Synthese (2007) 159:99–130
DOI 10.1007/s11229-006-9073-5

ORIGINAL PAPER

Language structure: psychological and social constraints

Gerhard Jäger · Robert van Rooij

Received: 18 May 2006 / Accepted: 12 June 2006 / Published online: 29 August 2006

Abstract In this article we discuss the notion of a linguistic universal, and possible sources of such invariant properties of natural languages. In the first part, we explore the conceptual issues that arise. In the second part of the paper, we focus on the explanatory potential of horizontal evolution. We particularly focus on two case studies, concerning Zipf's Law and universal properties of color terms, respectively. We show how computer simulations can be employed to study the large scale, emergent, consequences of psychologically and psychologically motivated assumptions about the working of horizontal language transmission.

Keywords Language universals · Evolution · Game theory

1 Language universals

There are around 5,000 spoken languages of the world today, and they are all different. Thus, natural languages are very diverse. However, most linguists share the view that the languages of the world, and the way they are used in different communities, have a lot in common. The properties that all languages have in common are called *language universals*. Linguists have claimed of many different (kinds of) things that they are language universals. For a simple example of a syntactic universal, it has been claimed that all languages spoken by humans have, and make a distinction between, nouns,

G. Jäger
Department of Linguistics and Literature,
University of Bielefeld,
PF 10 01 31,
D-33501 Bielefeld, Germany
e-mail: gerhard-jaeger@uni-bielefeld.de

R. van Rooij (✉)
Department of Humanities,
University of Amsterdam, Institute for Logic, Language and Computation,
Nieuwe Doelenstraat 15, Amsterdam, 1012 CP, The Netherlands
e-mail: R.A.M.vanRooij@uva.nl

verbs, and modifiers. Simple phonological universals have it that all languages have oral vowels, and that no language has the three vowel system */i/-/e/-/u/*. Some simple semantic universals are that all languages have color words for what we call "black" and "white", and that we don't have color words for interrupted spaces in the color vector space. Another one says that all languages have simple lexical items to express, for instance, negation, conjunction, and disjunction, mood, universal quantification, simple past, deontic necessity, and the comparative relation. These universals are all rather surfacy, stated in absolute terms, and categorial (not about use). As long as one stays at this superficial level, what is shared by languages is rather limited. There are various ways, however, in which these limitations can be overcome. Language universals can be stated in *implicational* rather than absolute terms, they can be stated in more *abstract* terms making use of a *hidden structure* (like phrase structure trees, and other abstract representational levels), and finally, they can be formulated in terms of *tendencies* or *frequency distributions* rather than in categorial terms. Within typological research going back to Greenberg (1963) most universals are rather surfacy but stated in implicational form, or, equivalently, in terms of preference hierarchies. One such a syntactic universal due to Greenberg, for instance, says that if a language has a marked singular, it also has a marked plural, but not vice versa. Another implicational one says that if a language has three vowels, they will be */i/-/a/-/u/*, and if five, they will probably be */i/-/e/-/a/-/o//u/* (Maddieson, 1982). Yet another one, but now of a semantic kind, says that if a language has three basic color words (remember that all languages have words for black and white), then the third one denotes "red"; if a language has five basic color words, the next colors are "green" and "yellow"; sixth and seventh are "blue" and "brown", respectively (Berlin & Kay, 1969). Other (one way) implicational universals are stated as tendencies. A well-known example is Greenberg's (1966) first universal: in declarative sentences with nominal subject and object, the dominant order is almost always one in which the subject precedes the object. Within generative linguistics (associated mostly with Chomsky and followers), language universals are typically absolute, but stated in a more abstract way, making use of things like phrase structure (rules), transformations and constraints thereof, and empty categories, i.e., syntactic categories void of phonological content and internal structure. One such universal, for instance, says that all NPs must have case, even if it need not be realized phonologically. In semantics, universals typically constrain which meanings can be expressed by "simple" worlds. These constraints on meanings are normally formulated in terms of logical features, such as monotonicity. One such semantic universal says, for instance, that all simple quantificational noun phrases must behave monotonically.

The above universals are all categorial in nature, and/or about language structure. But, we also have universals of language use. One example of such a proposed universal (tendency) of language use is that in co-operative conversational situations speakers give new information, and information that is relevant. We might call this a *pragmatic* universal, one that concerns language use, and not language structure. Another proposed pragmatic universal, one based on a standard truth-conditional view on semantics, is that—certainly if used as answer to an explicit question—speakers typically convey more information by their use of sentences that contain (the equivalent of) "or", or "some" than semantics would predict.

The above pragmatic universals still mainly concern speakers' knowledge of a language: a speaker is only *competent* if she knows how to use language in accordance with the above principles. Other universals of language use, however, have not much

to do with speaker's competence, and only seem to concern linguistic *performance*. These latter universals are mainly stated in terms of frequency distributions. The best known one is Zipf's law, which states that the frequency of a word decays as a (universal) power law of its rank.

2 Where do these universals come from

We have seen above that we have, roughly speaking, two kinds of universals: universals of *structure* of language, and universals of *use*. Given that grammar (phonology, syntax, semantics) is about structure, and pragmatics about use, it seems natural to assume that we can explain the universals of language use in these terms.

2.1 Pragmatic functional explanation

Within Gricean pragmatics, one tries to explain regularities in linguistic performance when language is used to exchange information in terms of general assumptions about people's behavior in such conversational situations. Grice's guiding idea was that the communicative goal of the speaker is to find an economical means of invoking specific ideas in the hearer, knowing that the hearer has exactly this expectation. In terms of principles of rational communicative behavior—implemented in terms of the Gricean maxims of conversation—one tries to explain why people can exchange ideas in an efficient (economical) but still effective and reliable way in conversation. More in general, pragmatics tries to explain regularities of language use in terms of their *communicative functions*. Grice (1967) strongly suggested that his maxims of conversation are not arbitrary conventions, but can be derived from very general assumptions of rationality and co-operativity that govern also aspects of non-linguistic behavior. In as far as this is possible, Gricean pragmatics can explain facts of linguistic use by reference to non-linguistic pressures and causes.

For instance, if used as answer to the question "Who came to the party?", sentences like "John *or* Bill came" and "*Some* of the boys came" convey more information than just what their standardly assumed semantic truth conditional meanings do: that at least John or at least Mary came, or that at least some of the boys came. One extra piece of information that is claimed to be conveyed is that the speaker does not know of anyone else (relevant in the conversation) that he or she came. This inference is not arbitrary, but based on the general conversational maxim that speakers should provide all relevant information they posses about the topic under discussion, at least, if it is required at the current point of conversation.

In Gricean pragmatics it is assumed that pragmatic universals owe their generality to the universality of usage principles that are not essentially linguistic, and everywhere exhibited. The fact that pragmatic universal principles are not essentially linguistic made some conclude that pragmatic universals have nothing to do with the linguistic system as such. But this leaves open the question of how to explain the linguistic universals.

2.2 Universals as language specific predisposition

Within generative linguistics it is normally assumed that linguistic universals should be explained by the principles of U(niversal) G(rammar). But as frequently observed

by functionalists (e.g. Comrie, 1981), without independent empirical support in favor of a specific UG, it is not clear why such an explanation of linguistic universals is not completely vacuous. Generative linguists point to the problem of language learning to counter this objection.

One of the most remarkable facts about human languages, or so it is standardly assumed in generative linguistics, is that children learn them in a short period of time. From the input data of the language they hear, children are able to learn this language in a limited number of years. This is taken to be a remarkable fact, because children generally get very little explicit instruction about how language can and should be used. In order to explain how children can become fluent speakers, they have to be able to form a "model" of the linguistic environment that not only makes grammaticality judgments about sentences they actually heard, but also about (many more) ones that they never heard before. In the fifties, Chomsky (1957) argued quite forcefully that the behavoristic learning theory that was popular at that time—assuming that people start out as a *tabula rasa*, and making use only of simple association and blind induction—could not account for this.[1,2] Somewhat later, this argument was backed up by Gold's (1967) results in formal learning theory. Stating it very sloppily, Gold proved that learning, or identifying, an arbitrary selected language among a collection of languages by "natural"[3] data is impossible in the complete absence of prior information. Chomsky's argument against the adequacy of simple behavioristic language learning and the results of Gold (and later results in learning theory) are clear and (should be) uncontroversial. What exactly we should conclude from this, however, is not.

Chomsky himself in the sixties declared the human ability of language learning despite the limited input ("poverty of the stimulus") to be *the* central fact that linguistics should explain. Chomsky's explanation of children's ability to learn their parents' language is well-known. First he claims that humans must have a *biological* basis for language: some mental capacities must come as part of the innate endowment of the mind in order for humans to be able to learn language. This claim by itself leaves open many possibilities of what this genetic endowment could be. For instance, it does not rule out that children can learn language because of some prewired general-purpose hypothesis-formulating abilities. Chomsky's second claim, however, was that the endowment for language is not general-purpose, but rather highly *specific* for language, a genetically given Language Acquisition Device (LAD). The main argument for such a language-specific LAD seems to be a negative one: that so far no-one has been able to explain that, and how, children could learn languages as complex as Dutch, German, and Chinese as efficiently as children actually can without assuming such a specific LAD.

