Probabilistic Approaches to Morphology

R. Harald Baayen

14th January 2002

1 Introduction

In structuralist and generative theories of morphology, probability is a concept that, until recently, has not had any role to play. By contrast, research on language variation across space and time has a long history of using statistical models to gauge the probability of phenomena such as t-deletion as a function of age, gender, education, area, and morphological structure. In this chapter, I discuss four case studies that illustrate the crucial role of probability even in the absence of sociolinguistic variation. The first case study shows that by bringing probability into morphological theory, the intuitive notion of morphological productivity can be made more precise. The second case study considers a data set that defies analysis in terms of traditional syntagmatic rules, but that can be understood as being governed by probabilistic paradigmatics. The third case study illustrates how the use of item-specific underlying features can mask descriptive problems that can only be resolved in probabilistic morphology. Finally, the fourth case study focuses on the role that probability plays in understanding morphologically complex words. However, before discussing the different ways in which probability emerges in morphology, it is useful to first ask the question of why probability theory, until very recently, has failed to have an impact in linguistic morphology, in contrast to, for instance, biological morphology.

To answer this question, consider the developments in information technology over the last 50 years, which have not been without consequences for the study of language. The computers of the fifties had comparatively reasonable computational capacity but very limited memory, on the order of 15KB (it was common then to count memory capacities in bits). For a program to work effeciently, it had to minimize storage. Programming languages such as Fortran, Cobol, Lisp, and Algol, were being developed, the latter being the first language (in 1958) with a formal grammar and the first language to allow recursive calling of functions.

With very few computers available for the academic community (by 1958, there were only some 2500 computers in use in the USA) many researchers had to carry out statistical analyses by hand, a tedious and laborious process — even though methods for applied statistics were generally designed to minimize calculations by adopting various kinds of simplifying assumptions (such as independence, linearity, and normality). Linguistic data in electronic form did not exist. Not surprisingly, the linguistic theories of the time took formal languages as the model for language, emphasizing the generative capacity of language, denying any role of importance to probability and statistics, and elevating economy of storage in memory to a central theorem.

Currently, the computers on our desk have vastly increased processing power and virtually unlimited memory. Carrying out a multivariate analysis is no longer a week's work. Statisticians have developed new, often computationally intensive, methods for analysing data that cannot be modeled adequately by the traditional statistical techniques (e.g., bootstrap and permutation methods). In addition, artificial neural networks have become standard tools in statistics. Some artificial neural network architectures have been found to be equivalent to existing statistical techniques. (See, for instance, Oja, 1982, for principal components analysis and Lebart, 1995, for correspondence analysis). Other network architectures have made it possible to estimate probabilities that cannot be calculated efficiently by standard analytical means (see, e.g., Mehta & Patel, 1986, and Clarkson, Fan, & Joe, 1993, for Fisher's exact test of indepen-

dence). Yet other network architectures such as feed-forward artificial neural networks provide genuinely new statistical tools for the flexible generalization of linear regression functions (see, e.g., Venables & Ripley, 1994, section 10.4).

Not only do we now have many more sophisticated statistical techniques, we also have an ever increasing amount of data. The early corpora for English, such as the Brown corpus (Kučera & Francis, 1967) comprised 1 million words, more recent corpora such as the British National Corpus (`http://info.ox.ac.uk/bnc/`), contain 100 million words, and the world wide web is enjoying increasing use as a data resource with for English an estimated 47,000,000,000 words in February 2000 (Grefenstette, 2000). Not surprisingly, these developments in technology and resources have left their mark in linguistics.

An area in linguistics in which these changes have had a very prominent impact is morphology. Early work on morphology in the generative framework focused on the properties of morphological rewrite rules (Aronoff, 1976, Selkirk, 1980), within a framework that, with the exception of Jackendoff (1975), proceeded on the assumption of a strict separation of the regular and the irregular. Although Bybee & Moder (1983) introduced morphological schemas to account for attraction phenomena among irregular forms in English, and although Bybee (1985) proposed to understand morphological phenomena in terms of similarities between stored representations in the lexicon, the study by Rumelhart and McClelland (1986a) turned out to be the most effective to challenge the classic view of morphology as a symbolic system. They showed that a very simple artificial neural network could map with a considerable degree of success English present tense forms on past tense forms without making a distinction between regular and irregular forms, and without formulating any explicit symbolic rules.

The original biological motivation for artificial neural networks stems from McCulloch & Pitts (1943). They published a seminal model of a neuron as a binary thresholding function in discrete time that has been very influential in the development of artificial neural networks. A report from 1948 (in Ince, 1992) shows that Alan Turing also developed the mathematics for networks of sim-

ple processing units (which he called unorganized machines) in combination with genetic algorithms (what he called a genetic search) as a model for understanding computation in the brain . This report did not attract attention at that time. It is only recently that it has become clear that Turing may have been the first connectionist (Copeland & Proudfoot, 1999), and it is only now that his ideas are implemented and studied with computer simulations (Teuscher, 2001). Real neurons are now known to be more complicated than the neurons of McCulloch and Pitts, Turing, or the ANNs used in statistics. The interest of the connectionist model for the creation of past tense forms of McClelland and Rumelhart (1986), therefore, resides not in the precise form of its network architecture, which is biologically implausible. The value of their study is that, by showing that a network of very simple processing units can perform a linguistic mapping, they have provided a powerful scientific metaphor for how neurons in the brain might accomplish linguistic mappings.

McClelland and Rumelhart's past-tense model met with fierce opposition. Pinker & Prince (1988) argued that it was fundamentally flawed in just about any conceivable way. Since then, the discussion has taken the shape of a stimulus-response series in which a paper in the symbolic tradition claiming that an artificial neural network cannot model a certain fact is followed by a study showing how that fact naturally follows once one adopts the right kind of connectionist architecture and training regime (see, e.g., MacWhinney & Leinbach, 1991; Plunkett & Juola, 2000). Brain imaging data from has been advanced as evidence against the connectionist account (Jaeger, Lockwood, Kemmerrer, Van Valin, & Murphy, 1996), without convincing the connectionist opposition (Seidenberg & Hoeffner, 1998).

Nevertheless, the symbolic position seems to be in something of a retreat. For instance, Pinker & Prince (1988) and Pinker (1991) flatly reject the connectionist approach. Pinker (1997) and Pinker (1999), however, allow for the possibility that irregulars are stored in some kind of associative memory, although the claim is maintained that language comprises a mental dictionary of memorized words on the one hand, and a mental grammar of creative rules on

the other. Marcus (2001) seems to go a step further by accepting connectionist models as enlightening implementational variants of symbolic systems. But he too claims that ANNs are incapable of explaining those crucial data sets that would reveal the supposedly symbolic nature of human language processing. Another index of the retreat of the narrow symbolic position is the emergence of stochastic optimality theory (Boersma, 1998; Zuraw, 2000), an extension of optimality theory incorporating a mechanism accounting for non-deterministic data, and of interpretations of optimality theory in which similarity spaces and attractors play a crucial role (Burzio, 2002).

There are at least three issues that play a role in the controversy between the connectionist and symbolic positions. The first issue is whether human cognition and language as a cognitive faculty are fundamentally symbolic in a nature. This is an issue about which I will remain agnostic.

A second issue is whether artificial neural networks are appropriate models for language. For instance, should ANNs be able to generalize outside the scope of their training space, as argued by Marcus (2001)? The answers to questions such as this depend on a host of assumptions about learning and generalizability in animals, primates, and humans, questions that go way beyond my competence and the scope of this chapter.

A third issue that is at stake here is whether language is, in its core, a deterministic phenomenon (one that can be handled by simple symbolic rules) or a probabilistic phenomenen (one for which such simple symbolic rules are inadequate). This is the issue addressed in this chapter. I will argue that the role of probability in morphology is far more pervasive than standard textbooks on morphology would lead one to believe. However, this chapter will not be concerned with how artificial neural networks deal with non-deterministic data, for two reasons. First, a good introduction to neural network theory requires a chapter of its own (see, for instance, McLeod, Plunkett, & Rolls, 1998). Second, given that the neural networks used to model language are *artificial* neural networks providing abstract statistical models for linguistic mapping problems, it makes sense to consider a broader range of statistical tools available at present

for understanding the quantitative structure of such problems. From this perspective, ANNs pair the advantage of maximal flexibility with the disadvantage of requiring considerable time with respect to training the model on the one hand, and a loss of analytical user control on the other: To understand how an ANN achieves a particular mapping itself requires application of conventional multivariate statistical techniques.

What I will therefore discuss in this chapter is some quantitative techniques for coming to grips with the probabilistic structure of morphological phenomena that, unlike ANNs, do not require extensive training time and that provide immediate insight into the quantitative structure of morphological data. I offer these techniques to the reader as useful analytical tools, without committing myself to any of these techniques as 'models of the mind'.

However, I will also indulge in speculating how these techniques might be articulated in terms of the spreading activation metaphor, the current gold standard in psycholinguistics for modeling lexical processing. I indulge in these speculations in order to provide some indication as to how the mental lexicon might deal with probabilistic phenomena without invoking complex statistical calculations. Those committed to a connectionist approach will have no difficulty reformulating my symbolic spreading activation models at the subsymbolic level. Those committed to optimality theory will find that my models can be reformulated within stochastic optimality theory. Both kinds of reformulation, however, come with the cost of increased complexity in terms of the numbers of formal parameters required to fit the data.

The remainder of this chapter is structured as follows. Section 2 illustrates how probability theory can help to understand a key issue in morphological theory, the enigmatic phenomenon of morphological productivity. Section 3 discusses two data sets illustrating the role of probability in the production of complex words. Finally, section 4 considers the role of probability during the comprehension of morphologically complex words.

2 Probability and productivity

Aronoff (1976) described productivity as one of the central mysteries of derivational morphology. What is so mysterious about productivity is not immediately evident from the definition given by Schultink fifteen years earlier. According to Schultink (1961), productivity is the possibility available to language users to coin, unintentionally, a number of formations which are in principle uncountable. The empirical problem that makes productivity so mysterious is that some word formation rules give rise to small numbers of words, while other word formation rules give rise to large numbers of formations. In English, there are many words in *-ness* (*goodness*), there are fewer words in *-ee* (*employee*), and hardly any in *-th* (*warmth*). As soon as we start observing how often different kinds of word formation patterns are realized, we see that productivity is graded or scalar in nature, with productive word formation at one extreme, semi-productive word formation in the middle, and unproductive word formation at the other extreme.

Productivity becomes an even more enigmatic notion once it is realized that unproductive word formation patterns can be fully regular (for instance, the Dutch suffix *-in* that creates female agent nouns, as in *boer*, 'farmer', *boerin*, 'female farmer'), while word formation need not be rule-governed in the traditional sense to be productive (see the discussion of linking elements in Dutch in section 3 below).

Some researchers (Schultink, 1961; and recently Bauer, 2001) have argued for a principled distinction between productive and unproductive word formation. An unproductive affix would be 'dead', it would not be part of the grammar. Productive affixes, on the other hand, would be 'alive', and the question of degrees of productivity would only arise for such living affixes. The problem with this view is that in practice it is very difficult to know whether an affix is truly unproductive. Consider, for instance, the following quote from (Bauer, 2001:206).

> Individual speakers may coin new words which are not congru-

> ent with currently predominating customs in the community as a whole. *The Oxford English Dictionary* credits Walpole with coining *gloomth* and *greenth* in the mid-eighteenth century, some 150 years after the end of the period of societal availability for *-th*; *greenth* appears to have survived into the late nineteenth century, but neither is now used, and neither can be taken to illustrate genuine productive use of *-th*.

Interestingly, a simple query of the world wide web reveals that words such as *coolth, greenth*, and even *gloomth* are used by speakers of English, even though *-th* is one of the well-worn examples of a supposedly completely unproductive suffix in English. Consider the following examples of the use of *coolth*.

