regularly suggested an inflectional form contrary to the actual form one would expect the latter to erode over time.

Ultimately, the only way to entirely allay this concern is to use actual behavioral data; that is, experimentally elicited nonword generalization. This comes with a complementary set of problems. The greatest problem is probably that of selecting a fair test set, as our artificial language simulations underscored. Our test of generalization on the extant lexicon was not plagued by any sampling problem because we simply used the entire population, but any human nonword performance data will necessarily be a comparatively modest sample. The 24-item test set devised by Marcus et al. (1995), which we described in the previous section, samples only a limited area of phonological space, but, given that it was designed with the dual-route account in mind, its use seems fair at least to dual-route models.

These 24 items, presented as lexical roots, were included in the production data we collected (described above). We thus extracted them for use in a further model comparison.

The use of experimental data also has consequences for the evaluation of model performance. Our simulations above scored model responses as either correct or incorrect. However, there seems to be no real way in which there

is a ''correct'' response to these 24 nonwords: as can be seen from Fig. 9 above, each stimulus item elicits a range of responses. On the full data set the average agreement between two raters lies at 49%, i.e., two ''typical raters'' agree on 49% of the 112 items (computed as the average between all 351 possible pairs of the 27 raters). This level of agreement is above chance agreement, as predicted by a base-rate model (computed by the percentages in the noncompound data set as set out in Table 1) which lies at 29% ($\kappa = .28$); but raters seem to disagree as much as they agree.

As can also be seen from Fig. 9, the overall distribution of responses clearly varies *between items;* this suggests fitting the model directly to the overall distribution for each item. With these general consideration in hand, we turn to the description of the modeling itself.

Models

We chose the GCM for these simulations. The GCM has only one free parameter (two in the dual-route version), in contrast to the numerous free parameters involved in network simulation. This makes the GCM far easier to use for detailed predictions about individual words. Given a network of particular hidden unit size and predetermined levels of training, one would still have to average results across many different initial random seeds in order to obtain a robust set of model predictions, particularly with respect to the low-frequency types (see also Error Analysis above). The single-route GCM, by contrast, makes a unique set of predictions for each value of the sensitivity parameter *s*. Furthermore, initial tests had shown the model to perform quite well on predictions of human performance.

Methods

The single-route GCM was given the entire vocabulary of 8598 nouns as its lexical memory; the dual-route GCM was given the same set minus the 571 /+s/ plurals to ensure that regular productivity was independent of the lexicon as stipulated by Marcus et al. (1995) (see Dual-Route Models above). Both models were tested on the unseen 24 nonwords. Performance was measured in terms of the model's ability to predict each item's *overall distribution of responses*. The GCM's output is a probability for each plural class, so this simply involved correlating the item distribution with the models' set of class probabilities, for each item in the test set.

We exhaustively sampled for all the best settings of the sensitivity parameter *s* (see above) and then—for the dual-route GCM—of the threshold *t*, choosing the values which gave optimal performance (i.e., the minimum RMS error) on the test set.

Results

This yields a value of $r^2 = .57$ for the single-route GCM and a value of $r^2 = .40$ for the dual-route GCM. Thus, the single-route model again outperforms the dual-route version, accounting for an additional 17% of the item variability. The comparative performance of both models is shown in

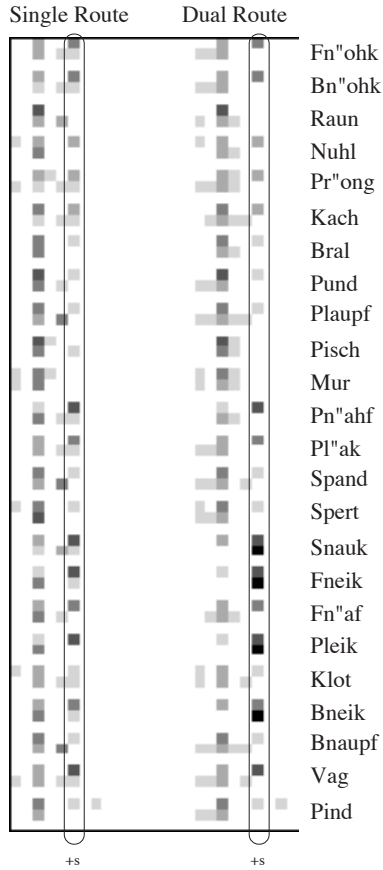

FIG. 10. Comparative model-fits for 24 ''Marcus Words''; from left to right are single-route GCM and dual-route GCM. The upper line of squares in every pair represents the distribution of responses over the 15 plural types in the human data; the line immediately below represents model output.