Hypothesizing the existence of an innate LAD raises the question where this acquisition device comes from. Chomsky himself—followed by other linguists like Bickerton, Newmeyer, and Lightfoot—used to be not very specific about this question but suggested that it might have evolved at once through a large mutation, or

[1] Perhaps more importantly, he also argued that English has the property of being self-embedding, and showed that languages with this property could not be described making use of probabilistic finite state machines. These probabilistic finite state machines play an important role in Shannon and Weaver's (1949) information theory that is closely related with behavioristic learning by association.

[2] It is not very surprising that this critique was of great importance to cognitive psychology. One might say that it triggered the "cognitive revolution" in psychology.

[3] Meaning that learners do not receive direct evidence about what is not in the language.

as a by-product of some other evolutionary development. In any case, it was denied that it was gradually selected for what most of us would think that our ability to use language is good for: communication in a knowledge-using, socially interdependent lifestyle (Pinker & Bloom, 1990).[4] According to (Pinker & Bloom, 1990), the LAD, or what they call the "human language faculy", is a biological adaptation that should be explained by natural selection. A similar view is suggested in various papers of Nowak and co-workers (e.g. Nowak, Komarova, & Niyogi, 2001).[5]

2.3 How specific is the LAD?

Pinker and Bloom make the standard generative linguistic assumption that the LAD should be characterized in terms of the principles of UG. Stated in learning-theoretical terms, it is assumed that this LAD puts a limitation on the learning space. Although a child cannot learn from "natural" data alone which one of a (rather large) set of logically possible languages is used in her linguistic environment, this becomes possible if we put severe a priori limitations on what this set could be. Thus, it is said that the set of *human* possible languages is much constrained and innately given and that it can be characterized in terms of the principles of Universal Grammar: only those languages are *humanly* possible that satisfy these principles. We can learn to make grammatical (syntactic, phonological) judgments exactly because of this limited learning space. Syntactic concepts as sentence, noun, and verb, for instance, are not abstracted from the data, but are explained by innate properties of the brain. The same holds for more abstract principles that make use of internal structure. Similarly in phonology: the inventories of speech sounds and the way they can be strung together into syllables is not something children have to learn (e.g. Chomsky & Halle, 1968). It is the structure of our innately given language-learning mechanism that explains the language universals. Conversely, through in-depth studies of individual languages linguists can discover the principles of Universal Grammar and thereby gain "deep insights" into the human mind.

It is highly controversial, however, whether the genetic endowment for language is as specific as it is sometimes proposed by generative grammarians. Moreover, there is one main problem with this way of accounting for syntactic learning. Children not only hear grammatical sentences as their input, but also ungrammatical ones. This has disastrous consequences if learning would go via Gold's procedure: children simply cannot deal adequately in that way with ungrammatical input.[6] Partly to overcome

[4] Adaptation is denied, because the principles of UG are claimed to be non-optimal for communication. However as noted by Briscoe and others, this is not really an argument against adaptation by natural selection, because the latter doesn't select for globally optimal systems, but just for locally optimal ones. Others have argued that because the language faculty is an undecomposable unit, gradual evolution is impossible, because no part of it would have any function in the absence of other parts. But as noted by Komarova and Nowak (2003b), also this argument is not valid, for in this sense human language is no different from other complex traits: *all* biological systems consist of specialized components.

[5] Hurford (1989) also claimed to explain the biological evolution of the LAD, that assumes a two-way mapping between forms and meanings, but Oliphant and Batali (1997) showed that we don't need such bias for bidirectionality in order to explain the bidirectionality of languages. For more discussion on, among others, the specificity and possible evolution of the LAD, see the articles in Christiansen and Kirby (2003).

[6] Though see Osherson, Stob, and Weinstein (1986) for some generalizations of the Gold paradigm that seek to handle non-ideal input.

this problem, Horning (1969) and others have argued that the LAD should not be characterized (completely) as a proper subset of the space of possible languages, but rather in terms of a prior preference ordering on the set of all grammars, or system of rules, that (weakly) generates a (context free) language. Thinking of the LAD in this way as an inductive bias,[7] it becomes possible for learners to make use of the fact that the linguistic input to a learner most likely corresponds to the frequency distributions found in the external language. Horning (1969) already showed that once this is taken into account,[8] learners are in principle able to identify (context free) languages. This result is important because it shows that for language learning to be possible, we are not forced to assume a LAD with specific linguistic constraints.[9] Unfortunately, the particular way in which Horning represented the inductive bias did not result in efficient language learning at all, and more standard ways of representing such a bias (in terms of Minimal Description Length) are not likely to enable learners to identify (or approximate) the target language successfully without any additional assumptions. However, it seems clear that we *can* make some additional assumptions that help learners a lot, and that this can explain at least some language universals.

2.4 The role of structure and of semantic input

Above we discussed two ways by which a grammar could, in principle, be learned from positive data alone. We either presume that the search space is severely restricted, or we assume that learners are endowed with a probabilistic inductive bias. According to both solutions to the problem of how to learn a language by limited evidence it is assumed that the linguistic input is just a set of unstructured strings, with no meanings.

Already very early on Chomsky (1962) suggested that even if children receive semantic input, this is largely irrelevant to solving the problem of language learning:

> ... it might be maintained, not without plausibility, that semantic information of some sort is essential even if the formalized grammar that is the output of the device does not contain statements of direct semantic nature. Here, care is necessary. It may well be that a child given only the input of LAD as nonsense elements would not come to learn the principles of sentence formation. This is not necessarily a relevant observation, however, even if true. It may only indicate that meaningfulness and semantic function provide the motivation for language learning, while playing no necessary part in its mechanism, what is what concerns us here. (Chomsky, 1962)

Thus, or so Chomsky suggested, we'd better think of language learning from a purely syntactic point of view. Semantic, or even pragmatic input has at most a motivational function but does not really help to solve the real problem, i.e., the problem of how children learn the principles of sentence formation, i.e., of syntax.

Suppose that the linguistic input is not just a sequence of strings of grammatical sentences, but rather a sequence of strings with a functor-argument structure, or even

[7] A special case of a prior distribution is one where a finite number of grammars is expected with equal probability and all other grammars are expected with zero probability, which is equivalent to a finite search space. See also Komarova and Nowak (2003a).

[8] Not making use of probability distributions, Gold assumed that the learner may see an unrepresentable sample of the language.

[9] This also means that few, if any, specific linguistic universal can be explained in terms of such a LAD in a direct way.

a sequence of strings together with a full meaning. In this way a learner receives much more input, and it seems obvious that in this way learning a language becomes much easier. But this is not the case, at least if meanings themselves can be represented by strings of symbols (e.g. in terms of first-order logic). In formal learning theory languages are thought of as in the theory of formal grammar. This means that a sentence is taken to be a finite string of symbols drawn from some fixed, finite alphabet, and a language is construed as a subset of all possible sentences. This definition allows for rich conceptions of sentences, for which function-argument structures and meanings are parts of sentences. Since structured objects of almost any nature can be collapsed into strings of symbols drawn from a suitable extended vocabulary, it is sufficiently general to construe a language as the set of such strings. On this more general view, what is meant by a language depends on how much information one puts into a signal. In the previous section we thought of a language as a set of unstructured strings without meaning. In that case, a learner is successful if she can uniquely identify the whole set of unstructured strings that constitute the well-formed sentences from a limited set of unstructured strings that form the input. However, when we take structure or even meaning into account, not only the input is richer, but also the task gets more demanding. If we assume that the input is a set of strings with functor-argument structure, it is only natural to ask from a learner not just to identify the correct set of well-formed sentences, but also that she should be able to give those sentences their correct functor-argument structure. It is natural to make a similar demand if a learner receives as input the set of well-formed strings together with their meanings: in that case the learner should be able to identify the set of all sentences together with their meanings. But once we strengthen the task of the learner analog to the received input, very little is gained. In fact, Gold's theorem applies in that case in precisely the same way as it did for the learning from and of unstructured strings: without prior knowledge, a language of structured strings, or of strings with meanings, is not identifiable by positive data alone.

But notice that these negative results are based on the assumption that there is no limit on (i) the set of grammatical categories that words could have, (ii) the set of meanings that could be expressed in natural languages, and (iii) the possible ways in which the elements of these sets of forms and meanings can be related to each other. Moreover, it has to be assumed that meanings are structured objects that can be represented by strings of symbols. All these assumptions are doubtful, however. The interesting thing is that giving up one or more of these implausible assumptions might simplify language learning considerably.[10] Unfortunately, giving up one or more of these assumptions by themselves, together with a general purpose learning mechanism, will still hardly be enough to explain linguistic rules, let alone linguistic universals.

3 Cultural evolution

Pinker and Briscoe propose that language universals, or a language specific learning device, can be explained evolutionarily by natural selection. But evolution is not restricted to genetic evolution. Evolution is always possible for any representational mechanism once we have either enough variation or the repeated introduction, for whatever reason, of new representational mechanisms, and a feedback loop between the occurrence of a specific representation and selective "success". In order

[10] See Kanazawa (1998) for some important results concerning (i).

for positive or negative feedback to be possible, we have to take the environment in which the representation takes some role into account as well. The environment can be the purely physical environment, but for our purposes the environment of other language users is more important. In the latter case, the feedback is then one either between an abstract learning mechanism and the stability of a language, or between the individual's use of a language and the use of it in a population. Both might help to explain optimization of certain properties of language.