(1) *Coolth, once a nonce word made on analogy with warmth, is now tiresomely jocular: The coolth of the water in the early morning is too much for me. Kenneth G. Wilson (1923?). The Columbia Guide to Standard American English.* 1993.
`www.bartleby.com/68/5/1505.htm`

(2) *Increase the capacity of your house to store coolth. (Yes, it is a real word.) Using the mass in the house to store coolth in the summer and heat in the . . .*
`www.tucsonmec.org/tour/tech/passcool.htm`

(3) *The combination of high-altitude and low-latitude gives Harare high diurnal temperature swings (hot days and cool nights). The team developed a strategy to capture night-time coolth and store it for release during the following day. This is achieved by blowing night air over thermal mass stored below the verandah's . . .*
`www.arup.com/insite/features/printpages/harare.htm`

(4) *Do we see the whiteness of the snow, but only believe in its coolth. Perhaps this is sometimes so; but surely not always. Sometimes actual coolth is . . .*
`www.ditext.com/sellars/ikte.htm`

(5) *Early drafts of Finnegans Wake- HCE . . . behaved in an ungentlemanly manner opposite a pair of dainty maidservants in the greenth of the rushy hollow, whither, or so the two gown and pinners pleaded . . .*
`www.robotwisdom.com/jaj/fwake/hce.html`

(6) *Macom — Garden realization ... realization 3. Delivery of carpet lawn - Fa Kotrba 4. Maintainance of greenth a) chemical treatment; weeding out; fertilization and plant nutrition; prevention of . . .*
`www.macom.cz/english/service.htm`

(7) *This year I discovered the Gothic novel. The first Gothic novel I read was "Melmoth the Wanderer." I read all 697 pages in about five days it was so good. In the Penguin Classics introduction to "Melmoth" it mentions other Gothic novels such as "The Italian," "Vathek" and "The Castle of Otranto." All of which I've since read and have discovered a new genre of fiction which I really enjoy. I've also had a new word added to my vocabulary: "Gloomth."*
`www.geocities.com/prozacpark/gothnovel.htm`

Example (1) is an example of the prescriptive view, mirroring on the web the quote from Bauer (2001). The second example shows how *coolth* is used to fill the lexical gap in the series *hot/heat, warm/warmth, cool/***coolth**, *cold/cold*. It is a technical term introduced to the reader as a real word of English. The writer of example (3) takes the use of *coolth* for granted, and the writer of example (4), in a discussion of Kant's theory of experience, seems to find the non-technical use of *coolth* unproblematic. Examples (5) and (6) illustrate the use of *greenth*, and the last example shows how modern speakers can even enjoy learning about *gloomth*. What these examples show is that forms such as *coolth*, *greenth*, and *gloomth* are occasionally used in current English, testifying to the residual degree of productivity of *-th* and the graded, scalar nature of productivity.

At first sight, it would seem that the degree of productivity of a word for-

mation pattern might be captured by counting the number of distinct formations, henceforth word types. The problem with type frequency as a measure of productivity is that unproductive patterns may comprise more types than productive patterns. In Dutch, for instance, the suffix *-elijk* occurs more often than the prefix *-her* (see, e.g., Baayen, 2001), but it is the latter and not the former which is generally judged to be productive. What we need, then, is a measure that captures the probability of new words, independently of the number of words that are already attested.

Two measures that formalize the notion of degree of productivity in terms of probability are available (Baayen & Renouf, 1996; Baayen 2001). They are based on the probability theory of the number of different species (types) observed among a given number of observations (tokens). First consider the 'productivity' of a fair dice. There are six types: 1, 2, 3, 4, 5, and 6. Imagine how many different types we count as we throw a fair dice 100 times. How many of these types may we expect to have been seen after N throws? In other words, how does the expected vocabulary size $\mathrm{E}[V(N)]$ increase as a function of the sample size N? The growth curve of the vocabulary size for a fair dice is shown in the upper left panel of Figure **??** using a solid line. The vertical axis plots the expected count of types. The horizontal axis plots the number of throws, i.e., the individual observations or tokens. The growth curve of the vocabulary shows that after 40 throws we are almost certain to have seen each side of the dice at least once. In fact, each type will probably have been counted more than once. This is clear from the dotted line in the graph, which represents the growth curve $\mathrm{E}[V(1,N)]$ of the hapax legomena, the types which occur exactly once in the sample. After 40 trials, it is very unlikely that there is a type left in the sample that has been observed only once.

Now consider the upper right panel of Figure **??**. Here we see the corresponding plot for some 6 million words from the British National Corpus (its context-governed subcorpus of spoken British English). Instead of throwing a dice, now imagine that we are reading through this subcorpus, word token by word token, keeping track of the number of different word types, and also

counting the hapax legomena. The upper right panel of Figure **??** shows that both the growth curve of the vocabulary and the growth curve of the hapax legomena increase as we read through the corpus. There is no sign of the growth curve of the vocabulary reaching an asymptote, as in the case of the fair dice. There is also no indication of the growth curve of the hapax legomena having an early maximum with the X-axis as asymptote as N is increased. This pattern is typical for word frequency distributions, irrespective of whether one is dealing with small texts of a few thousand words, or with huge corpora with tens or even hundreds of millions of words. It is also typical for the word frequency distributions of productive affixes. For instance, the nouns in *-ness* in the context-governed subcorpus of the BNC are charaterized by the growth curves of the vocabulary and the hapax legomena in the lower right panel. Conversely, the pattern show for the fair dice represents a frequency distribution prototypical for unproductive affixes. For instance, the nouns in *-th* in the context-governed subcorpus of the BNC have the growth curves shown in the lower left panel of Figure **??**.

PLACE FIGURE **??** APPROXIMATELY HERE

It turns out that the rate $P(N)$ at which the vocabulary size $V(N)$ increases is a simple function of the number of hapax legomena $V(1, N)$ and the sample size N:

$$P(N) = \frac{\mathrm{E}[V(1, N)]}{N} \tag{1}$$

(see, e.g., Good, 1953, or Baayen, 2001). Note that the growth rate of the vocabulary size is itself a function of the sample size. The rate at which the vocabulary size increases decreases through sampling time. Initially, nearly every word is new, but as we read on through the corpus, we will see more and more words that we have encountered before. Also note that the upper left panel of Figure **??** provides a clear illustration of the relation between the growth rate $P(N)$ and the growth curve of the number of hapax legomena $V(1, N)$. After 50 throws, the growth curve of the vocabulary is, at least to the eye, completely flat. After 50 throws, therefore, the growth rate of the vocabulary should be

very close to zero. Since after 50 throws the number of hapax legomena has become practically zero, $P(N)$ must also be zero, as required.

The growh rate $P(N)$ has a simple geometric interpretation: It is the slope of the tangent to the growth curve of the vocabulary size at sample size N. The growh rate $P(N)$ is also a probability, namely, the probability that, after having sampled N tokens, the next token to be sampled will represent a type that has not been observed among the previous N tokens. To see this, consider an urn with a fixed number of marbles. Each marble has one color, there are V different colors, N marbles, and $V(1)$ marbles with a unique color, i.e., with a color that no other marble has. The probability that the first marble drawn from the urn will have a color that will not be sampled again is $V(1)/N$. Since there is no reason to suppose that sampling the last marble will be different from sampling the first marble, the probability that the very last marble taken from the urn represents a color that has not been seen before must also be $V(1)/N$. This probability approximates the probability that, if marbles from the same population are added to the urn, the first added marble drawn from the urn will have a new color. The expectation operator in (1) makes this approximation precise.

In order to derive productivity measures from the growth rate of the vocabulary size, we consider the case where we sample a new token after having read through N tokens. Let $\{\mathrm{A}\}$ denote the event that this token represents a new type. We furthermore regard the vocabulary as a whole to be a mixture of C different kinds of words: various kinds of monomorphemic words (simplex nouns, adjectives, pronouns, ...), and many different kinds of complex words (compounds, nouns in *-ness*, verbs in *-ize*, adverbs in *-ly*, ...). Let $\{\mathrm{B}\}$ denote the event that $N+1$-th token belongs to the i-th mixture component of the vocabulary. The hapax-conditioned degree of productivity $\mathcal{P}^*(N,i)$ of the i-th mixture component is the conditional probability that the $N+1$-th token belongs to the i-th mixture component, given that it represents a type that has not been observed before. Let $V(1,N,i)$ denote the number of hapax legomena belonging to the i-th mixture component, observed after N tokens have been

sampled.

$$\begin{aligned}
\mathcal{P}^*(N,i) &= \mathrm{P}(\{B\}|\{A\}) \\
&= \frac{\mathrm{P}(\{B\} \cap \mathrm{P}(\{A\})}{\mathrm{P}(\{A\})} \\
&= \frac{\frac{\mathrm{E}[V(1,N,i)]}{N}}{\frac{\sum_{j=1}^{C} \mathrm{E}[V(1,N,j)]}{N}} \\
&= \frac{\mathrm{E}[V(1,N,i)]}{\mathrm{E}[V(1,N)]}. \qquad (2)
\end{aligned}$$

For *-th* and *-ness*, the values of $\mathcal{P}^*$ are 1.0e-13/32042 = 3.1e-18 and 158/32042 = 0.0049 respectively. Note that for spoken British English, the probability of observing new formations in *-th* is vanishingly small. It is probably only for written English that an extremely large corpus such as the world wide web (for which $N > 47000000000$, Grefenstette, 2000) succeeds in showing that there are unobserved formations, i.e., that there is some very small residual productivity for *-th*.

The category-conditioned degree of productivity $\mathcal{P}(N,i)$ of mixture component i is the conditional probability that the $N+1$-th token represents a new type, given that it belongs to mixture component (or morphological category) i. With N_i the number of tokens counted for the i-th mixture component, we have:

$$\begin{aligned}
\mathcal{P}(N,i) &= \mathrm{P}(\{A\}|\{B\}) \\
&= \frac{\mathrm{P}(\{A\} \cap \mathrm{P}(\{B\})}{\mathrm{P}(\{B\})} \\
&= \frac{\frac{\mathrm{E}[V(1,N,i)]}{N}}{\frac{N_i}{N}} \\
&= \frac{\mathrm{E}[V(1,N,i)]}{N_i}. \qquad (3)
\end{aligned}$$

Applied to *-th* and *-ness*, we obtain 1.0e-13/3512 = 2.8e-17 and 158/3813 = 0.04 as estimates of $\mathcal{P}$ for *-th* and *-ness* respectively.

To understand the difference between the interpretations of $\mathcal{P}$ and $\mathcal{P}^*$, it is important to realize that productivity is determined by a great many factors,

ranging from structural constraints and processing constraints to register and modality (Bauer, 2001; Plag, Dalton-Puffer, & Baayen, 1999). Since $\mathcal{P}^*$ estimates the contribution of an affix to the growth rate of the vocabulary as a whole, it is a measure that is very sensitive to the different ways in which non-systemic factors may affect productivity. For instance, the suffix *-ster* attaches productively to Dutch verbs to form female agent nouns (*zwem-er*, 'swimmer', *zwemster*, 'female swimmer'). However, even though *-ster* is productive, speakers of Dutch are somewhat hesitant to use it. Consequently, its contribution to the overall growth rate of the vocabulary is quite small.

The category-conditioned degree of productivity of a given affix does not take counts of other affixes into consideration. This measure is strictly based on the morphological category of the affix itself. It estimates its productivity, independently of the non-systemic factors. Hence, it provides a better window on the potentiality of the affix. Measured in terms of $\mathcal{P}$, *-ster* emerges with a high degree of productivity (see Baayen, 1994, for experimental validation).

The prominent role of the hapax legomena in both productivity measures makes sense from a processing point of view. The more frequent a complex word is, the more likely it is that it is stored in memory and the less likely it is that its constituents play a role during production and comprehension (Hasher & Zacks, 1984; Scarborough, Cortese, & Scarborough, 1977; Bertram, Schreuder, & Baayen, 2000). Conversely, the more infrequent words there are with a given affix, the more likely it is that its structure will be relevant during comprehension and production. The number of hapax legomena, the lowest-frequency words in the corpus, therefore provide a first approximation of the extent to which the words with a given affix are produced or accessed through their constituents.

Hay (2000) provides a more precise processing interpretation for the category-conditioned degree of productivity. This study brings together the insight that phonological transparency co-determines productivity and the insight that relative frequency is likewise an important factor. First consider phonological transparency. The more the phonological form of the derived word masks its

morphological structure, the less such a form will contribute to the productivity of the morphological category to which it belongs (see, e.g., Cutler, 1981; Dressler, 1985). Hay operationalizes the role of phonological transparency in terms of offset-onset probabilities, and then shows that the resulting juncture probabilities predict other properties of the words in which they occur, such as prefixedness ratings, semantic transparency ratings, and number of meanings.

Next consider relative frequency. The idea here is that the frequency relation between a derived word and its base should co-determine the parsability of that word. If the frequency of the derived word is substantially greater than that of its base, it is unlikely that the base will effectively contribute to the processes of production and comprehension. If, on the other hand, the frequency of the derived word is much lower than the frequencies of its constituents, it is much more likely that these constituents do have a role to play. Hay (2000) shows that relative frequency predicts pitch accent placement: Prefixes in word for which the derived frequency is greater than the frequency of the base are less likely to attract pitch accent than words for which the derived frequency is less than the base frequency. She also shows that t-deletion is more likely to occur in case the derived word is more frequent than its base. Finally, she shows that complexity ratings for such words tend to be lower than for words for which base frequency exceeds derived frequency.

Interestingly, Hay demonstrates that for a sample of twelve English derivational affixes the category-conditioned degree of productivity is a linear function of mean relative frequency and mean juncture probability of the formations in the corresponding morphological categories. In other words, the probability that a morphological category will give rise to new formations emerges as being demonstrably co-determined by the juncture probabilities of its members and the frequency relations between these members and their base words. Hay and Baayen (2002) provide a more detailed analysis of the correlation between relative frequency and the two productivity measures $\mathcal{P}$ and $\mathcal{P}^*$ for 80 English derivational affixes. Their results suggest that the degree of productivity of an affix correlates surprisingly well with the likelihood that it will be

parsed in comprehension.

Having outlined how probability theory can help to come to grips with the elusive notion of degrees of productivity, we now turn to consider the possibility that morphological regularity itself is probabilistic in nature.