Fig. 10, which plots the behavioral data against model performance. This also allows examination of the nature of model errors.

Discussion

The single-route model performs better than the dual-route model, despite the fact that the latter has an extra free parameter.[12] This result is not unex-

[12] One reviewer inquired whether the differences between models in our simulations were statistically significant. On statistical grounds, the model with fewer free parameters is preferable given equal performance. Hence tests for model comparisons such as Akaike's Information Criterion include a penalty term for model complexity. Given that the dual-route model actually performs *worse,* there is no point in applying this test.

pected, given the very specific distribution necessary for superior dual-route performance which we gained from the artificial language simulations. But because the models are being compared in their ability to fit human behavioral data, these comparisons provide support for a single-route account over the dual-route account. They complement the simulations testing generalization on the extant lexicon viewed as direct comparisons.

These results also confirm that the considerable variability in the data is not just noise, but rather is largely predictable on the basis of the structure of the language alone: the response distribution for each item matches well the respective class probabilities computed by the model. The variability in the data (as shown in Figs. 9 and 10), and the degree to which this is captured by the models, more generally makes a case for probabilistic accounts of inflection.

This would be a reorientation at least from the point of view of the dual-route account as previously characterized (Pinker & Prince, 1988; Pinker, 1991; Prasada & Pinker, 1993; Marcus et al., 1995). The dual-route account's pattern associator component can readily be interpreted probabilistically, as in our dual-route GCM; but the rule route excludes all other responses by definition. This makes the dual-route predictions entirely uniform wherever the rule is used, as can be seen in Fig. 10 on the items *Snauk, Fneik, Pleik, Bneik*. From the perspective of the dual-route account the variability in these cases is, in first instance, experimental noise. Possible paths along which the dual-route account might seek its own explanation are some kind of ''noisy'' threshold or individual subject variation. By contrast, the single-route GCM suggest that this variability is already largely explained by the items' position in phonological space. This additional challenge to the dual-route account can be met only by an explicit dual-route model with superior data fits.

CONCLUSIONS

This article countered two arguments put forth in support of the dual-route account of inflectional morphology. The minority default argument claimed that the mere existence of languages with low type frequency defaults indicated that inflection cannot be based on a single-route associative mechanism. Large-scale simulations with matched single- and dual-route models demonstrated that the structure of a given minority default system need not be any more compatible with a dual-route mechanism, thus undermining any support the argument might lend the dual-route account. The minority default argument flagged distribution dependence as a weakness of single-route models. Our simulations show that this ''weakness'' is shared by dual-route models and that dual-route models require very specific distributions, which we identified by means of artificial language simulations, for their performance to exceed that of otherwise matched single-route models.

The second argument examined was the symbolic rule's unifying and par-

simonious explanation of otherwise heterogeneous circumstances. Closer investigation of experimental data by Marcus et al. (1995), as well as data of our own, illustrate that these categories are, in fact, heterogeneous in one crucial way: namely with respect to the amount of irregular productivity they seem to allow.

For many of these categories—all those which do not comprise lexical roots—the fact that there is *any* irregular production is in conflict with the predictions of the dual-route account. Most importantly, however, the statistically significant difference in *degree* of irregular productivity *between categories* seems logically incompatible with the dual-route model in its current form. Differing amounts of irregular production for a given word can be achieved only by varying the parameter governing route interaction. But different settings for different circumstances defeat the unifying explanation through a single rule.

Finally, we compared the single- and dual-route GCM's ability to model human nonword generalization. On a set of 24 monosyllabic nonwords, designed by Marcus et al. (1995), the single-route version considerably outperformed the dual-route version. This concurs with the results of the model comparisons on previously unseen real words, except that there the discrepancy between the models had been less marked. Read together with the insights gained from the artificial language simulations, the agreement between these two complementary tasks strongly suggests that it is not merely an unlucky choice of data set that is to blame for the dual-route model's lesser performance.

The use of multiple models shed light on the relative importance of phonological (surface) similarity, type frequency, and higher order regularities. The nearest neighbor models, whose performance depends exclusively on phonological similarity, provided a baseline to which the relative contributions of type frequency—in the GCM—along with the ability to rerepresent the input as desired—in the network—could be compared. That the GCM, a model from the categorization literature, fared so well should be of interest beyond psycholinguistics.