How would we explain linguistic universals due to non-genetic evolution now? Such universals are then linguistic forms that emerge in all kinds of populations independent of their prior state, and once emerged are stable over time. Thus, for whatever reason new linguistic forms are introduced, once introduced they must be used, and thus reproduced by others in order to have a chance to become established in a population. In order for a linguistic form to become a universal, the introduction and stability under evolutionary pressure must even be inevitable. This means that other agents must both be *able* to use these linguistic forms, but also must have an *incentive* to do so. In order for an agent to be able to use a new linguistic form, the agent must (i) be able to *learn* how to use (and understand) this linguistic form (learnability), and (ii) the form must be relatively easy to use (not too costly to produce and process).

In the end, we end up with a language that is learnable, parsable, and useful in communication. This suggests that from a cultural evolution point of view, there might be three reasons why some linguistic forms are more successful than others: (i) some forms are better to *learn* than others; (ii) some forms are less costly in *processing*, and (iii) some forms are more *useful* in actual conversation. We believe that all three indeed have an influence on how language is used and structured, and we will discuss them in turn.

3.1 Vertical transmission: the learning bottleneck

Consider a set of points in a vector space. If this set is finite, it is always possible to describe this set by a look-up table. In case the set is very large, however, this description will be very large as well. When the set has some internal structure, the description of the set can be compressed, for instance by an equation or a function. If one wants to learn how to describe such a set without given the full set as data, however, one can only give a full description of the set if one makes generalizations. In formal learning theory it is assumed that language learning is not very different from learning such a (finite or infinite) set. Also here making generalizations is crucial. Making any generalization, however, is risky: one might make false predictions about points not in the input set. Of course, there are always several different generalizations possible, and the problem is which one to choose. Here is where the inductive bias comes in. It is well-established that one set of rules (the grammar or automaton) that describes (generates, or accepts) the input data and contains some generalizations is better than another if the former is simpler than the latter, where simplicity can be measured in terms of the length it takes to describe these rules (or automaton) together with the data. It is then assumed that by the inductive bias a learner adopts those set of rules that results in the shortest description of the data (e.g. Stolcke & Omohundro, 1994; Wolff, 1982). More likely than not, this procedure by itself still needs strong assumptions about what are possible languages (grammars) or of the inductive bias being used in order for learners to identify the target language by a limited set of (positive) data.

However, things are less hopeless if one does not stick with learning of an arbitrary target language by an individual in one generation, but considers *iterated learning*. Iterated language learning is studied extensively by Kirby and Hurford (2001), Brighton (2002), and others. The iterated learning model tries to factor out the effects of utility and psychological complexity to investigate to what extent language structure can be explained in terms of learnability restrictions alone. The idea is that learners have a learning mechanism (for instance one that makes use of the Minimum Description Length Principle) that is biased in favor of some hypotheses (grammars) over others. Moreover, it is typically assumed that the linguistic input consists of a set of sentence-meaning pairs. In this model, a child selects a grammar from a limited input set just like in standard learning procedures. Now, however, it is assumed that this same individual becomes an adult and uses this grammar to produce the (limited) input from which a child of the *next generation* has to induce a grammar. This child, in turn, becomes an adult producing input for the generation next to that, and so on. Children of all generations are assumed to share the same inductive bias for language learning. This last assumption is crucial. With a learning mechanism biased for (certain) generalizations from input data, it will always be the case that once an adult produces an input to a child of the next generation that has some internal structure—either by chance or because of the grammar used by this adult—the child that learns according to this bias will select a grammar in which the input is described in a rule-like way. Consequently—or at least for rules that are always used enough by adults to produce input for children—once a child selects a grammar that makes a certain generalization, this generalization will be picked up by children of the next generations because they use the same inductive bias as the adult. What we end up with in the end, although perhaps only after a great number of generations, is a language that is adapted to the learning mechanism: a language generated by a grammar that can in fact be identified by a child that makes use of the same learning mechanism as his parents and grand-parents.[11] Kirby (1999) shows that in terms of iterated learning quite a number of linguistic universals can be explained. Notice that in this way iterated learning constrained by poverty of input data can *explain* rather than has to postulate why languages make use of rules, can be given a short description, and can be learned from a limited set of data, to the extent that these latter features are not already built into the learning mechanism itself. Only those languages are humanly possible that are stable under cultural evolution due to iterated biased learning under a learning bottleneck.

As we have seen above, although it is very natural to assume that when children learn a language, they learn positive examples in combination with the meanings expressed by these examples, this by itself doesn't necessarily help a lot. Although they might observe what objects and actions are being referred to by the expressions they hear, Gold's result shows that this by itself won't enable them to correctly infer the grammar of the spoken language, now thought of as the whole set of rules that connects expressions of the language with their meanings represented by strings of symbols. However, what if the learner had some prior information on the set of meanings that could be expressed?[12] This would help a lot if the given semantic information is (type)-structured and is known to correspond closely with the syntactic structure of

[11] See also Zuidema (2003) for illuminating discussion.

[12] Indeed, something like this is generally assumed by proponents of cognitive grammar like Lakoff (1987): prior knowledge of syntactic principles is rejected in favor of innate knowledge of image-schematic structure. Chomskyans have argued that not much can be gained by this move: as long as this semantic prior knowledge is language specific, cognitive linguists are equally much committed to

the target language,[13] i.e., if we can assume that the language behaves compositionally. A language behaves compositionally if the semantic meaning of a complex expression is a function of the meanings of its constituent parts and the way in which they are put together. One can show that in natural circumstances a compositional language is advantageous under a learning bottleneck.

Suppose, for instance, that the shared meaning space T is a set and that a language can be thought of as a function that takes elements of T and expresses these via elements of M, the messages. For instance, if T is $\{a,b,c,d\} \times \{u,v\}$, each language would need eight different messages to be able to express all the elements of T. The set of all languages $\mathcal{L}$ is then the set of all injective functions from T to M. Some of these languages are "better" than others: one language, ℓ, assigns to each element of T a simple message: $\ell(a,u) = m_1, \ell(a,v) = m_2, \ell(b,u) = m_3, \ldots, \ell(d,v) = m_8$, where m_i is shorter than m_j iff $i < j$; another one, $\jmath$, used two symbols, α and β, to encode the meanings as follows: $\jmath(a,u) = \alpha\alpha\alpha$, $\jmath(a,v) = \alpha\alpha\beta, \ldots, \jmath(d,v) = \beta\beta\beta$; while yet another one makes use of four messages (α, β, γ and δ), but behaves compositionally by using the internal structure of the message: $\kappa(a,u) = \alpha\alpha, \kappa(a,v) = \alpha\beta$, $\kappa(b,u) = \beta\alpha, \ldots, \kappa(d,v) = \delta\beta$.

Which of those languages is to be preferred? Well, that all depends on what more is known about the meaning space. For instance, suppose that T can be thought of as naturally ordered by an x and y axis with obvious 0-points. In that case, the first coding is very systematic and efficient. If T is not ordered this way, the first language is still efficient, but not systematic and thus hard to learn. Suppose, as another example, that their exists a likelihood pattern with which the elements of T occur that can be described by a probability measure P such that $P(a,u) > P(a,v) > P(b,u) > \cdots > P(d,v)$. Now, language ℓ is again systematic and easy to learn, but also short and thus efficient. What if there are neither obvious ordered x- and y-axes with 0-points, nor a commonly shared probability measure over T? In that case only the latter, compositional language seems natural: it still describes the meanings in a systematic and compressed way by making use of the fact that each element of T has two features. In general, if the meaning space is known to have a (limited) set of features, compositional languages have an evolutionary advantage under a learning bottleneck, and, indeed, Kirby and Hurford (2001) show that under such a bottleneck, compositional languages are selected for by iterated learning in case children have to learn from a representable sample of the language. Thus, if we can assume that adult and child know that the meaning space is so structured and that semantic information is part of the input on the basis of which children learn their language, we can explain why languages evolve to become compositional, because in this way they can be acquired under a learning bottleneck.[14]

Footnote 12 continued
an innate UG as Chomskyans are. True, prior information is still assumed in this way, but not ones of such a specific kind as Chomskyans do. If one can explain universal features of languages in terms of these more general assumptions, certainly something is gained.

[13] See Dudau-Sofrnie, Tellier, and Tommasi (2003) for some formal results.

[14] The assumptions of this explanation—that adult and child know the structure of the meaning space, and that all semantic information expressed by a sentence is part of the input, and does not have to be inferred—are rather strong, however, and disputable. It seems that also at least part of the structure of the meaning space has to be learned, or constructed (for some recent experiments see Kirby (2005)), and that what is intended to be expressed by a sentence has to be inferred. In fact, people like Tomassello point to the ability to make the latter kind of inferences as that feature of humans that sets us apart from (other) animals.