3 Probability in morphological production

In this section, I introduce two data sets illustrating the role of probability in the production of morphologically complex words. The first data set concerns the production of linking elements in Dutch nominal compounds. The second data set addresses the selection of voice specification of syllable final obstruents in Dutch.

3.1 Linking elements in Dutch

The immediate constituents of nominal compounds in Dutch are often separated by what I will refer to as a linking element. Whether a linking element should be inserted, and if so, which linking element, is difficult to predict in Dutch. To see this, consider the compounds in (1).

(1)	*schaap-herder*	"sheep-herder"	'shepherd'
	schaap-S-kooi	"sheep-S-fold"	'sheepfold'
	schaap-EN-vlees	"sheep-EN-meat"	'mutton'

The same left constituent appears without a linking element, with the linking element *-s-*, and with the linking element *-en-*. Intensive study of this phenomenon has failed to come up with a set of rules that adequately describe the distribution of linking elements in Dutch compounds (see Krott, Baayen, & Schreuder, 2001, for discussion and further references). This suggests that the appearance of linking elements is fairly random, and that the use of linkers is unproductive. This is not the case, however. Linking elements are used productively in novel compounds, and there is substantial agreement among

speakers as to which linking element is most appropriate for a given pair of immediate constituents. The challenge that the linking elements of Dutch pose for linguistic theory is how to account for the paradox of an apparently random morphological phenomenon that nevertheless is fully productive.

The key to solving this paradox is to exchange a syntagmatic approach for a paradigmatic approach, and to exchange greedy learning for lazy learning. A syntagmatic approach assumes that it is possible to formulate a generalization describing the properties that the context should have for a given linking element to appear. A paradigmatic approach assumes that the set of compounds similar to the target compound requiring the possible insertion of a linking element, its compound paradigm, forms the analogical basis from which the probabilities of the different linking elements are derived. The syntagmatic approach is most often coupled with greedy learning, in the sense that once the generalization has been abstracted from a set of examples, these examples are discarded. In fact, researchers working in the tradition of generative grammar tend to believe that learning a rule and forgetting about the examples that allowed the rule to be deduced go hand in hand. Pinker (1991, 1997, 1999), for instance, has argued extensively for a strict division of labor between rules accounting for what is productive and regular on the one hand, and storage in memory for what is unproductive and irregular on the other hand. Conversely, the paradigmatic, analogical approach is based on the insight that learning may involve a continuous process driven by an ever-increasing instance base of exemplars. In this approach, it may even be harmful to forget individual instances. This kind of learning, then, is lazy in the sense that it does not attempt to formulate a rule that allows the data to be discarded. Greedy learning, once completed, requires little memory. By contrast, lazy learning assumes a vast storage capacity.

Two mathematically rigorously defined approaches to the modeling of paradigmatic analogy are available: Analogical Modeling of Language (AML, Skousen 1989, 1993), and the many machine learning algorithms implemented in the TiMBL program of Daelemans, Zavrel, van der Sloot, & van den Bosch (2000).

Both AML and TIMBL determine the choice of the linking element for a given target compound on the basis of the existing compounds that are most similar to this target compound. The two methods differ with respect to what counts as a similar compound. AML makes use of a similarity metric that also plays a role in quantum mechanics (Skousen, 2000). We will return to AML below. In this section, I describe the IB1-IG metric (Aha, Kibler, & Albert, 1991; Daelemans, Van den Bosch, & Weijters, 1997) available in TIMBL.

In formal analogical approaches, the question of which linking element to choose for a compound amounts to a classification problem: Does this compound belong to the class of compounds selecting *-en-*, to the class of compounds selecting *-s-*, or to the class of compounds with no overt linking element, for notational convenience henceforth the compounds with the linking element *-0-*. In order to establish which class a compound for which we have to determine the linking element belongs to, we need to define the properties of compounds on which class assignment has to be based. In other words, we need to know which features are relevant, and what values these features might have. Now consider Table 1, which lists a hypothetical instance base with 10 compounds. In this example, there are 5 features: the modifier, the head, the nucleus of the modifier, the onset of the head, and the coda of the head. The values of the feature 'Nucleus' are the vowels *aa, a, ou,* and *e*. The values of the feature 'Modifier' are the left constituents *schaap, lam, paard, koe,* and *varken*. What we want to know is what the most probable linking element is for the novel compound *schaap-?-oog* (sheep's eye).

PLACE TABLE 1 APPROXIMATELY HERE

To answer this question, we need to know which exemplars in the instance base are most relevant. We want to discard exemplars that are very different from the target compound, and we want to pay special attention to those compounds that are very similar. In other words, we need a similarity metric, or, alternatively, a distance metric. The IB1-IG distance metric is based on a simple distance metric (known as the simple matching coefficient or Hamming dis-

tance) that tracks the number of features that have different values. Let X denote the target compound, and let Y denote an exemplar in the instance base. The Hamming distance between these two compounds, $\Delta(X, Y)$, is defined as the number of features with mismatching values:

$$\Delta(X, Y) = \sum_{i=1}^{n} \mathrm{I}_{[x_i \neq y_i]}. \tag{4}$$

In the present example, the number of features n equals 5. The value of the i-th feature of X is denoted by x_i. The operator $\mathrm{I}_{[z]}$ evaluates to 1 if the expression z is true, and to 0 otherwise. This metric allows us to group the compounds in the instance base according to their distance to the target compound. For instance, *schaap-en-bout* is at distance 3 from the target *schaap-?-oog*, because these two compounds mismatch with respect to the features Head, Onset(2), and Coda(2). We can now determine the set $\mathcal{S}$ of compounds that, given the features and their values, are most similar to the target compound. The distribution of linking elements in this set of nearest neighbors determines the probabilities of selection. Denoting the cardinality of $\mathcal{S}$ by S, the probability of the linking element *-en-* is given by the proportion of compounds in $\mathcal{S}$ that select *-en-*:

$$\mathrm{P}(L = \text{-}en\text{-}) = \sum_{i=1}^{S} \frac{\mathrm{I}_{[L_i = \text{-}en\text{-}]}}{S}. \tag{5}$$

The IB1-IG distance measure (Daelemans, Van den Bosch, & Weijters, 1997) improves considerably upon (4) by weighting the features for their relevance, using the information-theoretic notion of entropy. The entropy $H(L)$ of a distribution of linking elements L, with J different linking elements,

$$H(L) = -\sum_{j=1}^{J} p_j \log_2 p_j, \tag{6}$$

is a measure of uncertainty about which linking element to choose in the situation that no information is available about the values of the features of a given word. For the data in Table 1,

$$H(L) \quad = \quad -[\mathrm{P}(L = en) \log_2(\mathrm{P}(L = en)) + \mathrm{P}(L = s) \log_2(\mathrm{P}(L = s)) +$$

$$
\begin{aligned}
& + \mathrm{P}(L = \emptyset) \log_2(\mathrm{P}(L = \emptyset))] \\
= \; & -[\, 0.4 \log_2(0.4) + 0.5 \log_2(0.5) + 0.1 \log_2(0.1)] \\
= \; & 1.36.
\end{aligned}
$$

The degree of uncertainty changes when extra information is provided, for instance, the information that the value of the feature Modifier is *schaap*:

$$
\begin{aligned}
& H(L|\text{Modifier} = \textit{schaap}) = \\
= \; & -[\mathrm{P}(L = \textit{en}|\text{Modifier} = \textit{schaap}) \log_2(\mathrm{P}(L = \textit{en}|\text{Modifier} = \textit{schaap})) + \\
& + \mathrm{P}(L = s|\text{Modifier} = \textit{schaap}) \log_2(\mathrm{P}(L = s|\text{Modifier} = \textit{schaap})) + \\
& + \mathrm{P}(L = \emptyset|\text{Modifier} = \textit{schaap}) \log_2(\mathrm{P}(L = \emptyset|\text{Modifier} = \textit{schaap}))]] \\
= \; & -[0.5 * \log_2(0.5) + 0.25 * \log_2(0.25) + 0.25 * \log_2(0.25)] \\
= \; & 1.5.
\end{aligned}
$$

We can gauge the usefulness or weight w_i of a feature F_i for predicting the linking element by calculating the probability-weighted extent to which knowledge of the value v of F_i decreases our uncertainty:

$$
w_i = H(L) - \sum_{v \in F_i} \mathrm{P}(v) H(L|v). \tag{7}
$$

For the feature Modifier, v ranges over the values *schaap, lam, paard, koe,* and *varken*. Note that when v has the value *schaap*, the probability $\mathrm{P}(v)$ equals $4/10$. Because $H(L|v) = 0$ when $v \neq \textit{schaap}$ (the other modifiers all occur with just one linking element, so there is absolute certainty about the appropriate linking element in these cases), the information gain weight for the feature Modifier is

$$
\begin{aligned}
w_{\text{Modifier}} &= H(L) - \mathrm{P}(\text{Modifier} = \text{schaap}) H(L|\text{Modifier} = \textit{schaap}) \\
&= 1.36 - 0.4 * 1.5 \\
&= 0.76.
\end{aligned}
$$

The feature with the lowest information gain weight is Coda(2), which is not surprising as there is no obvious phonological reason to suppose the coda of

the second constituent to codetermine the choice of the linking element. Crucially, when applied to a realistic instance base, the information gain weights are a powerful means for establishing which features are important for understanding the quantitative structure of a data set.

When applied to an instance base of Dutch compounds as available in the CELEX lexical database (Baayen, Piepenbrock, and Gullikers, 1995), it turns out that, of a great many features, the Modifier and Head features have the highest information gain values (1.11 and 0.41) respectively, and that other features, such as whether the first constituent bears main stress, have a very low information gain weight (0.07). When we modifiy our distance metric by weighting for information gain,

$$\Delta(X, Y) = \sum_{i=1}^{n} w_i \mathrm{I}_{[x_i \neq y_i]}, \tag{8}$$

and when we choose the linking element on the basis of the distribution of linking elements in the set of compounds with the smallest distance Δ, some 92% of the linking elements in Dutch compounds are predicted correctly (using ten-fold cross-validation). Experimental studies (Krott, Baayen, & Schreuder, 2001a, Krott, Schreuder, & Baayen, 2001b) have confirmed the crucial importance of the Modifier and Head features. Apparently, the compounds sharing the modifier constituent (the left constituent compound family), and to a lesser extent the compounds sharing the head constituent (the right constituent compound family), form the analogical exemplars on which the choice of the linking element in Dutch is based. What we have here is paradigmatically determined selection instead of syntagmatically determined selection. Instead of trying to predict the linking element on the basis of specific feature values of the surrounding constituents (its syntagmatic context), it turns out to be crucial to zoom in on the constituents themselves and the distributional properties of their positional compound paradigms.

PLACE FIGURE **??** APPROXIMATELY HERE

To see how such paradigmatic effects might be accounted for in a psycholinguistic model of the mental lexicon, consider Figure **??**. Figure **??** outlines the functional architecture of a spreading activation model for paradigm-driven analogy using the small instance base of Table 1. At the left hand side of the graph, the modifier (labeled LEFT) and the head constituent (labeled RIGHT) are shown. These two lexemes are connected to their positional compound paradigms, listed in the center. The weights on the connections to the compounds in the instance base are identical to the information gain weights of the left and right constituents. The different linking elements are displayed at the right hand side of the graph. Each linking element is connected with the compounds in which it appears. Activation spreads from the left and right constituents to the compounds in the paradigmatic sets, and from there to the linking element. The linking element that receives the most activation is the one selected for insertion in the novel compound *schaap-?-oog*. Krott, Schreuder, and Baayen (2001b) show that an implemented computational simulation model along these lines provides excellent fits to both the choices and the times required to make these choices in experiments with novel Dutch compounds.

This example shows that morphological network models along the lines proposed by Bybee (1985, 1995a, 2001) can be made precise and that, once formalized, they have excellent predictive power. It should be kept in mind, however, that the present model presupposes considerable structure in the mental lexicon, both in terms of the specific connectivity required and in terms of the specific information gain weights on these connections. In this light, it makes more sense to speak of an analogical or paradigmatic rule for the selection of the linking element rather than of a network model, because we are dealing not with undifferentiated connectivity in an encompassing network for the whole mental lexicon, but with highly structured connectivity in a sub-network dedicated to the specific task of selecting the appropriate linking element for Dutch compounds.

The next section provides a second example of a phenomenon that turns out to be analogical in nature, the morphophonology of final obstruents in Dutch.

3.2 Syllable-final obstruents in Dutch

The feature [voice] is distinctive in Dutch. This is illustrated in (2) for the Dutch nouns */rat-en/* and */rad-en/*. When the alveolar stop is voiceless, the noun means 'honey combs', when the stop is voiced, the noun means "councils". When the stop is in syllable-final position, as in isolated singular forms, the distinction between the voiced and voiceless obstruent is neutralized, and in phrase-final position both obstruents are realized as voiceless.