The key difference between this research and previous work has been its focus on predictions at the individual word level as opposed to qualitative, global phenomena (U-shaped learning, minority defaults, etc.) or properties of classes of words (''low-similarity words,'' ''high-similarity words,'' ''high frequency types,'' etc.). This shift in focus revealed novel facts about German plural morphology: the fact that a system viewed as predominantly arbitrary is, in fact, very predictable; that irregular productivity varies systematically; and that many forms are productive, giving rise to considerable variability in human generalization, which itself can be fairly well explained in terms of items' neighborhood characteristics. The focus on predictions for individual words was also central to the reevaluation of the minority default argument and the unification argument because it drew attention to

implementational considerations and their consequences. Finally, it enabled the first direct quantitative comparisons of competing models of inflection. That a phenomenon as complex as language can be tested at such a detailed level suggest exciting possibilities for the future.

APPENDIX A

The following words and nonce words were presented to participants. A question mark indicates the lack of a defined gender. In these cases, one of the three German genders was presented, with the particular choice randomized and counterbalanced across subjects.

? AP
der DAX
das Kfz
die SZ
? AOL
? HRK
das BAFöG
der PC
die EDV
? OPM
? OAU
die GmbH
das LKA
die PS
der IWF
der Pkw
die AG
die ZVS
? ZHS
der BRDler
Familie Schulz
Familie Weiss
Familie Hahn
Familie Meier
Familie Hansen
Familie Wolf
Familie Kahr
Familie Dachs
Familie Schmidt
Familie Müller
Klaus
Ulrike

Thomas
Hans
Michael
Beate
Christine
Peter
? Hacker
? Benser
? Muffer
? Paller
? Schnuckler
? Schlunzen
das Maxen
? Relpen
das Wulsen
das Rennen
das Wettsingen
der (VW) Polo
der Mercedes
der Opel
der Walkman
der (VW) Käfer
die Disco
das Fax
der Profi
der Prof
der Sozi
der Schieri
das Tele
der Direx
der Juso
die MuGe
der Aku
der Abbrand
das Bistum
das Mahnmal
das Schicksal
der Abseit
? Bral
? Kach
? Klot
? Mur
? Nuhl
? Pind

? Pisch
? Pund
? Raun
? Spand
? Vag
? Spert
? Bnaupf
? Bneik
? Bnöhk
? Fnähf
? Fneik
? Fnöhk
? Pläk
? Plaupf
? Pleik
? Pnähf
? Pröng
? Snauk
der Lift
die Alm
der Karfunkel
der Fakir
das Ekzem
der Pier
die Replik
die Viper
die Pizza
? Ferrsuhr
? Ringboss
? Laupnahl
? Ravocht
? Blaumde
? Haunter
? Zumsel
? Fagelzeng

REFERENCE

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B. Petrov & F. Csáki (Eds.), *2nd International Symposium on Information Theory*. Akadémiai Kiadó, Budapest.

Alegre, M., & Gordon, P. (1999). Frequency effects and the representational status of regular inflections. *Journal of Memory and Language,* **40,** 41–61.

Augst, G. (1979). Neuere Forschung zur Substantivflexion. *Zeitscrift fuer germanistische Linguistik,* **7,** 220–232.

Breiman, L., Friedman, J., Olshen, R., & Stone, C. (1984). *Classification and regression trees*. Belmont, CA: Wadsworth.

Bullinaria, J. (1997). Modelling, reading, spelling and past tense learning with artificial neural networks. *Brain and Language,* **59,** 236–266.

Bybee, J. (1995). Regular morphology and the lexicon. *Language and Cognitive Processes,* **10,** 425–455.

Bybee, J., & Moder, C. (1983). Morphological classes as natural categories. *Language,* **59,** 251–270.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Clahsen, H. (1999). Lexical entries and rules of language: A multidisciplinary study of german inflection. *Behavioral and Brain Sciences,* **22,** 991–1010.

Fodor, J., & Pylyshyn, Z. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition,* **28,** 3–71.

Hahn, U., & Chater, N. (1998). Rules and similarity: Distinct? Exhaustive? Empirically distinguishable? *Cognition,* **65,** 197–230.

Hahn, U., & Nakisa, R. (in preparation). On the notion of ''regular'' or ''default.''

Hahn, U., Nakisa, R., Bailey, T., Holmes, M., Kemp, D., & Palmer, L. (1998). Experimental evidence against the dual-route account of inflectional morphology: Frequency and similarity. In *Proceedings of the 20th Annual Meeting of the Cognitive Science Society* (pp. 472–477). Mahwah, NJ: Erlbaum.