3.2 Horizontal transmission: pressure of language use

In the previous section we have seen that linguistic structure can to some extent be explained by a learnability constraint: for a language to be adaptive, or stable, it has to be learnable by children based on a limited set of data: the learning bottleneck. But, obviously, learnability is not the only constraint we have to enforce upon successful languages. First of all, a language must not only be learnable, it must also be *expressive*.[15] A language that is able to express only a few distinct meanings is easy to learn, but not very useful from a communicative (or representational) point of view. Second, an expressive language must overcome some other bottlenecks as well. Perhaps the most obvious bottleneck is that agents have to be able to use and comprehend language rather rapidly. For one thing, it is crucial for speakers that it doesn't take too much time to express their believes, desires, and intentions, and this puts a pressure on language form, most obviously on phonetic articulation. For another, it is important for hearers to be able to figure out the communicative intention of a speaker expressed in a sentence in a rapid, but still reliable way. This involves not only *pragmatic reasoning* to calculate what the speaker could have meant but did not explicitly say, but first of all to be able to *parse* the sentence to determine the conventional, semantic meaning of the sentence rapidly, in real-time.

There is a long tradition explaining universal features of language in terms of difficulties of encoding and decoding of information. Zipf (1949) proposed to explain some such features in terms of a unifying *principle of least effort*. This principle works in contrasting directions for speaker and hearer, however, and this appears at many levels. As for the phonological level, speakers want to minimize articulatory effort and hence encourage brevity and phonological reduction. Hearers want to minimize the effort of understanding and hence desire explicitness and clarity. As for the lexical level, speakers tend to choose the most frequent words, because the availability of a word is positively correlated with its frequency. The effort for the hearer, however, has to do with determining what the word actually means. The higher the ambiguity of a word, the higher the effort for the hearer. As for syntax, one can imagine that especially the constraint that the sentences of a language can be parsed rapidly has left their mark on languages. In this section we will discuss those functional pressures on language structure, starting with the latter.

3.2.1 Hearer-based motivations

Parsing pressure. Generative linguists normally think of a grammar as that what constitutes the agent's (purely linguistic) knowledge of the language she speaks, her linguistic *competence*. Still, on that view, the grammatical form of a sentence is also important for language *performance*: both in speech production as in speech comprehension. Many linguists have tried to explain certain (universal) grammatical constructions, or constraints, by pointing out that these constructions and constraints enable hearers to parse and process sentences of a language rapidly in real-time. A grammatical form that allows for rapid processing has an advantage: it handles the parsing pressure rather well.

[15] Of course, to measure the expressiveness of a language, one has to assume an independently given meaning space. But as noted in the previous footnote, it seems not unnatural to assume that the meaning space itself has to be constructed, learned, or has to evolve together with the language, perhaps fueled by functional motivations.

What a grammatical system has to explain, for instance, is the fact that (la) and (lb) are grammatical, but (1c) is not.

(1)a. Who did John meet?
b. Who did you believe John met __?
c. *Who did you believe the fact that John met __?

Ross (1967) proposed to account for this by his so-called "Island constraints". A set of constraints that forbids movement of, for instance, *wh*-expressions from complex noun-phrases like "the fact that ..." in (1c). Constraints like this play an important role in transformational grammar: they limit the generative power of full transformational grammars and so help to make them psychologically more respectable.[16] But, of course, there are many constraints on movement one could think of that would give rise to such limitations, so we still would like to have an independent motivation for these Island constraints. Berwick and Weinberg (1984) have argued that these constraints (or the more general principle of Subjacency) can be motivated by parsing pressure: the parser is able to represent the left context in a finite way required for parsing decisions only if the constraint is satisfied.

According to Hawkins (1994), there is pressure to shape language so that the hearer can determine the structure of the sentence as rapidly as possible. On the assumption that speakers, when they have a choice, follow parsing preferences, it follows that facts about grammar will reflect these preferences. One such fact, for instance, is that long (or heavy) elements tend to come after short (or light) ones in V(erb) O(bject) languages. In general, Hawkins argues that word-order facts can be reduced to effects of parsing pressure to a great extent.

But how can parsing pressure have any effect on linguistic structure at all? Even if all human languages are shown to be optimal with respect to parsing pressure, this does by itself not show why existing languages are so optimal. We still need to explain how processing optimization plays a role in selecting languages, what Kirby (1999) calls the *problem of linkage*. He proposes to solve this problem by linking parsing with acquisition. The parser is proposed to works as a filter between the utterances used in a speech community and the input actually picked up by a child to acquire a language. If the parser prefers some utterances above others, it is argued that the former will have a greater impact on the child's shaping of principles of sentence formation than the latter. Over many generations it means that languages adapt to the parsing pressure.

This line of explanation thus ultimately explains the adaptation for parsability as a result of vertical transmission. Alternatively, a case can be made that this effect is a consequence of selection during horizontal transmission. Briefly put, language users show a tendency to repeat linguistic material from the immediately preceding context. If a certain item, like a word, a phrase, or also a certain construction, has been used before, the likelihood for it to be used again has increased. This effect als works in a

[16] Peters and Ritchie (1973) proved that Chomsky's (1965) transformational grammars had the generative power of full Turing machines. This had the devastating consequences that (i) the notion of grammaticality that followed from that model might be undecidable, but also (ii) that the language is not learnable. This was a serious blow to the intended psychological interpretation of that model: such an interpretation could only be saved by putting severe constraints on the generative power of such a grammar. Ross' Island Constraints were one of the first of this kind. Other generative linguists, like Gazdar, Pullum, Sag, and Klein (1985), were more radical and proposed to get rid of transformations altogether.

bidirectional way: somebody who just heard (and processed) a certain construction is more likely to use it herself than somebody in a neutral context (see for instance Branigan, Pickering, and Cleland (2000) for experimental evidence). A prerequisite for such a *priming* effect to occur is, however, that the priming trigger has not just been heard but also processed.

Applied to priming of syntactic constructions, this means that constructions that are harder to parse, and which are thus more likely to be not understood, are less likely to be repeated than easier constructions. This will result in a skewing of the frequency of constructions in actual language use in favor of structures which are easy to parse. Over an extended period of time, such a skewing of frequencies can lead to a change in grammar.

Perspectual distinctiveness and semantic comprehension. If the function of language is to transfer information, it must be clear to the hearer what the speaker wants to communicate by her use of an utterance. But this means that the utterance has to be well-understandable: from the articulational level to the pragmatic one. This constraint can have important effects on linguistic forms.

It has been found by linguists that the vowel system of human languages show a number of regularities: some vowels occur more often than others, and some combinations also occur more often than others. It has often been suggested (e.g. Jakobson, 1968) that these regularities can be given a functional motivation: languages that observe these regularities are optimized for acoustic distinctiveness between the different vowels in the repertoire. In this way we can explain why, if a language has three vowels, these typically take the shape of */i/, /a/*, and */u/*, i.e., occupy the areas farthest apart in phonological space. But how should we think of this functional explanation? Intuitively, languages don't satisfy the above functional constraint because humans explicitly optimize the vowel systems they learn: humans don't do that, they normally just try to imitate the vowel system of their peers. Instead, languages satisfy the above functional constraints because they limit the ways sound systems could change. As evolution theory clearly explains, only changes that result in sound systems that handle the functional pressure better than the current one have a chance to be adopted by a population of language users. This general point has been clearly shown in simulations performed by de Boer (2001). The vowel system with which his simulation runs end up with are very similar, and they are very close to the vowel system consisting of */i/, /e/, /a/, /o/, /u/*, the vowel system that occurs most frequently in human languages (in 88 percent of the languages that have five vowels).

Similarly, language forms should facilitate the recovery of the semantic content from the syntactic structure. It has been argued that some universals of syntactic forms can be explained in this way: violation of the universal would make such recovery more difficult. In syntax, for instance, it is natural to assume that case marking evolved because it helps listeners to identify the argument-places of the noun phrases in sentences with a (di)transitive verb (e.g. Comrie, 1981). As for semantics, according to Croft (2000) there exists a tendency in linguistic communities to avoid complete synonymy: alternative forms are used later to express more specific distinct meanings.[17]

[17] This subsection might misleadingly suggest a straightforward form–function correlation. Of course—as is now generally recognized with Optimality Theory—an explanation of linguistic form in terms of communicative function can in general not be that simple. Sometimes, a linguistic form that could be motivated by some functional constraint(s) doesn't respect other constraints, and it depends

3.2.2 Speaker-based motivations

Economy, and text frequency. Speakers have a tendency to transfer their communicative intentions with the least amount of effort as possible. This is arguably important in semantics and pragmatics to explain the evolution of short expressions whose meaning depends on context (cf. van Rooij, 2004) and in syntax to explain, for instance, differential case marking (cf. Jäger, 2004), but is most clearly visible in phonology. There is abundant experimental evidence that expressions are phonologically reduced under repetition (Fowler & Housum, 1987). One might expect that this is related to the informational value of an expression—repeated expressions are arguably more expected and thus less informative (this idea has been pursued in Lindblom (1990)). However, Bard et al. (2000) showed that phonological reduction under repetition is an automatic process that is independent of the informational state of the listener (or the speakers' conception of this state).