(2)

FORM	TRANSLATION	VOICING
/rat-en/	'honey comb'-PLURAL	voiceless
/rat/	'honey comb'	voiceless
/rad-en/	'council'-PLURAL	voiced
/rat/	'council'	voiceless

Traditionally, this phenomenon is accounted for by assuming that the obstruents in */rat-en/* and */rad-en/* are specified as being underlyingly voiceless and voiced respectively, with a rule of syllable-final devoicing accounting for the neutralization of the voice distinction in the singular (e.g., Booij, 1995). Whether the final obstruent in a given word alternates between voiced and voiceless is taken to be an idiosyncratic property of that word that has to be specified lexically, although it has been noted that fricatives following long vowels tend to be underlyingly voiced and that bilabial stops tend to be voiceless following long vowels (Booij, 1999).

Ernestus & Baayen (2001a) report that there is far more structure to the distribution of the voice specification of final obstruents in the lexicon of Dutch than expected on the basis of this standard analysis. Figure **??** summarizes some of the main patterns in the data for three major rime patterns by means of the barplots at the left hand side. The data on which these graphs are based are some 1700 monomorphemic words attested in the Dutch part of the CELEX lexical database. These words are nouns, verbs, or adjectives ending in an obstruent that has both voiced and voiceless counterparts in Dutch, and that are attested with a following schwa-initial suffix. For these words, we there-

fore know whether they have an alternating or a non-alternating final obstruent. The top left panel of Figure **??** plots the percentage of words exhibiting voice alternation (% voiced) as a function of the kind of obstruent (bilabial (P) and alveolar (T) stops, labiodental (F), alveolar (S), and velar (X) fricatives) for words with a rime consisting of a short vowel followed by a sonorant consonant, followed by the final obstruent. Note that as we proceed from left to right, the percentage of words with underlying voiced obstruents increases. The center left panel shows a fairly similar pattern for words ending in a long vowel that is directly followed by the final obstruent without any intervening consonant. The bottom left panel visualizes the distribution for words ending in a short vowel immediately followed by the final obstruent. For these words, we observe a u-shaped pattern. (A very similar pattern characterizes the subset of verbs in this database of monomorphemic words.) Ernestus and Baayen show that the quality of the vowel, the structure of the coda (does a consonant precede the final obstruent, and if so, is it a sonorant), and the type of final obstruent, are all significant predictors of the distribution of the percentage of voicing in Dutch.

PLACE FIGURE **??** APPROXIMATELY HERE

The right panels of Figure **??** present the corresponding barplots for the data obtained in a production experiment in which participants were asked to produce the past tense form for some 200 artificially created but phonotactically legal pseudo-verbs. The past tense suffix was selected because it has two allomorphs the selection of which depends on whether the final obstruent alternates. If the final obstruent alternates, the past tense suffix has the form *-de* and the final obstruent is realized as voiced. If the final obstruent does not alternate, the appropriate past tense suffix is *-te*, and the final obstruent is realized as voiceless. By asking participants to produce the past tense form, the status assigned to the final obstruent of the pseudo-verb, alternating or not alternating, can be determined simply on the basis of the form of the past-tense suffix.

Interestingly, the percentages of verbs for which the participants used the past tense suffix *-de* reflect to a considerable degree the percentages of the words with alternating final obstruents in the lexicon shown in the left panels. Note that even the u-shaped pattern in the lower left panel is present to some extent in the lower right panel. This pattern of results is incompatible with theories that hold that the selection of the past tense suffix crucially depends on the availability of a lexically specified feature marking the verb as underlyingly voiced. After all, the participants in the experiment were asked to produce the past tense for pseudo-verbs, forms that are not available in the lexicon and for which no such lexically specified feature is available. We are therefore faced with the question how the participants might have arrived at their choice of the allomorph of the past tense suffix for these pseudo-verbs.

PLACE FIGURE **??** APPROXIMATELY HERE

Figure **??** illustrates the kind of lexical connectivity required for a spreading activation network to predict the choice of the past tense allomorph. Completely analogous to the spreading activation model outlined in Figure **??** for modeling the choice of the linking element in Dutch compounds, activation spreads from the phonological feature values (left) to the words sharing these feature values (center) and from there to the voicing specification of the final obstruent and of the past tense allomorph (right). As before, this model embodies an analogical, paradigmatic rule. It presupposes that vowel length, coda structure, and type of obstruent can be identified for any given input form, and that the speaker has learned that it is these features that are primarily relevant for the voicing alternation. Once the values of these features are activated, the paradigms of words in the lexicon sharing these feature values are co-activated, proportionally to the weights $w_1, w_2, \ldots$. The support from the weighted paradigmatic cohorts determines the probability of selecting [-voice] or [+voice]. Note that this probability is **not** determined by the proportion of words with exactly the same values as the target for the features vowel length, coda structure, and type of obstruent. Words which share only two feature val-

ues, and even words that share only one feature value, also co-determine the probabilities of a voiced or voiceless realization.

A formal definition of these probabilities proceeds as follows. Let F denote the number of features, and let $\bar{v}$ denote the vector

$$\bar{v} = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_F \end{pmatrix}$$

specifying for the target word the value v_i of each feature F_i. We define n_{ijk} as the number of lexemes with value j for feature i that support exponent k, the exponents in this example being the voicing outcomes Voiced and Voiceless. The support s_k that exponent k receives given target $\bar{v}$ and weights w equals

$$s_k = \sum_{i=1}^{F} w_i n_{iv_i k}, \tag{9}$$

and the probability p_k that it will be selected, given K different exponents, is

$$p_k = \frac{s_k}{\sum_{m=1}^{K} s_m}. \tag{10}$$

The maximum likelihood choice of this spreading activation model is the exponent for which p_k is maximal. This maximum likelihood choice is the choice that the model predicts that the majority of the participants should opt for.

When we set the weights w to the information gain weights (7), the maximum likelihood prediction of the model coincides with the majority choice of the participants in 87.5% of the experimental pseudoverbs. When we optimize the weights using the simplex algorithm of Nelder & Mead (1965), this accuracy score improves to 91.7%. The by-word probabilities for voicelessness in this model correlate well with the proportions of participants selecting the voiceless allomorph of the past tense suffix, *-te* ($r = 0.85, t(190) = 22.1, p < 0.0001$). That the maximum likelihood choices of the model are similar to those made by the participant is illustrated in Figure **??**, the dendrogram for the hierarchical clustering of the participants and the model. (The clustering is based

on a distance matrix of by-participant pairwise proportions of pseudoverbs which differ with respect to the assignment of voice.) The participants are labelled 1–28, the model is labelled 29, and can be found in the lower center of the tree diagram. In a dendrogram such as this, participants who are very similar will be in a very similar position in the tree. For instance, participants 20 and 25 are very similar, the group formed by participants 20 and 25 is in turn very similar to participant 12. These participants are very different from participants 4 and 6. One has to traverse the tree almost to its root to go from participant 25 to participant 4. In other words, vertical traversal distance, labelled Height in Figure **??**, tells us how dissimilar two participants are. The position of the model (29) near to participants 15, 5, and 27, shows that the model's behavior is quite similar to that of a number of actual participants. If the model had occupied a separate position in the dendrogram, this would have been an indication that its voicing predictions might be in some way fundamentally different from those of the participants, which would have been a source of worry about the validity of the model as a model of how speakers of Dutch arrive at their choice of the past tense suffix.

PLACE FIGURE **??** APPROXIMATELY HERE

Summing up, the distribution of voice alternation for final obstruents in the lexicon of Dutch is far from random. Speakers of Dutch make use of this information when confronted with novel (pseudo)verbs (see Ernestus & Baayen, 2001b, 2002, for evidence that even the voice specification of existing words is likewise affected by the distributional properties of voicing in the lexicon). The probability that a speaker of Dutch will select the voiced or voiceless allomorph of the Dutch past tense suffix can be approximated with a reasonable degree of accuracy on the basis of only three parameters, one for each relevant feature.

There are many other formal quantitative models that provide good fits to the present data (see Ernestus & Baayen, 2001a, for detailed discussion), two of which are of special interest. Boersma (1998) proposes a stochastic version

of optimality theory (SOT) in which constraints are assigned a position on a hierarchy scale. The exact position of a constraint on this scale is stochastic, i.e., it varies slightly from instance to instance, according to a normal distribution. The positions of the constraints are determined by Boersma's gradual learning algorithm (see also Boersma & Hayes, 2001). This algorithm goes through the list of forms in the model's input, adjusting the constraints at each step. If the model predicts an outcome that is at odds with the actually attested outcome, the positions of the constraints are adjusted. Those constraints that are violated by the actual input form are moved down. At the same time, those constraints that are violated by the words that the model thought were correct instead of the actual input form are moved up in the hierarchy. The simplest model that provides a reasonable fit to the data is summarized in Table 2. It has 10 constraints, and hence 10 parameters. The maximum likelihood predictions of this model correspond for 87% of the experimental pseudoverbs with the majority choice of the participants, a success rate that does not differ significantly from the success rate (91%) of the spreading activation model ($p > .25$, proportions test).

Given that SOT and the spreading activation model have the same observational adequacy, the question arises which model is to be preferred for this specific data set. SOT implements a greedy learning strategy, in that the individual examples to which the model is exposed are discarded. It is a memory-less system that presupposes that it is known beforehand which constraints might be relevant. The spreading activation model, by contrast, crucially depends on having in memory the phonological representations of Dutch monomorphemic lexemes. Interestingly, this requirement does not add to the complexity of the grammar, as the phonological form of monomorphemic words must be stored in the lexicon anyway. Since there is no intrinsic advantage to greedy learning for these data, Occam's razor applies in favor of the spreading activation model as the more parsimonious theory, at least for this data set.

PLACE TABLE 2 APPROXIMATELY HERE

The interactive activation model is in its turn challenged by a lazy learning model with no parameters at all, Analogical Modeling of Language (AML, Skousen 1989, 1993). In what follows, I give a procedural introduction to AML. For a rigorous mathematical analysis of the statistical properties of AML, the reader is referred to Skousen (1992), and for the use of this natural statistic in quantum physics, to Skousen (2000).

Like the interactive activation model, AML requires an input lexicon that specifies for each lexeme the values of a series of features describing its final syllable together with the underlying voice specification of the final obstruent. Table 3 lists some examples of an instance base with features Onset, Vowel Type, Vowel (using the DISC computer phonetic alphabet), Coda structure (whether a pre-final consonant is present and, if so, whether it is a sonorant or a stop), and final obstruent.

PLACE TABLE 3 APPROXIMATELY HERE

When AML has to predict the voice specification for a pseudoverb such as *puig*, it considers the exemplars in its lexicon for all possible supracontexts of the target. A supracontext of the target is the set of exemplars (possibly empty) that share a mimimum number (possibly even zero) of feature values with the target. Table 4 lists all the supracontexts of *puig*. The first supracontext has distance 0: The values of all its features are fixed. That is, in order for a word to belong to this supracontext, it must share *puig*'s values for all its five features. Because *puig* is not a real word of Dutch, this fully specified supracontext is empty. The next 5 supracontexts have distance 1. They contain the exemplars that share 4 feature values, and that differ from the target at (at most) one position. This position is indicated by a hyphen in Table 4. The next 10 supracontexts cover the sets of exemplars that have two variable positions. The final supracontext has five variable positions. As we move down Table 4, the supracontexts become less specific in their similarity requirements, and contain nondecreasing numbers of exemplars. The columns labeled 'voiced' and 'voiceless' tabulate the number of exemplars in a given context that have the

corresponding voice specification. Thus, there are 8 exemplars in the second supracontext, 7 of which are underlyingly voiced. The last, most general supracontext at the bottom of the table covers all 1684 words in the lexicon, of which 583 are voiced.

PLACE TABLE 4 APPROXIMATELY HERE

Supracontexts can be *deterministic* or *non-deterministic*. A supracontext is deterministic when all its exemplars support the same voice specification (e.g., `p long L - -`), otherwise, it is non-deterministic (e.g., `- long L None -`). When predicting the voice specification for a new word that is not yet in the model's instance base, AML inspects only those supracontexts that are *homogeneous*. All deterministic supracontexts are homogeneous. A non-deterministic supracontext is homogeneous only when all more specific supracontexts that it contains have exactly the same distribution for the voice specification. Consider, e.g., the non-deterministic supracontext `- long L - x`, which has the outcome distribution $(7, 1)$. It contains the more specific supracontext `- long L None x`. This supracontext is more specific because the fourth feature has the specific value `None`. This supracontext is also non-deterministic and it has the same outcome distribution $(7, 1)$. The supracontext `- long L - x` has one other more specific supracontext, `p long L - x`. This supracontext is the empty set, and does not count against the homogeneity of the more general supracontexts of which it is a subset. Therefore, the supracontext `- long L - x` is homogeneous. It is easy to see that the non-deterministic supracontext `- long L None -` is heterogeneous, as there is no other more specific supracontext with the distribution $(38, 34)$. The homogeneous contexts jointly constitute the analogical set on which AML bases its prediction. Intuitively, one can conceptualize the homogeneity of a supracontext as indicating that there is no more specific information (in the form of a more fully specified supracontext) with contradicting distributional evidence. In other words, distributional evidence tied to more specific supracontexts blocks contradicting distributional evidence from less specific supracontexts from having an analogical

contribution.