Hahn, U., Nakisa, R., & Plunkett, K. (1997). The dual-route model of the english past-tense: Another case where defaults don't help. In *Proceedings of the GALA '97 Conference on Language Acquisition*. Human Communications Research Centre: The University of Edinburgh.

Kim, J., Pinker, S., Prince, A., & Prasada, S. (1991). Why no mere mortal has ever flown out to center field. *Cognitive Science,* **15,** 173–218.

Köpcke, K. (1988). Schemas in german plural formation. *Lingua,* **74,** 303–335.

Köpcke, K. (1993). *Schemata bei der Pluralbildung im Deutschen*. Tübingen: Gunter Narr.

Lee, B. (1996). On the processing of regular and irregular inflections: The symbolist-connectionist debate revisited. In C. Koster & F. Wijnen (Eds.), *Proceedings of the Groningen Assembly on Language Acquisition*.

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

Marcus, G. (1998). Can connectionism save constructivism? *Cognition,* **66,** 153–182.

Marcus, G., Brinkmann, U., Clahsen, H., Wiese, R., Woest, A., & Pinker, S. (1995). German inflection: The exception that proves the rule. *Cognitive Psychology,* **29,** 189–256.

Mugdan, J. (1977). *Flexionsmorphologie und Psycholinguistik*. Tübingen: Gunter Narr.

Nakisa, R. C., Plunkett, K., & Hahn, U. (2000). A cross-linguistic comparison of single and dual-route models of inflectional morphology. In P. Broeder & J. Murre, (Eds.), *Cognitive models of language acquisition*. Cambridge, MA: MIT Press.

Nakisa, R., & Hahn, U. (1996). Where defaults don't help: The case of the german plural system. In *Proceedings of the 18th Annual Meeting of the Cognitive Science Society*. Mahwah, NJ: Erlbaum.

Nosofsky, R. (1986). Attention, similarity and the identification-categorization relationship. *Journal of Experimental Psychology: General,* **115,** 39–57.

Nosofsky, R. (1988a). Exemplar-based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **14,** 700–708.

Nosofsky, R. (1988b). Similarity, frequency and category representations. *Journal of Experimental Psychology: Learning, Memory and Cognition,* **14,** 54–65.

Nosofsky, R., Clark, S., & Shin, H. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition,* **15,** 282–304.

Pinker, S. (1991). Rules of language. *Science,* **253,** 530–535.

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

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

Plunkett, K., & Marchman, V. (1993). From rote learning to system building. *Cognition,* **48,** 21–69.

Plunkett, K., & Nakisa, R. (1997). A connectionist model of the arabic plural system. *Language and Cognitive Processes,* **12,** 807–836.

Prasada, S., & Pinker, S. (1993). Generalization of regular and irregular morphological patterns. *Language and Cognitive Processes,* **8,** 1–56.

Quinlan, R. (1992). *C4.5 software: Programs for machine learning*. San Mateo, CA: Morgan Kaufmann.

Rueckl, J., Mikolinski, M., Raveh, M., Miner, C., & Mars, F. (1997). Morphological priming, fragment completion, and connectionist networks. *Journal of Memory and Language,* **36,** 382–405.

Rumelhart, D., & McClelland, J. (1986). On learning past tenses of english verbs. In J. McClelland & D. Rumelhart (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition. Vol 2: Psychological and biological models*. Cambridge, MA: MIT Press.

Russ, C. (1989). Die pluralbildung im deutschen. *Zeitschrift für germanische Linguistik,* **17,** 58–67.

Seidenberg, M., & Bruck, M. (1990). *Consistency effects in the generation of past tense morphology*. Paper presented at the 31st Meeting of the Psychonomic Society, New Orleans.

Sereno, J., & Jongman, A. (1997). Processing of english inflectional morphology. *Memory & Cognition,* **25,** 425–437.

Smolensky, P. (1988). On the proper treatment of connectionism. *Behavioral and Brain Sciences,* **11,** 1–74.

Stanners, R., Neiser, J., Hernon, W., & Hall, R. (1979). Memory representation for morphologically related words. *Journal of Verbal Learning and Verbal Behavior,* **18,** 399–412.

Taft, M. (1979). Recognition of affixed words and the wordfrequency effect. *Memory & Cognition,* **7,** 263–272.

Wurzel, W. (1981). *Grundzuege einer deutschen Grammatik chapter: Phonologie: Segmentale Struktur* (pp. 898–990). Berlin: Akademie Verlag.

(Accepted March 30, 2000)