If speakers do this systematically, it doesn't seem to be a very daring hypothesis to say that this leaves its mark on the language system. In particular, it will result in the reduced forms of (older) expressions that are frequently used. But this suggests that historical paths are possible where linguistic items that are often used get shorter, though not the other way around. This is exactly what we find in language change: full lexical items (like nouns and verbs) for instance can become "grammaticalized" to grammmatical items (like auxiliary verbs or pronouns), which may be further reduced to affixes (which in turn may disappear altogether). Such historical paths are frequently attested, while the reverse direction (from affix via clitic to grammatical word further to content word) do not occur.[18]

Another consequence of the tendency to phonological reduction is that in a stable state, there must be a correlation between frequency of words and their length in the sense that the most frequent words are the shortest ones. This correlation has been noticed already by Zipf (1935):

> "In view of the evidence of the stream of speech we may say that the length of a word tends to bear an inverse relationship to its relative frequency; and in view of the influence of high frequency on the shortenings from truncation and from durable and temporary abbreviatory substitution, it seems a plausible deduction that, as the relative frequency of a word increases, it tends to diminish in magnitude. This tendency of a decreasing magnitude to result from an increase in relative frequency, may be tentatively named the Law of Abbreviation." (Zipf, 1935:38)

Prestige and social status. In the first section of this paper we stated the tendency of speakers of providing relevant information as a pragmatic universal. Indeed, it is standardly assumed within pragmatics that communication is a co-operative affair: we exchange information in order to help the hearer. It is clear that receiving relevant truthful information is beneficial to hearers. But this by itself doesn't explain

Footnote 17 continued
on the interaction between those constraints whether the linguistic form will be part of the language or not.

[18] We are of course not claiming that grammaticalization can exclusively be explained by phonological reduction—it can't. Our point here is that there is a well-attested general pattern of unidirectional language change that involves phonological reduction, while no such processes are known that would lead would reverse phonological reduction.

why *speakers* typically present such truthful and relevant private information. What could be the advantage for the speaker to do so? It is clear that once the preferences of hearer and speaker are perfectly aligned, the extra energy of sending a message might still be worth its while for the speaker. However, it is doubtful that these pre-conditions are always met in actual conversation. But how, then, can we still explain communicative behavior in these situations? Reciprocity (Trivers, 1971) and correlation (Skyrms, 1996) have been suggested to overcome the similar, though more general, problem of explaining co-operative behavior of selfish agents. Though both explanations certainly play a role, neither of them can explain why speakers are sometimes so eager to provide relevant information, even without checking whether the receiver is really co-operative.

There is another way to solve the above problem that explains the lack of reluctance of speakers to share information, i.e., that explains our above pragmatic universal. According to it, sending messages that contain relevant, or useful, information is beneficial to the speaker because it *increases her social status*.[19] If the usefulness (in whatever way) of this increased social status is higher than the cost of sending the message and sharing the information, speakers have an incentive to give away useful private information. In this way, co-operative communicative behavior is explained, rather than presupposed.

Increase of social status due to communication need, of course, not only be due to the *content* of what one says. Sometimes the *form* itself can already have the desired effect. Sociolinguists have proposed, for instance, that the introduction and use of polite pronouns and the use of dialect can be explained by social usefulness, while in van Rooij (2003) it is argued that the complex forms in terms of which polite requests are made can be explained in terms of social costs. Both explanations are *speaker-based*: the choice of expression used in communication is directly related to the speaker's interest.

4 Two case studies

Functionalist explanations of linguistic universals are sometimes criticized on methodological grounds. Such explanations, one might argue, are necessarily "postdictions" because the consequences of a would-be functional explanation—the linguistic universals at hand—are already known in advance, and we cannot conduct experiments to test whether natural language would look differently if a certain side condition were different.

One way out of this kind of deadlock is to construct models of the investigated domain and to experiment with the model. This approach has become increasingly popular among functionalists in recent years. There are some attempts to come up with explicit mathematical models of the dynamics of language use (including language learning), like the mentioned work by Nowak and his co-workers (who uses evolutionary game theory and Markov processes to model iterated language acquisition), and the work of Nettel (Nettle, 1999a, b) who works with social network theory.

Another, perhaps less explanatory but currently more viable way is to use computer models of language use. Here too, the large scale consequence of varying side

[19] This point was suggested by Zahavi and Zahavi (1997) and has been elaborated extensively by Dessalles (1998).

conditions can be studied, and thus functional explanations be tested. Within the last 10 years or so, an immense amount of research has been conducted on this basis (see for instance the collections Cangelosi and Parisi (2002) and Kirby and Christiansen (2003)). In the sequel, we will present two such studies to illustrate the potential of computer simulations for the explanation of linguistic universals.

4.1 Zipf's law

Most linguistic reserach on language universals is concerned with invariant properties of linguistic *competence* across languages. Linguistic performance, however, also exhibits universal characteristics.

The most famous of those is perhaps the already mentioned "Zipf's Law" which states that the frequencies of different words in a sufficiently large corpus of texts obey a power law. Simply put, if you order the words according to their frequency in a text, the first in the list occurs approximately twice as often as the second one, three times as often as the third one, ten times as often as the tenth one etc. If f_i is the frequency of word ω_i in a given corpus and r_i its rank (word ω_i is the r_ith in the list of words if ordered according to their frequency), the law says that there is a constant k such that

$$\forall i : f_i = k \cdot r_i^{-1}$$

For example, in the SUSANNE corpus of written English (which is freely available from Geoffrey Sampson's web site http://www. grsampson.net/RSue.html), the most frequent word, *the*, occurs 9,573 times, which is about twice as much as the frequency of the second most common word, *of*, which occurs 4,682 times, while the word with rank 10, *was*, occurs 1,233 times, word number 100 (*years*) 115 times etc. The frequency of all words in this corpus are plotted against their rank in Fig. 1. The straight line corresponds to the prediction of Zipf's law.

Mandelbrot (1954) notes that the actual frequency distribution are usually not a perfectly straight line, as Zipf's law predicts. He comes up with a more general parameterized law:

$$f_i = P \cdot (r + \rho)^{-B}.$$

Zipf's law is a special case of Mandelbrot's law if $P = k$, $B = 1$ and $\rho = O$. However, different values might lead to a better fit of the data. For instance, the frequency distribution in SUSANNE can be approximated quite nicely with the values $P = 40,000$, $B = 1.15$, and $\rho = 10$. This can be seen in Fig. 2.

There is an ongoing debate how Zipf's/Mandelbrot's law should be explained. Miller (1957) for instance notes that a Zipf style distribution is to be expected if language production is conceived as a stochastic process. Put simply, monkeys on a typewriter will produce texts obeying Zipf's law, provided everything between two white spaces is counted as a word. However, human speech is quite different in character from the monkey-typewriter scenario, and it is questionable how far Miller's observation carries as an explanation.

It is interesting and suggestive that Zipf's law is not restricted to language but recurs in various other social contexts. For instance the size of the cities in a country has a Zipf-style distribution. (New York is about twice as large, in terms of inhabitants, as Los Angeles, three times as large as Chicago, four times as Housten etc.; Berlin is

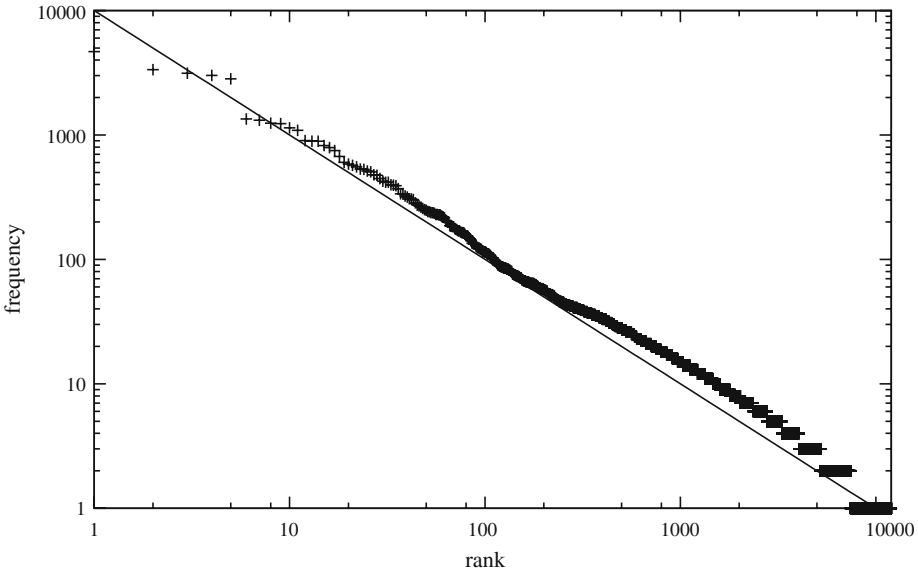

Fig. 1 Zipf's law—frequencies and ranks of the words in the SUSANNE corpus. For better visibility, both scales are logarithmic