PLACE TABLE 5 APPROXIMATELY HERE

Table 5 lists the exemplars that appear in the analogical set. AML offers two ways for calculating the probabilities of the outcomes (voiced or voiceless), depending on whether the size of the supracontexts is taken into account. In occurrence-weighted selection, the contribution of an exemplar (its similarity score in the model) is proportional to the count of different supracontexts in which it occurs. The third column of Table 5 lists these counts, and the fourth column the proportional contributions. The probability of a voiced realization using occurrence weighted selection is 0.75, as the summed count for the voiced exemplars equals 25 out of a total score of 100. Applied to all experimental pseudoverbs, the maximum likelihood choice of AML with occurrence weighted selection agrees with the majority choice of the participants in 166 out of 192 cases, an accuracy score of 86% that does not differ significantly from the accuracy score of 91% obtained with the spreading activation model ($X^2(1) = 1.6761, p = 0.1954$).

When we use size weighted selection, the contribution of an exemplar is proportional to the sum of the sizes of the supracontexts in the analogical set in which it appears. This size weighted selection amounts to using a squaring function for measuring agreement, similar to the quadratic measure of agreement found in Schrödinger's wave equation (Skousen, 2000). The exemplar *buig*, for instance, occurs in 4 homogeneous supracontexts, the homogeneous supracontexts in Table 4 with the $(7, 1)$ distribution. The size of each of these 4 supracontexts is 8, hence, *buig* now contributes a count of 32 instead of 4. In the case of the exemplar *poog*, the two homogeneous supracontexts in which it appears both have a size of 1. Hence, the contribution of *poog* remains proportional to a count of 2. The probability of a voiced realization using size-weighted selection is 0.84. The accuracy score of AML with respect to the complete set of experimental pseudoverbs is again 86%. Although AML seems to be slightly less accurate than the spreading activation model, the fact that AML

is a parameter-free model, i.e., a model with no parameters that the analyst can twiddle to get the model to better fit the data, makes it a very attractive alternative.

It is important to realize that AML bases its predictions on the local similarity structure in the lexicon given a target input. There are no global computations establishing general weights that can subsequently be applied to any new input. It is not necessary to survey the instance base and calculate the information gain weight for, say, the onset. Likewise, it is not necessary to establish a priori whether constraints pertaining to the onset should or should not be brought into a stochastic optimality grammar. (The only requirement is the a-priori specification of a set of features and their values, but this minimal requirement is a prerequisite for any current theory.) What I find interesting is that the microstructure of local similarities as captured by the analogical set of AML is by itself sufficient to capture the support in the language for the voice specification for a given target word. The absence of a role for the onset follows without further specification from the fact that the supracontexts containing the onset (the supracontexts with an initial p in Table 4) are either very sparsely populated or heterogeneous. Althought generalizations in the form of abstract rules may provide a good first approximation of morphological regularities, for more precise prediction it is both necessary and, surprisingly, sufficient to take into account the microstructure of the similarity space around individual words. Global similarity structure is grounded in the local similarity structure around individual words.

3.3 Discussion

The case studies surveyed in this section have a number of interesting consequences for linguistic theory. A first observation concerns the notion of productivity. Regularity, of the kind that can be captured by symbolic rules, is often seen as a necessary condition for productivity. The Dutch linking elements, however, are productive without being regular in this sense. Similarly,

the morphophonological voicing alternation of obstruents in Dutch also enjoys a fair degree of productivity, even though standard analyses have treated it as idiosyncratic and lexically specified. To understand the basis of productivity, the paradigmatic, probabilistic dimension of morphological structure is crucial.

A second observation is that rejecting syntagmatic symbolic rules as the appropriate framework for the analysis of a given morphological phenomenon does not imply embracing subsymbolic connectionism. The present examples show that a symbolic approach in which paradigmatic structure provides a similarity space over which probabilities are defined can provide an excellent level of granularity for understanding the role of probability in language production. This is not to say that the present data sets cannot be modelled by means of subsymbolic artificial neural networks. To the contrary, artificial neural networks are powerful non-linear classifiers, whereas the classification problems discussed in this section are trivial compared to the classification problems that arise in, for instance, face recognition. Artificial neural networks have as disadvantage that they require large numbers of parameters (the weights on the connections) that themselves reveal little about the linguistic structure of the data, unlike the information gain weights in the spreading activation model. The hidden layers in an artificial neural network often provide a compressed re-representation of the structure of the data, but the cost in terms of the number of parameters and the complexity of the training procedure are high. And it is not always clear what one has learned when a three-layer network successfully maps one type of representation onto another (see Forster, 1994). For those who take the task of morphological theory as part of linguistics to be to provide the simplest possible account for (probabilistic) phenomena in word formation, artificial neural networks are probably not the analytically most insightful tool to use. However, those who view morphology as part of cognitive science may gladly pay the price of greater analytical complexity, especially when more biologically realistic neural networks models become available.

A third observation concerns the notion of learning. Traditional symbolic approaches such as the one advocated by Pinker (1991) are based on the a-priori assumption that greedy learning is at the core of the language faculty. The gradual learning algorithm of Boersma (1998) is in line with this tradition: Occurrences leave a trace in the positions of the constraints, they themselves need not be stored in memory. TIMBL and AML, by contrast, are based on lazy learning, with extensive storage of exemplars in memory and similarity-based reasoning taking the place of abstract rules.

A question that arises here is to what extent these models provide a reasonable window on language acquisition. It should be noted at the outset that all models discussed here share, in their simplest form, the assumption that it is known at the outset which features are relevant and what values these features can assume, and that this knowledge is constant and not subject to development over time. This is a strong and unrealistic assumption. Granted this assumption, SOT, TIMBL, and AML can all approximate acquisition as a function of the input over time. Given which constraints are relevant for a given linguistic mapping, SOT can chart how the positioning of constraints develops over time. What TIMBL requires for modeling classificatory development is a continous re-evaluation of the information gain weights as the instance base is increased with new exemplars. AML, by contrast, predicts changing classificatory behavior as a function of a changing lexicon without further assumptions, a property it shares with connectionist models of language acquisition.

A related final question concerns whether there are differences in the extent to which different models depend on a-priori assumptions. All models reviewed here, including SOT and AML, do not differ with respect to the minimal levels of representation they require. The differences between these models concern how they make use of these representations and what happens with the examples from which a given mapping has to be learned. Both TIMBL and AML instantiate lazy learning algorithms that do not require any further a-priori knowledge. The same holds for connectionist models. SOT instantiates a greedy learning algorithm, an algorithm that does require the a-priori

knowledge of which constraints are potentially relevant, or, at the very least, that does require considerable hand-crafting in practice. For instance, there is no constraint *iuyF[+voice] in the SOT model summarized in Table 2, simply because it turned out to be unnecessary given the other constraints that had already been formulated. Theoretically, it might be argued that every combination of feature values (e.g., T, or SonO) and outcome values (e.g., [-voice]) is linked automatically with a (supposedly innate, universal) constraint, with irrelevant constraints dropping to the bottom of the grammar during learning. Under this view, SOT would not depend on a-priori knowledge either, although such a SOT grammar is encumbered with an enormous pile of useless constraints lying inactive at the bottom of the ranking. Note, however, that TIMBL and AML are encumbered in a different way, namely, with useless features that are themselves harmless but that render the on-line calculation of the similarity space more complex. Similarly in connectionist networks, such useless features add noise to the system, delaying learning and slowing down convergence.

Summing up, from a statistical point of view, SOT, AML and TIMBL, and connectionist models as well, all have their own advantages and disadvantages as explanatory frameworks for the data sets discussed. One's choice of model will in practice be determined by one's view of the explanatory value of these models as instantiations of broader research traditions (optimality theory, machine learning, and cognitive science) across a much wider range of data sets.

4 Probability in morphological comprehension

In the previous section, we have seen how probabilities based on counts of word *types* falling into different similarity classes play a role in the production of morphologically complex words. In this section, we shall see that probabilities based on *token* frequencies play a crucial role in solving the ambiguity problem in morphological segmentation. Consider the examples in (3).

(3)				
	acute+ness	32	a+cuteness	39
	expert+ness	25	ex+pert+ness	68
	intent+ness	29	in+tent+ness	57
	perverse+ness	31	per+verse+ness	66
	sacred+ness	27	sac+redness	56
	tender+ness	28	tend+er+ness	55
	prepared+ness	19	prep+a+red+ness	58

The first column lists correct segmentations for a number of words with the suffix *-ness*, the third column lists some incorrect or implausible segmentations. I will explain the interpretation of the numbers in columns two and four below. How do we know, upon reading a string such as `preparedness`, which of the segmentations in (4) is the one to choose?

(4)		
	preparedness	19
	prepared+ness	27
	pre+pared+ness	56
	prep+a+red+ness	58
	prep+a+redness	58
	pre+par+ed+ness	71
	pre+pa+red+ness	78
	pre+pa+redness	78
	pre+pare+d+ness	78
	prep+are+d+ness	78
	prepare+d+ness	78

Some of these segmentations can be ruled out on the basis of combinatorial restrictions. For instance, the verb form *are* (past tense of *be*, or singular of the noun denoting an area of 100 m^2) in *prep+are+d+ness* does not combine with the suffix *-d*. Other segmentations in (4), however, are possible albeit implausible. E.g., *((pre((pare)d))ness)* has the same structure as *((pre((determine)d))ness)*, but it is unlikely to be the intended reading of *preparedness*. Why are such forms implausible? In other words, why are their probabilities so low? In various com-

putational approaches (e.g., probabilistic context-free grammars, probabilistic head lexicalized grammars, and data-oriented parsing, as described in the introductory chapter of this book), the probability of the whole, P(preparedness), is obtained from the probabilities of combinations of its constituents, some of which are listed in (5).

(5) P(prepared, ness)
P(pre, pared)
P(pared, ness)
P(pre, pared, ness)
...
P(prepare, d)
P(d, ness)
P(prepare, d, ness)

Crucially, these probabilities are estimated on the basis of the relative token frequencies of these bigrams and trigrams in large corpora, with the various approaches using different subsets of probabilities and combining them according to different grammars (for an application to Dutch morphology, see Heemskerk, 1993; for a memory-based segmentation system using TIMBL see Van den Bosch & Daelemans, 2000). In this section, I consider the question how human morphological segmentation might be sensitive to probabilities of combinations of constituents.

It is important to realize that if the brain does indeed make use of probabilities, then it must somehow keep track of (relative) frequency information for both irregular and completely regular complex words. The issue of whether fully regular complex words are stored in the mental lexicon, however, is hotly debated. Pinker (1991, 1997), Marcus, Brinkman, Clahsen, Wiese, and Pinker (1995), Clahsen, Eisenbeiss, and Sonnenstuhl (1997), and Clahsen (2000) have argued that frequency information is stored in the brain only for irregular complex words, and not at all for regular complex words. They argue that regular morphology is subserved by symbolic rules that are not sensitive to the fre-

quencies of the symbols on which they operate, and that irregular morphology is subserved by a frequency-sensitive associative storage mechanism.

The only way in which probabilities might play a role for regular complex words in this dual route model is at the level of the rules themselves, i.e., rules might differ with respect to their probability of being applied. Consider Table 6, which lists the frequencies of the singular, plural, and diminutive forms of the Dutch nouns *tong*, 'tongue', and *gast*, 'guest', as listed in the CELEX lexical database. By assigning different probabilities to the rules for diminutivization and pluralization, this approach can account for the lower probabilities of diminutives compared to plurals. However, it cannot account for the differences in the probabilities of the two plural forms. Even though the lexemes *tong* and *gast* have very similar probabilities, the former occurs predominantly in the singular, and the latter predominantly in the plural. I will refer to *tong* as being a singular-dominant noun and to *gast* as a plural-dominant noun. What the dual route model predicts is that such differences in frequency dominance are not registered by the brain, and hence that such differences do not affect lexical processing. Note that this amounts to the claim that the brain has no knowledge of the probability that a particular noun co-occurs with the plural suffix. The only frequency count that should be relevant in the dual route approach is the stem frequency, the summed frequency of the singular and the plural forms.

PLACE TABLE 6 APPROXIMATELY HERE

There is massive evidence, however, that the claim that frequency information is retained by the brain for irregular words only is incorrect (see, e.g., Taft, 1979, Sereno & Jongman, 1995, Bertram, Laine, Baayen, Schreuder, & Hyönä, 1999, Bertram, Schreuder, & Baayen, 2000, Baayen, Schreuder, De Jong, & Krott, in press). These studies show that probabilistic information in the form of knowledge of co-occurrence frequencies of constituents forming complex words must be available to the brain. But how then might the brain make use of this information? Although I think that formal probabilistic models

provide excellent mathematically tractable characterizations of what kinds of knowledge are involved in morphological segmentation, I do not believe that, e.g., the algorithms for estimating the parameters of hidden markov models can be mapped straightforwardly onto the mechanisms used by the brain. The way the brain solves the ambiguity problem may well be more similar to a dynamic system in which a great many interdependent morphological units compete to provide a segmentation that spans the target word in the input. In what follows, I will outline a model implementing a simple dynamic system, and I will show that it provides a reasonable framework for understanding some interesting aspects of the processing of Dutch and German plural nouns.