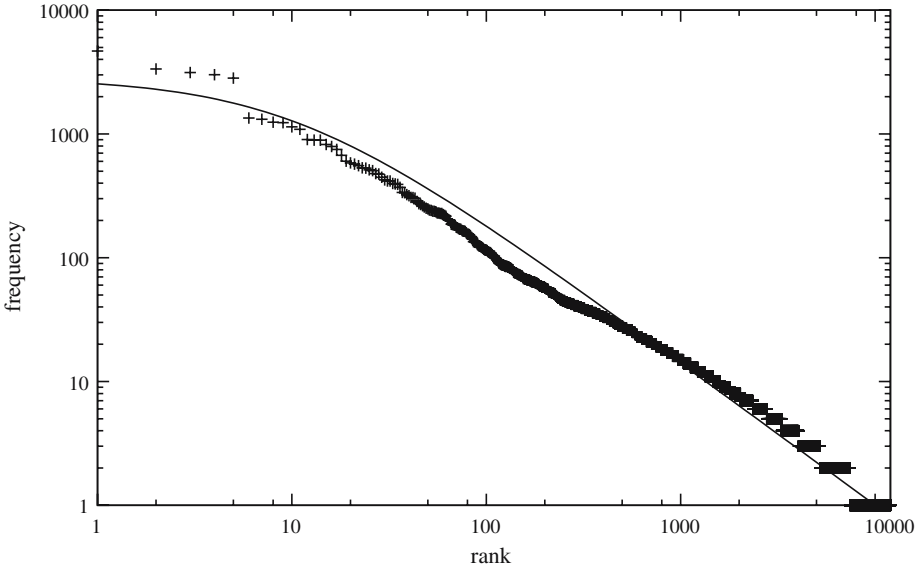

Fig. 2 Mandelbrot's law—frequencies and ranks of the words in the SUSANNE corpus

twice as large as Hamburg, three times as Munich, four times as Cologne).[20] Also, the frequency of citations in scientific journals follows Zipf's law.

Gabaix (1999) proposes an explanation for Zipf's law in the context of city sizes. He considers the following assumptions:

[20] It is a curious fact that Zipf's law also applies quite well to Dutch cities, except that Amsterdam is too small—it should have one million inhabitants rather than mere 700,000.

1. The growth rate of each city is a stochastic variable.
2. The mean value and the variance of the growth rates of all cities are identical.

This essentially says that all cities in a country grow approximately equally fast. If these conditions are met, the distribution of city sizes will approximate a Zipf-like distribution over time.

A case can be made that the frequency of words in a corpus is governed by a mathematically similar mechanism, despite all differences. A speaker/writer will tend to use those words with high likelihood that she heard or used in the immediately preceding context. There are two mechanisms which are responsible for that:

1. *Pycholinguistic priming*: Words (as well as other linguistic units like meanings, morphemes or syntactic constructions) remain activated for a certain period after they have been processed by a language user, be it actively or passively. Units that are activated in this way are more likely to be used than non-activated items.
2. *Discourse coherence*: As Tullo and Hurford (2003) point out, discourse coherence leads to a similar inertia effect. Given that we used the word *Zipf* quite frequently in the preceding paragraphs, it is very likely that we will also use it in the next sentence, for instance.

We can model this inertia phenomenon by the assumption that the probability of the usage of a word is proportional to its frequency of use in the preceding discourse. Hence the expected "growth rate" of the frequency of a word—its frequency in a given stretch of discourse, divided by its frequency in the preceding discourse, is identical for all words. The variance of these growth rates is higher for rare than for frequent words though, so it is questionable whether Gabaix' model is applicable. As an analytic question, we have to leave this open at this point, but computer simulations confirmed that the "linguistic inertia" model predicts a Mandelbrot distribution to a good approximation.

Tullo and Hurford (2003) performed computer simulations where, starting from a pool of random words, a new word was produced by drawing a random word from the set of previously used words in the entire history of the experiment. This setup as well as various variations invariably led to a Zipf-like distribution of the overall word use.

Inspired by this work, we performed a similar experiment. The setup can be summarized as follows:

1. A vocabulary of 100,000 words is given.
2. An initial word sequence of 100 tokens is drawn at random from the vocabulary (with replacement, i.e., the same word may occur several times in the sequence).
3. A word is added according to the following rule:
   (a) With a probability of 1/8, a random word is drawn from the vocabulary and appended to the sequence.
   (b) With a probability of 7/8, a random word is drawn from the last 100 words and appended to the sequence.
4. Step 3 is repeated 100,000 times.

So unlike in Tullo and Hurford's experiments, we limited the impact of the discourse history to the last one hundred words to model the rapid decay of priming effects. Also (as Tullo and Hurford in one of their experiments), we took into account

that the choice of words is not solely goverened by the discourse history. For instance, when we used the word *Mandelbrot* for the first time in this section, we did it because we wanted to talk about Mandelbrot's work, regardless of whether he had been mentioned before in the discourse.

The overall frequencies of the about 12,000 word types that were used during the simulation are plotted against their rank in Fig. 3. The results closely approximate a Mandelbrot distribution. Varying the parameters in the simulation (vocabulary size, windows size etc.) has no crucial effect on the qualitative outcome.

To summarize this discussion, Zipf/Mandelbrot's law can be seen as a consequence of the fact that the probability of the usage of a word is positively correlated to its frequency in the preceding discourse. This line of explanation of a linguistic universal has several noteworthy features:

- Zipf's law is not assumed to be part of the linguistic knowledge of competent speakers, no matter how abstract the notion of knowledge is conceived. Rather, it is an *emergent* phenomenon.
- The law is an invariant of the *usage* of language, not of its structure.
- It emerges via *horizontal* transmission. The small-scale mechanisms that lead to the large scale effect—discourse coherence and priming—are usually not asymmetric. They can be reflexive—the behavior of a speaker is influenced by her own previous behavior, and they can be symmetric. This sets this approach apart from models of cultural language evolution that focus on *vertical* transmission via language acquisition. (Language acquisition is necessarily irreflexive and largely asymmetric.)

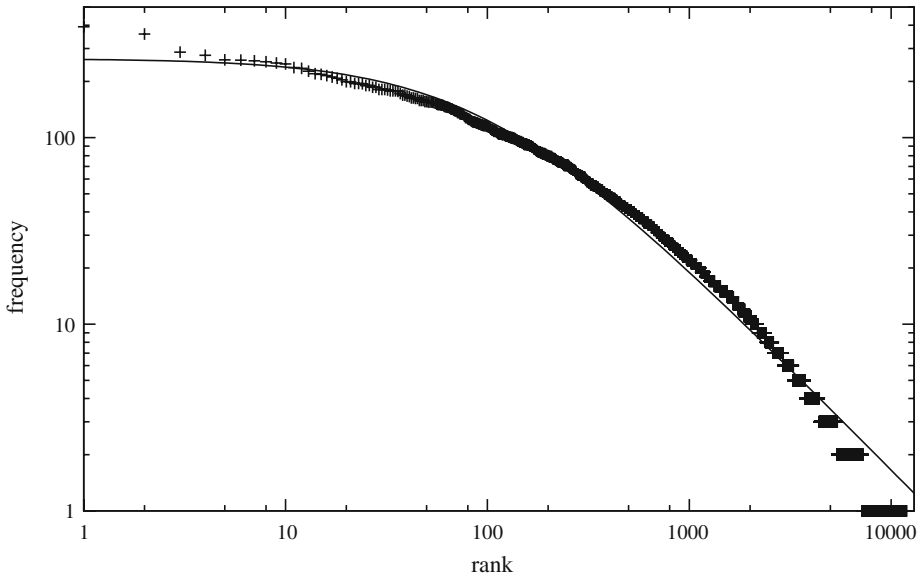

Fig. 3 Results of the simulation. The solid line gives the prediction of the Mandelbrot distribution for the parameters $P = 42,000$, $B = 1.1$, and $\rho = 100$

4.2 Convex meanings

Model theoretic semantics imposes very few restrictions on possible relations between forms and meanings. Even if a recursive syntax and an assignment of types to expressions are given, any mapping from expressions to elements of the domain of the respective type is a licit interpretation function. It has frequently been observed that human languages are much more restricted in this respect. In fact, this lack of restrictiveness has been used as one of the main arguments of cognitive semanticists against model-theoretic semantics. According to the cognitivist point of view (see for instance Lakoff (1987)), semantics maps expressions to mental representations of aspects of the world. Thus possible meanings (in the objectivist sense) must be representable mentally, and these representations must be learnable.

We believe that this is undoubtedly true, but that this picture is incomplete. Interpretations are plastic entities that are constantly being modified and reshaped. The need to communicate with an interpreted language exerts an evolutionary pressure on the relation between form and meaning. Therefore natural languages must be evolutionarily stable with respect to these forces, and this is an important source of semantic universals.

One of the best studied areas of cognitive semantics is the inventory of *color terms* in natural languages. We will use this domain to illustrate the potential of evolutionary explanations in semantics.

We need some basic insights into the psychology of color perception before we can start with the linguistic aspects. Physically, a color is characterized by a mixture of wave lengths of visible light. So in principle, the physical color space has infinitely many dimensions, one for each wave length between infra-red and ultra-violet. However, perception massively reduces the dimensionality. There is agreement in the literature that the psychological color space has just three dimensions. Various co-ordinate systems have been proposed that can be transformed into each other. One possibility (due to Smith (1978)) uses the dimensions of *brightness, hue* and *saturation*. Brightness measures the overall intensity of light—with black being the one extreme and white the other. Hue and saturation form a polar co-ordinate system. These two dimensions jointly form a circle. Saturation measures the distance from the center of the circle, and hue the angle of the connection between center and the point in question. The center of the circle is occupied by grey, while the periphery corresponds to the spectral colors—red, orange, yellow, green, blue and purple.

The Commission Internationale d'Eclairage (CIE) proposed another co-ordinate system, which uses the intensity of the three basic colors *red, green* and *blue* as dimensions. Yet another system, which goes back to Hering (1878), uses the red–green axis, the blue–yellow axis and the black–white axis.

There are natural transformations between these co-ordinate systems. Crucially, these transformations preserve the essential geometric structure of the color space. Notions like "connected region" or "convex region" are thus independent of the frame of reference.

Since the classical study of Berlin and Kay (1969), the inventory of color terms has been a topic of intensive study. We only focus on two prominent findings here. Also, following Berlin and Kay, we concentrate on *basic color terms* of a language, i.e., those terms that are morphologically simple and part of the general vocabulary of the language.

- *Color categories are always convex regions of the color space*. First, this means that the meanings of color terms are *regions*. This seems trivial, but according to model theoretic semantics, every set of points in the color space would be a licit meaning of a color word, including entirely discontinuous ones. Second, the meanings are *connected*. No color category ever subsumes a shade of yellow and a shade of blue without subsuming a shade of green as well, for instance. Third, color categories are *convex*, i.e., if two points belong to a category, every point in between belongs to this category as well. For instance, there are no languages with a category for saturated colors (that would include red and green but exclude grey, which is in between).
- There is a close connection between the number of basic color terms of a language, and the meanings of those terms. This can be formulated in terms of implicational universals. The overarching pattern is that there are six salient points in the color space, that correspond to prototypical instances of the English color terms *black, white, red, green, blue* and *yellow*. All inventories of color terms partition this set of salient colors in one way or another. Some prominent universals are:

  – If a language has two basic color terms, their meanings subsume {white, yellow, red} and {black, blue, green}.[21]
  – If a language has three basic color terms, they mean {white}, {red, yellow}, and {black, blue, green}.
  – If a language has five basic color terms, they mean {white}, {black}, {red}, {yellow}, and {blue, green}.
  – If a language has six color words, they mean {white}, {black}, {red}, {green}, {blue}, and {yellow}.

  Languages with more than six basic color terms add intermediate regions like brown, grey, orange, pink and purple to the inventory. No language appears to have more than 12 basic color terms.

Let us consider the problem of color categories from the point of view of a language designer. Suppose we need a code that enables two agents, perhaps technical devises, to exchange information about this three-dimensional color space. We only have a limited vocabulary at our disposal. The meaning space is continuous here (or at least very large, if we grant that humans can only discriminate similar colors up to a point). Therefore perfect communication, i.e., guaranteed unique reference to any arbitrary color, cannot be achieved because there are many more meanings than forms. As code designers we have to compromise. Let us say that our goal is to maximize the average similarity between the intention of the sender and the interpretation of the receiver. What would be an optimal code?

The problem has the shape of a signaling game. Nature chooses some point in the color space, according to some fixed probability function p_i. The sender S knows the choice of Nature, but the receiver R does not. S is allowed to send one out of a fixed number of signals to R. R in turn picks a point in the color space. S and R have the joint goal to maximize the similarity between Nature's choice and R's choice.

Formally, we say that M is the set of meanings (points in the color space), S is a function from M into some finite set F of forms, and R is a function from F to M. The

[21] We use the English color words here to refer to the focal colors, not to the entire region that corresponds to their denotation in English.

utility of the communicators can be defined as

$$u(S,H) = \sum_{m \in M} p_i(m) \cdot sim(m, R(S(m)))$$

where *sim* is a function that measures the similarity between two points. The precise definition of similarity need not concern us here. All what matters is that the similarity between two points is a monotonically decreasing function of their distance in the color space. By convention, similarity is always positive, and every point has a similarity of 1 to itself.

We did not specify above whose utility function this is because we assume that the interests of S and R are completely identical. Also, the signals themselves come with no costs. (The game us thus a cheap talk game of complete co-operation.)

Suppose S knows R, the interpretation function of the receiver. What would be the best coding strategy then? For each possible signal f, S knows R's interpretation, namely $R(f)$. So for a given choice m of nature, S should choose the signal f that maximizes $sim(m, R(f))$. In other words, each form f corresponds to a uniqe point $R(f)$, and each input meaning m is assigned to the point which is most similar to m, or, in other words, which minimzes the distance to m. This kind of partitioning of the input space is called a *Voronoi Tesselation*. An illustration of such a tesselation is given in Fig. 4.

Okabe, Boots, and Sugihara (1992) prove the following lemma (quoted from Gärdenfors (2000)):

Lemma 1 *The Voronoi tessellation based on a Euclidean metric always results in a partioning of the space into* convex *regions.*

The best response of R to a given coding function S is less easy to describe. Intuitively, the receiver should map each form f to the center of the region which the sender maps to f. Formally we can state that for each form f, it should hold that

$$R(f) = \arg\max_m \int_{S^{-1}(f)} p_i(m')\, sim(m, m')\mathrm{d}m'.$$

The precise nature of such an optimal sender's response depends on the details of nature's probability function p_i, the similarity function, and of course the geometry of

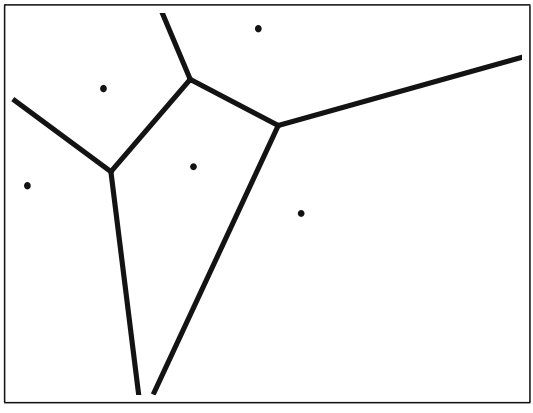

Fig. 4 Voronoi tesselation of a two-dimensional Euclidean space

the underlying space. Suffice it to say that for a closed space and continuous p_i and *sim*, the existence of such a best response is always guaranteed. Under the same restrictions, it can also be shown that there exists a (possibly non-unique) pair $\langle S, R\rangle$ such that the utility $u(S, R)$ is maximal. Obviously, such a pair must be a Nash equilibrium, i.e., S is a best response to R and vice versa.

These considerations show that it is sensible to carve up the color space into convex regions for the purposes of discrete communication. The question is why natural languages behave in such a seemingly sensible way.

A first step towards an answer is the insight that the above utility function does not just measure the efficiency of communication. Human communication is essentially *bidirectional*. A speaker always monitors and thus interprets her own utterance.[22] Also, psycholinguistic priming works both ways. Producing a certain piece of linguistic structure facilitates the subsequent use of the same piece of structure (Bock, 1986). So production primes production. However, hearing and interpreting a piece of structure likewise facilitates the subsequent production of the same piece of structure (see for instance Branigan et al., 2000). So interpretation can prime production as well. Taken together this means that an utterance by a speaker has a twofold priming effect on the speaker herself: both the production itself and the self-monotoring, which is an act of interpretation, has an impact on the probabilities of the subsequent behavior of that very speaker.

It is also crucial to observe that priming does not only occur under complete identity but also under similarity. Hence, if a speaker codes a certain meaning m with the form f, she will automatically interpret her own utterance and map f to some, possibly different, meaning m'. If m and m' are similar, the productive association $m \to f$ and the interpretive association $f \to m'$ will reinforce each other via priming. The stronger the similarity is, the stronger is the mutual reinforcement of the coding and the decoding strategy of the language user. The utility function above, which measures the average degree of similarity between input to coding and output from decoding, is positively correlated to the degree of mutual reinforcement between S and R via priming.

The degree of reinforcement in turn determines the probability with which a certain coding/decoding strategy is used in subsequent utterances. The utility function is thus directly related to the "fitness" of a strategy—strategies with a high fitness increase their probability to be used in communication. Given this, we are in a position to apply the concepts of *Evolutionary Game Theory* (EGT) here. Under this interpretation of game theory, games are played iteratively, and the utility of a strategy at a given point of time is nothing but its probability to be used at the next point in time. The original motivation for EGT comes from biology. In the biological context, strategies are genetically determined, and the probability of a strategy in a population corresponds to the abundance of individuals with the respective genetic endowment in the population.

A state of a dynamic system is *evolutionarily stable* if it does not change under the evolutionary dynamics, and if it will return into that state if minor perturbations ("mutations" in the biological context) occur. This is not the right occasion to go into the mathematical details of EGT. However, it is easy to see that an evolutionarily stable state[23] must be a Nash equilibrium. Otherwise, there would be a better response

[22] This is uncontroversial in psycholinguistics; see for instance Levelt (1983).

[23] See Maynard Smith (1982).

to an incumbent strategy than the incumbent counter-strategy. A mutant "playing" this better strategy would then obtain a higher utility than the incumbent and thus spread.