4.1 A dynamic system for morphological segmentation

Whereas computational parsing models in linguistics have succesfully used token-count based probabilities of occurrence, psycholinguistic research on the segmentation problem has focused on the role of form similarity. In the Shortlist model of the segmentation of the auditory speech stream (Norris, 1994; Norris, McQueen, & Cutler, 1996), for instance, the lexical representations that are most similar to the target input are wired into a connectionist network implementing a similarity-based competition process. The resulting model is very sensitive to differences in form between lexical competitors, and captures important aspects of auditory lexical processing. However, the authors of Shortlist have not systematically addressed how to account for the word frequency effect in auditory word recognition (see, e.g., Rubenstein & Pollack, 1963).

MATCHECK (Baayen, Schreuder, & Sproat, 1997, Baayen & Schreuder, 1998, 2000) implements an approach to the segmentation problem in which form similarity and token-frequency based probability simultaneously play a role. This model is a dynamic system articulated within the interactive activation framework. It shares with dynamic systems in general the properties that its behavior depends crucially on the initial condition of the model, that it is de-

terministic, and that there is order in what seems to be chaotic behavior. The components of the model are an input lexicon, a mechanism for ascertaining whether a lexical representation should be taken into account as a candidate for a segmentation, and a competition mechanism. In what follows, I outline the architecture of Matcheck for the visual modality.

The input lexicon contains form representations for stems, affixes, and full forms, irrespective of their regularity. Each representation w has an initial activation level $a(w, 0)$ equal to its frequency in a corpus. The initial probability of a lexical representation $p_{w,0}$ is its relative frequency in the lexicon.

The mechanism for determining whether a lexical representation should be taken into account as a possible constituent in a segmentation makes use of an activation probability threshold $0 < \theta < 1$. Only those lexical representations with a probability $p_{w,t} \geq \theta$ at timestep t are candidates for inclusion in a segmentation.

The competition mechanism consists of a probability measure imposed on the activation levels of the lexical representations, combined with a similarity-based function that determines whether the activation of a given lexical representation should increase or decrease. The activation probability of representation w at timestep t is

$$p_{w,t} = \frac{a(w,t)}{\sum_{i=1}^{V} a(w_i,t)}, \tag{11}$$

with V the number of representations in the lexicon. The details of the criteria for whether the activation of a lexical representation increases or decreases need not concern us here. What is crucial is that a lexical representation that is aligned with one of the boundaries of the target word is allowed to increase its activation until its activation probability has reached the threshold θ. Once this threshold has been reached, activation decreases. Given a decay rate δ_w ($0 < \delta_w < 1$) for representation w, the change in activation from one time step to the next is defined as

$$a(w,t) = a(w,0) + \delta_w\{a(w,t-1) - a(w,0)\}. \tag{12}$$

Because $\delta_w < 1$, the activation at each successive timestep becomes smaller

than it was at the preceding timestep. Asymptotically, it will decrease to its original resting activation level. Activation increase, which occurs at the timesteps before w has reached threshold, is also defined in terms of δ_w,

$$a(w,t) = \frac{a(w,t-1)}{\delta_w}, \tag{13}$$

but now we divide by δ_w instead of multiplying by δ_w. Consequently, the activation at timestep t becomes greater than the activation at timestep $t-1$. Because activation increase and activation decrease are both defined in terms of δ_w, words with a decay rate close to 1 are slow to decay and slow to become activated. Conversely, words which decay quickly (δ_t close to 0) also become activated quickly. Note that if the activation level of a lexical representation increases while the activation levels of the other representations remain more or less unchanged, its probability increases as well.

The key to the accuracy of MATCHECK as a segmentation model lies in the definition of the activation and decay parameter δ_w, which differs from representation to representation. It is defined in terms the frequency f_w and the length L_w of a lexical representation w, in combination with three general parameters (α, δ, ζ), as follows:

$$\delta_w = f(g(\delta, \alpha, w), \zeta, \delta, w), \tag{14}$$

with

$$g(\delta, \alpha, w) = \delta \frac{1}{1 + \alpha \frac{\log(L_w+1)}{\log(f_w)}}, \tag{15}$$

and, with T denoting the target word, and using d as shorthand for $g(\delta, \alpha, w)$,

$$f(d, \zeta, \delta, w) = \begin{cases} d + (1-d)\left(\frac{|L_w - L_T|}{\max(L_w, L_T)}\right)^{\zeta} & \text{iff } \zeta > 0 \\ \delta & \text{otherwise.} \end{cases} \tag{16}$$

The parameter δ ($0 < \delta < 1$) denotes a basic activation and decay rate that is adjustable for each individual word. Lexical representations with higher frequencies receive higher values for δ_w. They have a reduced information load, and become activated less quickly than lower-frequency representations. The

parameter α ($\alpha \geq 0$) specifies how the frequency and length of the representation should be weighted, independently of its similarity to the target word. The parameter ζ ($\zeta \geq 0$) determines to what extent the relative difference in length of a lexical competitor and the target word should affect the activation and decay rate of w. Increasing α or ζ leads to a decrease of δ_w, and hence to more rapid activation and decay. Finally, constituents that are more similar in length to the target word have a smaller δ_w than shorter constituents (for experimental evidence that longer affixes are recognized faster than shorter affixes, see Laudanna, Burani and Cermele, 1995).

Figure **??** illustrates the activation dynamics of MATCHECK. On the horizontal axis, it plots the timesteps in the model. On the vertical axis, it plots the activation probability. The solid horizontal line represents the activation probability threshold $\theta = 0.3$. Consider the left panel, which plots the activation curves for the singular-dominant plural *tongen*. The curved solid line represents the plural suffix *-en*, which, due to its high frequency, has a high initial probability. During the first timesteps, the many words with partial similarity to the target, such as *long*, 'lung', become activated along with the constituents themselves. Hence, although the activation of *-en* is actually increasing, its activation probability decreases. The base *tong* starts out with a very low initial probability, but due to its greater similarity to the plural form, it reaches threshold long before *-en*. Upon having reached threshold, its activation begins to decay. The subsequent erratic decay pattern for *tong* is due to the interference of a lexical competitor, not shown in Figure **??**, the orthographic neighbor of the plural form, *tonnen* (see, e.g., Andrews, 1992, Grainger, 1990, and Grainger & Jacobs, 1996, for experimental evidence for orthographic neighborhood effects). Thanks to the probability measure imposed on the activations of the lexical representations, the decay of the activation of *tong* makes activation probability available for the lower-frequency plural form, which now can now reach threshold as well. This point in model time (timestep 19) is represented by a vertical solid line. This is the first timestep in the model at which a full spanning of the input is available. In other words, the first 'parse' to become available for *tongen* is

the plural form. Once both the singular and plural form have entered activation decay, the plural affix finally reaches threshold. It is only now that stem and suffix provide a full spanning of the input, so the second parse to become available arrives at timestep 41. The following erratic activation bumps for the base noun arises due to the difference in the speed with which the singular and plural forms decay, and have no theoretical significance.

The central panel in Figure **??** reveals a similar pattern for the plural-dominant plural *gasten*. However, its representations reach the threshold at an earlier timestep. The singular *tong* reaches theshold at timestep 13 and its plural *tongen* becomes available at timestep 18, whereas the corresponding model times for *gast* and *gasten* are 10 and 14 respectively.

The right panel of Figure **??**, finally, shows the timecourse development for a very low-frequency singular-dominant noun plural, *loepen*, 'magnification glasses'. Note that the first constituent to reach threshold is the plural suffix, followed by the base, which jointly provide a full spanning of the target long before the plural form reaches threshold.

Two questions arise at this point. First, how well does MATCHECK solve the ambiguity problem? Second, how good is MATCHECK at modeling actual processing times in psycholinguistic experiments? The first question is addressed in Baayen and Schreuder (2000). They showed, for instance, that for 200 randomly selected words of lengths 5–8 with on average 3 incorrect segmentations, the first segmentation to be produced by MATCHECK was correct in 194 cases, of which 92 were due to the full form being the first to become available, and 102 to a correct segmentation becoming available first. The model times listed in examples (3) and (4) illustrate how MATCHECK tends to ranks correct segmentations before incorrect segmentations, using the parameter settings of Baayen & Schreuder (2000). For instance, the full form *preparedness* and the correct parse *prepared-ness* become available at timesteps 19 and 27 respectively, long before the first incorrect parse *pre-pared-ness* (timestep 56) or the complete correct segmentation *prepare-d-ness* (timestep 78). Because the model gives priority to longer constituents, a parse such as *prepared-ness*, which is based on the

regular participle *prepared*, becomes available before the many incorrect or implausible parses containing shorter constituents. The presence of the regular participle *prepared* in the lexicon protects the model against having to decide between alternative segmentations such as *prepare-d-ness* and *pre-pared-ness*. In other words, paradoxically, storage enhances parsing. In MATCHECK, storage in memory does not just occur for its own sake, its functionality is to reduce the ambiguity problem.

PLACE FIGURE **??** APPROXIMATELY HERE

We are left with the second question, namely, whether the model times produced by MATCHECK have any bearing on actual human processing latencies. The next two sections address this issue by means of data sets from Dutch and German.

4.2 Regular noun plurals in Dutch

Baayen, Dijkstra, and Schreuder (1997) studied regular Dutch plurals in *-en* and their corresponding singulars using visual lexical decision. The visual lexical decision task requires participants to decide as quickly and accurately as possible whether a word presented on a computer screen is a real word of the language. Response latencies in visual lexical decision generally reveal strong correlations with the frequencies of occurrence of the words. This particular study made use of a factorial experimental design contrasting three factors: Stem Frequency (high versus low), Number (singular versus plural), and Dominance (singular dominant versus plural dominant). Within a stem frequency class, the average stem frequency was held constant in the mean, as illustrated in Table 6 above for the nouns *tong* and *gast*. The right panel of Figure **??** provides a graphical summary of the pattern of results. The horizontal axis contrasts singulars (left) with plurals (right). The dashed lines represent plural dominant singulars and plurals. The solid lines represent singular dominant singulars and plurals. The lower two lines belong to the nouns with a

high stem frequency, and the upper two lines to the nouns with a low stem frequency.

What this graph shows is that high frequency nouns, irrespective of number, are processed faster than low-frequency nouns. It also reveals a frequency effect for the plural forms. For each of the two stem frequency conditions, singular-dominant plurals are responded to more slowly than plural-dominant plurals. Speakers of Dutch are clearly sensitive to how often the plural suffix *-en* co-occurs with particular noun singulars to form a plural. Especially the fast response latencies for plural dominant plurals bear witness to the markedness reversal studied by Tiersma (1982): While normally the singular is the unmarked form both with respect to its phonological form as with respect to its meaning, many plural dominant plurals are semantically unmarked compared to their corresponding singulars (see Baayen, Dijkstra, and Schreuder, 1997, for further discussion). This plural frequency effect, however, is completely at odds with the dual route model advanced by Pinker and Clahsen and their coworkers. Finally, Figure **??** shows that for noun singulars it is the frequency of the lexeme, i.e., the summed frequencies of the singular and the plural forms, that predicts response latencies, and not so much the frequencies of the singular forms themselves. If the frequencies of the singulars as such would have predicted the response latencies, one would have expected to see a difference in the response latencies between singular dominant and plural dominant singulars. The observed pattern of results has been replicated for Dutch in the auditory modality (Baayen, McQueen, Dijkstra, and Schreuder, 2002), and for Italian (visual modality) by Baayen, Burani, and Schreuder (1997).

PLACE FIGURE **??** APPROXIMATELY HERE

The left panel of Figure **??** summarizes the results obtained with MATCHECK for the parameter settings $\delta = 0.3, \theta = 0.25, \alpha = 0.4, \zeta = 0.3$, and $\rho = 3$. This last parameter determines the granularity of the model, i.e., the precision with which the timestep at which the threshold is reached is ascertained (see Baayen, Schreuder, & Sproat, 2000). The lexicon of the model contained the

186 experimental words and in addition some 1650 randomly selected nouns as well as various inflectional and derivational affixes. The frequencies of the affixes were set to the summed frequencies of all the words in the CELEX lexical database in which they occur. Noun singulars received the summed frequencies of the singular and plural forms as frequency count, the assumption being that in a parallel dual route model the processing of the (globally marked) plural leaves a memory trace on the (globally unmarked) singular. Noun plurals were assigned their own plural frequency. The lexicon also contained a dummy lexeme consisting of a series of x characters, which received as frequency count the summed frequencies of all words in the CELEX database that were not among the 1650 randomly selected nouns. This ensured that all words in the lexicon had initial probabilities identical to their relative frequencies in the corpus on which the CELEX counts are based.

Interestingly, it is impossible to obtain the pattern shown in the left panel of Figure **??** with just these settings. The reason for this is that the singular form, thanks to its cumulated frequency, is a strong competitor of the plural form. Although plural-dominant plurals reach the threshold before singular-dominant plurals with the same stem frequency, as desired, they also reach the threshold well after the corresponding singulars. Indistinguishable processing times for singulars and plurals, as observed for plural-dominant singulars and plurals in the high stem frequency condition in the experiment, cannot be simulated.

The adaptation of MATCHECK that leads to the pattern of results actually shown in the left panel of Figure **??** is to enrich the model with a layer of lexemes in the sense of Aronoff (1994) or lemmas in the sense of Levelt (1989). The representations of the singular and plural forms have pointers to their lexeme, which in turn provides pointers to its associated semantic and syntactic representations. The lexemes serve a dual function in MATCHECK. Their first function is to accumulate in their own activation levels the summed activation levels of their inflectional variants. Once the lexeme has reached threshold activation level, a response can be initiated. This allows the model to take into account the combined evidence in the system supporting a given lexeme.

The second function of the lexemes is to pass on part of the activation of the (marked) plural form to the (unmarked) singular form. I assume that it is this process that leaves a memory trace with the singular that results over time in the summed frequency of the singular and the plural form being the predictor of response latencies for singulars. To implement this modification, we have to revise the definition of increasing activation given in equation (13) for the representation of the singular as follows. Let w_s denote the representation of a singular, and let i range over its n inflectional variants. For the present data, $n = 1$, as there is only one inflectional variant other than the singular itself, namely, the plural. We define $a(w_s, t)$ as follows:

$$a(w_s, t) = \frac{a(w_s, t-1)}{\delta_{w_s}} + \log(a(w_s, t-1)) \sum_{i=1}^{n} a(w_i, t)^{\lambda}, \tag{17}$$

with λ the parameter determining how much activation flows from the plural to the singular. In the simulation leading to the left panel of Figure **??**, λ was set to 1.1.

How well does this model approximate the observed patterns in the experimental data? A comparison of the left and right panels of Figure **??** suggests that the model provides a good fit, an impression that is confirmed by Table 7: The same main effects and interactions that are significant in the experiment are also significant according to the model. The model also captures that high-frequency plural dominant singulars and plurals are processed equally fast ($t(45.99) = -0.82, p = 0.416$ for the response latencies, $t(35.44) = -0.86, p = 0.398$ for the model times, Welch two-sample t-tests), in contrast to the low-frequency plural dominant singulars and plurals ($t(41.65) = -2.38, p = 0.022$ for the response latencies, $t(32.33) = -4.394, p = 0.0001$ for the model times).

PLACE TABLE 7 APPROXIMATELY HERE

MATCHECK provides this good fit to the experimental data with parameter settings that make it make the most of the available full form representations. The only word for which it is a parse into base and affix that reached

the threshold before the full form was the plural form *loepen*, for which the activation probabilities over time were shown in the right panel of Figure **??**. In other words, if the model has to parse, it can do so without any problem; otherwise, the stem will tend to contribute to the recognition process only through its contribution to the activation of the lemma.

The conclusion that parsing as such plays a minor role for this data set also emerged from the mathematical modeling study of Baayen, Dijkstra, and Schreuder (1997). Their model, however, was based on the assumption that there is no interaction between the representations feeding the direct access route and those feeding the parsing route. The present model, by contrast, implements the idea that all lexical representations are in competition and that the direct route and the parsing route are not independent. In fact, by allowing activation to spread from the plural to the base, and by allowing the evidence for the word status of a target in lexical decision to accumulate at the lexeme layer, there is much more synergy in this competition model that in the earlier mathematical model.

4.3 Noun plurals in German

Clahsen, Eissenbeiss, and Sonnenstuhl (1997) and Sonnenstuhl & Huth (2001) report that in German high-frequency plurals in *-er* are responded to faster in visual lexical decision than low-frequency plurals in *-er*, while matched high and low-frequency plurals in *-s* are responded to equally fast. They attribute this difference to the linguistic status of these two plurals. The *-s* plural is argued to be the default suffix of German (Marcus et al, 1995), the only truly regular plural formative. According to these authors, plurals in *-s* are therefore not stored, which would explain why high and low frequency plurals in *-s* require the same processing time, a processing time that is determined only by the speed of the parsing route and the resting activation levels of the representations on which it operates. Plurals in *-er*, by contrast, are described as irregular. These forms must be stored, and the observed frequency effects for

high and low frequency plurals reflect this.

In the light of their theory, it is not suprising that Clahsen, Eissenbeiss, and Sonnenstuhl (1997) suggest that Dutch plurals in *-en* might not be regular but rather irregular (Clahsen, Eissenbeiss, and Sonnenstuhl, 1997). This hypothesis, which does not make sense from a linguistic point of view (see Baayen, Schreuder, De Jong, & Krott, 2002), is inescapable if one accepts the dual route model. However, the dual route model faces many problems.

Burzio (2001), for instance, points out that many morphologically regular past-tense forms are phonologically irregular, while many morphologically irregular past-tense forms are phonologically regular, an inverse correlation between morphological and phonological regularity. In a dual route model, morphological rules should be able to coexist with phonological rules, leading to a positive correlation between morphological and phonological regularities, contrary to fact.

Behrens (2001) presents a detailed examination of the German plural system that shows that the *-s* plural does not fulfill the criteria for instantiating a symbolic rule, a conclusion that is supported by her acquisition data based on a very large developmental corpus.

Other problems for the dual route model are pointed out by Ramscar (2001) and Hare, Ford, and Marslen-Wilson (2001). Hare et al. (2001) report frequency effects for regular past-tense forms in English that had unrelated homophones (e.g., *allowed* and *aloud*). Particularly interesting is the study by Ramscar, who shows that past-tense inflection in English is driven by semantic and phonological similarity, instead of by encapsulated symbolic rules that would be sensitive to only the phonological properties of the stem.

If the separation of rule and rote in the dual route model is incorrect (see also Bybee, 2001), the question arises why there seems to be no frequency effect for the German *-s* plural. In what follows, I will show that a simulation study with MATCHECK sheds new light on this issue.

The data set that serves as our point of departure is that of Experiment 1 of the study by Sonnenstuhl & Huth (2001). Table 8 summarizes the design of this

experiment: six affix classes (plurals in *-s*, plurals in *-er*, and four sets of plurals in *-n*/*-en*, grouped by gender and the presence or absence of a final schwa), each of which is crossed with plural frequency (high versus low) while matching for Stem frequency. Table 9 summarizes the pattern of results obtained by means of a series of t-tests on the item means, and the right panel of Figure **??** displays the results graphically. The high and low frequency plurals in *-s* elicited response latencies that did not differ significantly in the mean. The same holds for the set of plurals in *-en* of non-feminine schwa-final nouns (labelled (4) in the tables and figure). The authors forward these results as evidence that plurals in *-s* are not stored in the German mental lexicon. A question that remains unanswered in this study is why the supposed default plural suffix, the rule-governed formative par excellence, is the one to be processed most slowly of all plural classes in the study.

PLACE TABLE 8 APPROXIMATELY HERE

PLACE TABLE 9 APPROXIMATELY HERE

To study this data set with MATCHECK, a lexicon was constructed with the 120 singulars and their corresponding plurals that were used in this experiment. Some 2100 randomly selected words from the CELEX lexical database were added to the model's lexicon, including representations for the plural suffixes *-en*, *-n*, *-er*, and *-s* as well as for other inflectional suffixes such as *-d* and *-e*. All lexical entries received a frequency based on the frequency counts in the German section of the CELEX lexical database. Because these counts are based on a small corpus of only 6 million words, they were multiplied by 7 to make them similar in scale to the Dutch frequency counts, which are derived from a corpus of 42 million words. Suffixes received frequencies equal to the summed frequency of all the words in which they occur as a constituent. The singulars of the target plural forms were assigned the summed frequency of their inflectional variants, the lemma frequency of CELEX on which the high and low frequency sets of plurals in the Sonnenstuhl & Huth (2001) study

were matched, just as in the simulation study of Dutch plurals in the preceding section. Again, a dummy lexeme was added with a frequency equal to the summed frequencies of the words not explicitly represented in the model's lexicon. Parameter values that as a first step yield a good fit to the experimental data are $\delta = 0.28, \theta = 0.20, \rho = 3, \alpha = 0.7, \zeta = 0.9$, and $\lambda = 0.5$. This fit is shown in the left panel of Figure **??**. As shown by the Welch two-sample t-tests listed in Table 9, those frequency contrasts that were found to be statistically significant in the experiment are also significant in the model, and those contrasts that are not significant in the experiment are likewise not significant in the simulation.

What is disturbing about these results is that the ordering in model time seems to be wrong. Compare the left and right panels of Figure **??**. According to the model, the plurals in *-s* should be processed more quickly than any other kind of plural, but in fact they elicit the longest response latencies. Conversely, the plurals in *-er* show up in the simulation with the longest model times, even though in fact they were responded to very quickly.

PLACE FIGURE **??** APPROXIMATELY HERE

Why do we get this reversed pattern in MATCHECK? A possible answer to this question can be found by considering the average family size of the six plural classes. The family size of an inflected simplex noun is the type count of the derived words and compounds in which that noun occurs as a constituent. Various studies have shown that, other things being equal, words with a large morphological family are responded to more quickly than words with a small morphological family (Schreuder & Baayen, 1997, Bertram, Baayen, & Schreuder, 1999, De Jong, Schreuder, & Baayen, 2000, De Jong, Feldman, Schreuder, Pastizzo, & Baayen, 2001). The family size effect is semantic in nature, and probably arises due to activation spreading in the mental lexicon to morphologically related words. Interestingly, Table 10 shows that the nouns with the *-er* plural have the highest family size in this data set, while the nouns with the *-s* plural have the lowest family size. Thus, plurals in *-s* might elicit long response la-

tencies because of their small morphological families. Conversely, the plurals in *-er* might be responded to quickly thanks to their large families.

PLACE TABLE 10 APPROXIMATELY HERE

It is possible to test this hypothesis by asking ourselves whether we can systematically reorder the lines in the left panel on the basis of the family counts listed in Table 10 such that a pattern approaching that in the right panel is obtained. It turns out that this is not possible using the counts of family members for the individual words, probably because these counts (based on a corpus of only 6 million words) introduce too much noise compared to the model times of MATCHECK. A mapping is possible, however, on the basis of the class means, using the transformation $h(t_{ij})$ for the j-th plural in the i-th plural class with family size V_i and model time t_{ij}. Let $x_i = \log(V_i) - 1.2$, the distance of log family size from a baseline log family size of 1.2. The further the family size of a plural class is from this baseline, the further the distance y_i that it will shifted:

$$y_i = 1.7 * e^{|x_i|} * s(x_i), \tag{18}$$

with $s(x) = 1$ if $x > 0$ and $s(x) = -1$ if $x < 0$. Adjustment with the mean of the shifts y_i leads to the transformation

$$h(t_{ij}) = t_{ij} - y_i - \frac{\sum_k^n y_k}{n}. \tag{19}$$

Application of (19) results in the pattern shown in the central panel of Figure **??**, a reasonable approximation of the actually observed pattern represented in the rightmost panel. Although a more principled way of integrating MATCHECK with subsequent semantic processes is clearly called for, the present result suggests that differences in morphological family size may indeed be responsible for the difference in ordering between the first and third panels of Figure **??**. This family size effect that we observe here underlines that the plurals in *-er* are tightly integrated in the network of morphological relations of German, and that, by contrast, the plurals in *-s* are rather superficially integrated and relatively marginal in the language. This is not what one would expect for

a default suffix in a dual route model, especially if default status is meant to imply prototypical rule-based processing rather than marginal rule-based processing applicable mainly to recent loans, recent abbreviations, interjections, and other normally uninflected words.

These modeling results receive further support by a simulation of Experiment 4 of the study by Clahsen et al. (1997), who contrasted plurals in *-s* with plurals in *-er*. As in the Sonnenstuhl & Huth (2001) study, a significant frequency effect was observed only for the plurals in *-er*. Interestingly, application of MATCHECK to this data set with exactly the same parameter values leads to the same pattern of results, with a significant difference in modeltime for the plurals in *-er* ($t(15.45) = 4.13, p = 0.0008$) but not for the plurals in *-s* ($t(15.52) = 1.60, p = 0.1290$).

We are left with one question. Why is it that MATCHECK does not produce a frequency effect for plurals in *-s*? Consider Figure **??**, which plots the timecourse of activation for the German nouns *Hypotheken* (left panel) and *Taxis* (right panel). The pattern in the left panel is typical for the plurals in *-en* and for those plurals in *-er* for which no vowel alternation occurs. Crucially, the base and the plural form become active one after the other, indicating that there is little competition between them. In the case of plurals in *-s*, however, the base and the plural become active at more or less the same time — they are in strong competition masking the effect of the frequency of the plural. The reason for this strong competition is the small difference in length between the singular forms and their corresponding plurals in *-s*. Recall that the activation/decay rate of a word as defined in equations (14)–(16) depends on its length in relation to the length of the target word. The more similar a word is in length to the target, the faster it will become active. This property contributes to the segmentation accuracy of MATCHECK as reported in Baayen & Schreuder (2000), and it is responsible for the absence of a frequency effect for *-s* plurals in the present simulation study. Note that this property of MATCHECK is of the kind typically found in dynamic systems in general, in that it allows a tiny difference in the initial conditions (here the length of the base) to lead to quite different

outcomes (here the presence or absence of a frequency effect).

PLACE FIGURE ?? APPROXIMATELY HERE

From a methodological point of view, it is important to realize that null-effects, in this case the absence of a frequency effect for German plurals in *-s*, should be interpreted with caution. Clahsen and his co-workers infer from this null effect that German *-s* plurals do not develop representations of their own. We have seen that this null effect may also arise as a consequence of the differences between the forms of the *-s* and *-en* suffixes and the kinds of words that they attach to (see Laudanna & Burani, 1995, and Bertram, Schreuder, & Baayen, 2000, for the processing consequences of affix-specific differences). Another example of an argument based on a null effect can be found in Frost, Forster, and Deutsch (1997), who claim that connectionist models would not be able to model the presence and absence of semantic transparency effects in priming experiments in English and Hebrew respectively. As shown by Plaut and Gonnerman (2000), the null effect of transparency in Hebrew emerges in connectionist simulation studies as a consequence of the difference in morphological structure between English and Hebrew.

4.4 Discussion

This section addressed the role of probability in morphological comprehension. Techniques developed by computational linguists make profitable use of co-occurrence probabilities to select the most probable segmentation from the set of possible segmentations. It is clear from the computational approach that knowledge of co-occurrence probabilities is indispensible for accurate and sensible parsing.

From this perspective, the hypothesis defended by Pinker and Clahsen and their co-workers that co-occurrence knowledge is restricted to irregular complex words is counterproductive (see also Bybee, 2001). Bloomfieldian economy of description cannot be mapped so simply onto human language processing. Much more productive are, by contrast, those psycholinguistic studies

that have addressed in great detail how form-based similarity affects the segmentation process. These studies, however, seem to implicitly assume that the segmentation problem can be solved without taking co-occurrence probabilities into account.

MATCHECK is a model in which probabilities develop dynamically over time on the basis of both frequency of (co-)occurrence and form similarity. I have shown that this model provides reasonable fits to experimental data sets, and that it provides an explanation of why frequency effects for regular plurals may be absent even though these plurals are stored in the mental lexicon. The mechanisms used by MATCHECK to weight the role of frequency and similarity are crude and in need of considerable refinement. Nevertheless, I think that the model provides a useful heuristic for understanding how the brain might use probability in morphological comprehension to its advantage.

5 Concluding remarks

In this chapter, I have presented some examples of how concepts from probability theory and statistics can be used to come to grips with the graded nature of many morphological data. I have first shown that the graded phenomenon of morphological productivity can be formalized as a probability, a probability that is itself grounded, at least in part, in junctural phonotactic probabilities and parsing probabilities. Following this, I have discussed examples of morphological regularities that are intrinsically probabilistic in nature, outlining how simple spreading activation architectures (symbolic connectionist networks) might capture the role of probability while avoiding complex statistical calculations. Finally, I have shown that part of the functionality of the storage of regular complex forms in the mental lexicon may reside in contributing to resolving parsing ambiguities in comprehension.

Not surprisingly, I agree with researchers such as Seidenberg & Gonnerman (2000) and Plaut & Gonnerman (2000) that traditional, non-probabilistic theories of morphology are inadequate in the sense that they cannot handle graded

phenomena in an insightful way. Seidenberg, Plaut, and Gonnerman seem to suggest, however, that the graded nature of morphology shows that the sub-symbolic connectionist approach to language is the *only* way to go. This is where they and I part company. In this chapter, I have shown that computational models of analogy provide excellent analytical tools for understanding the role of probability in morphology.

Acknowledgements

I am indebted to Kie Zuraw and Janet Pierrehumbert for their careful and constructive criticism, to Wolfgang Dressler, Mirjam Ernestus, Robert Schreuder, and Royal Skousen for their comments and suggestions for improvements, and to Stefanie Jannedy, Jennifer Hay, and Rens Bod for their wonderful editorial work. Errors in content and ommissions remain the responsibility of the author.

Modifier (1)	Head (2)	Nucleus (1)	Onset (2)	Coda (2)	*L*	translation
schaap	*bout*	aa	b	t	*-en-*	'leg of mutton'
schaap	*herder*	aa	h	r	*-0-*	'shepherd'
schaap	*kooi*	aa	k	i	*-s-*	'sheep fold'
schaap	*vlees*	aa	v	s	*-en-*	'mutton'
lam	*bout*	a	b	t	*-s-*	'leg of lamb'
lam	*vlees*	a	v	s	*-s-*	'lamb'
lam	*gehakt*	a	g	t	*-s-*	'minced lamb'
paard	*oog*	aa	-	g	*-en-*	'horse eye'
koe	*oog*	oe	-	g	*-en-*	'cow's eye'
varken	*oog*	e	-	g	*-s-*	'pig's eye'

Table 1: Features and their values for a hypothetical instance base of Dutch compounds. *L* denotes the linking element. The numbers in parentheses refer to the first and second constituents.

Constraint	Gloss	Position
*P[+voice]	no underlyingly voiced bilabial stops	-173.5
*T[+voice]	no underlyingly voiced alveolar stops	-217.5
*S[+voice]	no underlyingly voiced aveolar fricatives	-512.2
*F[−voice]	no underlyingly voiceless labiodental fricatives	-515.4
*X[−voice]	no underlyingly voiceless velar fricatives	-515.6
*V:O[−voice]	no underlyingly voiceless obstruents following long vowels	-516.4
*iuyO[−voice]	no underlyingly voiceless obstruents following [i, u, y]	-516.7
*VO[+voice]	no underlyingly voiced obstruents following short vowels	-517.0
*SonO[−voice]	no underlyingly voiceless obstruents following sonorants	-1300.1
*OO[+voice]	no underlyingly voiced obstruents following obstruents	-1302.1

Table 2: Constraints and their position in SOT for the voice specification of final obstruents in Dutch.

Lexeme	Onset	Vowel Type	Vowel	Coda	Obstruent	Voicing
aap	empty	long	a	None	P	voiceless
aard	empty	long	a	Sonorant	T	voiced
aars	empty	long	a	Sonorant	S	voiced
aas	empty	long	a	None	S	voiced
abrikoos	k	long	o	None	S	voiced
abt	empty	short	A	Obstruent	T	voiceless
accent	s	short	E	Sonorant	T	voiceless
accijns	s	short	K	Sonorant	S	voiced
accuraat	r	short	a	None	T	voiceless
acht	empty	short	A	Obstruent	T	voiceless
...	...	...	...	...	...	...
zwijg	zw	long	K	None	X	voiced
zwoeg	zw	iuy	u	None	X	voiced
zwoerd	zw	iuy	u	Sonorant	T	voiced

Table 3: Feature specifications in the lexicon for AML.

supracontext					distance	voiced	voiceless	homogeneity
p	long	L	None	x	0	0	0	empty
-	long	L	None	x	1	7	1	**homogeneous**
p	-	L	None	x	1	0	0	empty
p	long	-	None	x	1	1	0	**homogeneous**
p	long	L	-	x	1	0	0	empty
p	long	L	None	-	1	0	1	**homogeneous**
-	-	L	None	x	2	7	1	**homogeneous**
-	long	-	None	x	2	65	1	heterogeneous
-	long	L	-	x	2	7	1	**homogeneous**
-	long	L	None	-	2	38	34	heterogeneous
p	-	-	None	x	2	2	1	heterogeneous
p	-	L	-	x	2	0	0	empty
p	-	L	None	-	2	0	1	**homogeneous**
p	long	-	-	x	2	1	0	**homogeneous**
p	long	-	None	-	2	4	8	heterogeneous
p	long	L	-	-	2	0	2	**homogeneous**
-	-	-	None	x	3	107	4	heterogeneous
-	-	L	-	x	3	7	1	**homogeneous**
-	-	L	None	-	3	38	34	heterogeneous
-	long	-	-	x	3	65	1	heterogeneous
-	long	-	None	-	3	261	188	heterogeneous
-	long	L	-	-	3	38	38	heterogeneous

p	-	-	-	x	3	2	1	heterogeneous
p	-	-	None	-	3	7	34	heterogeneous
p	-	L	-	-	3	0	2	**homogeneous**
p	long	-	-	-	3	7	11	heterogeneous
-	-	-	-	x	4	126	5	heterogeneous
-	-	-	None	-	4	409	636	heterogeneous
-	-	L	-	-	4	38	38	heterogeneous
-	long	-	-	-	4	300	231	heterogeneous
p	-	-	-	-	4	13	59	heterogeneous
-	-	-	-	-	5	583	1101	heterogeneous

Table 4: Supracontexts for the pseudoverb *puig*.

exemplar	voicing	occurrence weighted		size weighted		
		count	contribution	count	composition	contribution
buig	voiced	4	0.10	32	8-8-8-8	0.12
duig	voiced	4	0.10	32	8-8-8-8	0.12
huig	voiced	4	0.10	32	8-8-8-8	0.12
juich	voiceless	4	0.10	32	8-8-8-8	0.12
poog	voiced	2	0.05	2	1-1	0.01
puist	voiceless	2	0.05	4	2-2	0.02
puit	voiceless	4	0.10	6	2-2-1-1	0.02
ruig	voiced	4	0.10	32	8-8-8-8	0.12
tuig	voiced	4	0.10	32	8-8-8-8	0.12
vuig	voiced	4	0.10	32	8-8-8-8	0.12
zuig	voiced	4	0.10	32	8-8-8-8	0.12
P(voiced)			0.75			0.84

Table 5: The exemplars predicting the voice specification for the pseudoverb *puig*. The column labeled 'composition' specifies the sizes of the supracontexts to which an exemplar belongs.

word	inflection	frequency	probability
tong	singular	2085	4.96e-05
tongen	plural	179	0.43e-05
tongetje	diminutive	24	0.06e-05
		2288	5.45e-05
gast	singular	792	1.89e-05
gasten	plural	1599	3.80e-05
gastje	diminutive	0	0
		2391	5.69e-05

Table 6: Frequencies in a corpus of 42 million tokens as available in the CELEX lexical database of the singular, plural, and diminutive forms of the Dutch nouns *tong*, 'tongue', and *gast*, 'guest', and the corresponding probabilities.

	expected		observed	
	F	p	F	p
Number	64.1	0.000	46.4	0.000
Dominance	27.3	0.000	27.1	0.000
StemFreq	65.7	0.000	90.4	0.000
Number:Dominance	13.9	0.000	15.8	0.000
Number:StemFreq	7.3	0.007	7.9	0.005
Dominance:StemFreq	0.6	0.434	0.1	0.828
Number:Dominance:StemFreq	0.1	0.459	0.6	0.451

Table 7: F-values for 1 and 178 degrees of freedom and the corresponding p-values for the model times (expected) and the reaction times (observed).

suffix	class	frequency	F(stem)	F(pl)	RT	model time
-er	(1)	low	112	13	556	11.3
-er	(1)	high	112	43	510	9.3
-s	(2)	low	105	11	591	15.5
-s	(2)	high	97	50	601	15.2
-en	(3) [+fem,+e]	low	107	22	568	12.1
-en	(3) [+fem,+e]	high	111	59	518	10.6
-en	(4) [-fem,+e]	low	178	100	571	14.2
-en	(4) [-fem,+e]	high	209	170	544	13.0
-en	(5) [+fem,-e]	low	115	11	588	15.7
-en	(5) [+fem,-e]	high	116	66	548	12.8
-en	(6) [-fem,-e]	low	110	17	594	15.1
-en	(6) [-fem,-e]	high	112	84	535	12.7

Table 8: German plurals cross-tabulated by class and frequency. Response latencies and frequencies of stem and plural as tabulated in Sonnenstuhl & Huth (2001).

suffix	class	observed			expected		
		t	df	p	t	df	p
-s	(1)	0.31	9	$p = 0.760$	0.53	16.59	0.601
-er	(2)	4.92	9	$p = 0.001$	2.35	17.93	0.030
-en	(3)	6.44	9	$p < 0.001$	2.95	16.70	0.009
-en	(4)	1.51	9	$p = 0.166$	1.19	17.00	0.248
-en	(5)	4.39	9	$p = 0.002$	2.80	17.06	0.012
-en	(6)	4.19	9	$p = 0.001$	2.82	12.32	0.015

Table 9: T-tests by affix. The t-tests for the observed data are as reported in Sonnenstuhl & Huth (2001), the t-tests for MATCHECK are Welch two-sample t-tests based on the model times.

Plural	Class	Family Size
-er	(1)	12.5
-s	(2)	1.7
-en	(3)	8.8
-en	(4)	3.1
-en	(5)	2.6
-en	(6)	3.9

Table 10: Mean family size of the six German plural classes in the Sonnenstuhl & Huth (2001) study.