As already mentioned, every Nash equilibrium in our signaling game amounts to a Voronoi tesselation of the meaning space. This means that every evolutionarily stable state induces a partition of the meaning space into convex categories.[24]

For the purpose of illustration, we did a few computer simulations of the dynamics described above. The meaning space was a set of squares inside a circle. The similarity between two squares is inversely related to its Euclidean distance. All meanings were assumed to be equally likely. The experiments confirmed the evolutionary stability of Voronoi tesselations. The graphics in Fig. 5 show stable states for different numbers of forms. The shadings of a square indicates the form that it is mapped to by the dominant sender strategy. Black squares indicate the interpretation of a form under the dominant receiver strategy.

To sum up so far, we assumed that a production strategy and an interpretation strategy reinforce each other the more similar the input for production and the interpretation of the corresponding form are on average. This induces an evolutionary dynamics. The attractors of this dynamics, i.e., the evolutionarily stable states, are those where the meaning space is carved up into convex regions. The convexity of categories is a well-established fact in the domain of color terms. Gärdenfors (2000) argues that many other semantic domains are organized as conceptual spaces with a geometric structure. He assumes that natural meanings, i.e., natural properties and concepts, are always composed from convex regions of some conceptual space.

An explanation for the implicational universals regarding the substantive content of color terms is much harder to come by, and we will not attempt one. However, we will show with a toy example how this kind of implicational universals may be induced by the underlying ("Nature's") probability distribution over the meaning space. The remainder of the section is pure speculation though; we do not know what this probablity distribution really is, and we don't see an easy way to determine it.

This being said, let us assume that the meaning space, despite being continuous, contains finitely many, in fact very few, elements that are highly frequent, while all other meanings are so rare that their impact on the average utility is negligible. For the sake of concreteness, let us suppose that the meaning space forms a circle, and that there are just four meanings that are frequent. Let us call them Red, Green, Blue and Yellow. Inspired by the middle plane of the color space (supressing the brightness dimension), we assume that all four prominent meanings are close to the periphery, and they are arranged clockwise in the order Red, Yellow, Green, and Blue. Their positions are not entirely symmetric. Rather, they are arranged as indicated in Fig. 6.[25]

Since the similarity between two points is inversely related to their distance, it follows that Blue and Green are more similar to each other than Red and Yellow,

[24] This does not only hold for evolutionarily stable *states* but also for evolutionarily stable *sets*, i.e., sets of states that are attainable from each other in the evolutionary dynamics.

[25] We would like to stress that this example is designed for presentational purposes. We chose the parameters so that (1) the structure of the example is fairly simple, and (2) that the configuration of equilibria resembles the universal patterns of color words. In this way we achieve something like an existence proof: it is possible to derive in our setup the kind of implicative universals that typologists actually observed. No claim is made, however, that our choice of parameters has anything to do with the true psychology of color cognition.

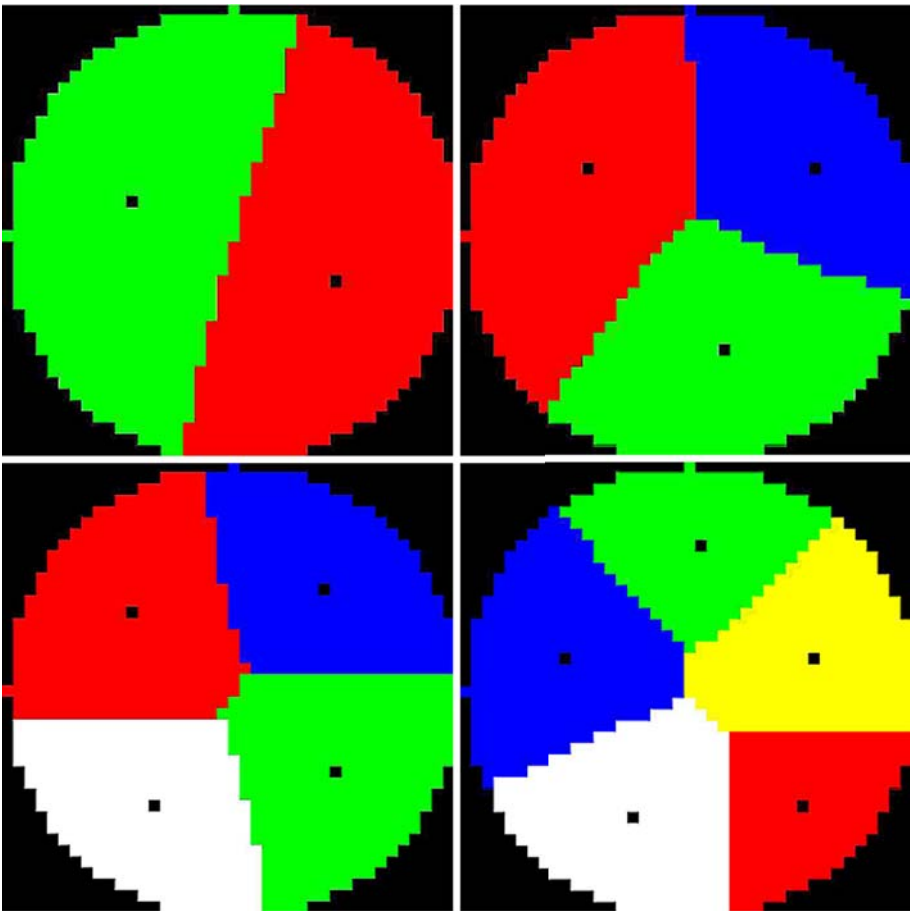

Fig. 5 Evolutionarily stable states of the signaling game with a uniform probability distribution over meanings

which are in turn more similar than the pair Green/Yellow and the pair Red/Blue. The pairs Blue/Yellow and Red/Green are most dissimilar.

Following Nosofsky (1986),[26] we assume that the similarity between two points is a negative exponential function of the squared distance between them

$$sim(m_1, m_2) = \mathrm{e}^{-\frac{\mathrm{d}(m_1,m_2)^2}{\sigma^2}}.$$

The constant σ determines how fast the similarity decays with growing distance.

We finally assume that the probabilities of all four prominent meanings are close to 25%, but not completely equal. For the sake of concreteness, let us assume that

$$p_i(\text{Red}) > p_i(\text{Green}) > p_i(\text{Blue}) > p_i(\text{Yellow}).$$

Now suppose the sender has just two forms at her disposal. What are the Nash equilibria of this game? Since this is an instance of the more general kind of game considered

[26] Cited after Gärdenfors (2000).

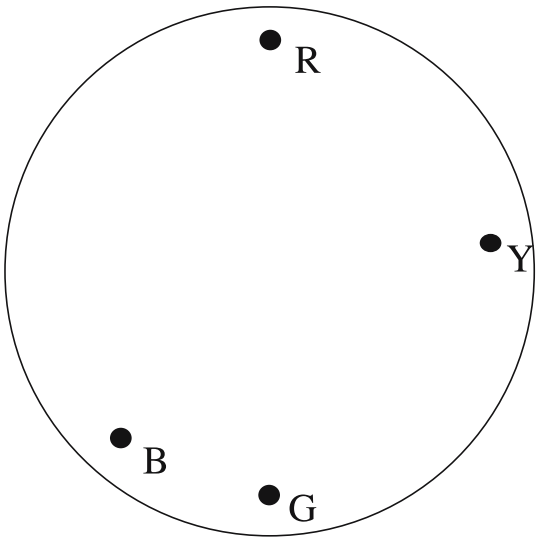

Fig. 6 Schematic arrangement of the four prominent meanings in the example

above, the Sender strategy in a Nash equilibrium is based on a Voronoi tesselation of the meaning space which is induced by the range of the receiver strategy.

A Voronoi tesselation induces a partition of the set {Red, Yellow, Green, Blue}. The partition {Red, Green}, {Blue,Yellow} is excluded because it is not convex. This leaves seven possible partitions:

1. *{Red, Yellow, Green, Blue}/{ }* This is a so-called *pooling equilibrium* in which no real communication takes place because the sender always uses the same signal regardless of Nature's choice. It is a weak Nash equilibrium; i.e., one of the strategies involved is not the *unique* best response to the other player's strategy. In our case, this applies at least to the receiver. Since in the equilibrium, the sender always sends the same signal, it does not matter for the utility how the receiver would interpret the other, unused signal.
2. *{Red}/{Yellow, Green, Blue}* If the sender strategy partitions the meaning space in this way, the best response of the receiver is to map the first signal to Red and the second one to the point that maximizes the average similarity to the elements of the second partition. If the probabilities of Yellow, Green and Blue are almost equal, all other points have a probability close to 0 and σ is sufficiently large, the average similarity to Yellow, Green and Blue is a function with three local maxima, that are located close to Yellow, Green and Blue, respectively. (The distance between these maxima and the focal points becomes negligible if σ is large enough.) So if the sender uses this partition, the best response of the receiver is to map the first form to Red and the second to the point with the highest average similarity to Yellow, Green and Blue. This is one of the mentioned three local maxima. Since, by assumption, Green is more probable than Yellow and Blue, the maximum close to Green is the global maximum. But this entails that the sender strategy is not the best response to the receiver strategy because Yellow is closer to Red than to Green, and thus the best response of the sender would be to express Yellow with the first function, alongside with Red. So this partition does not give rise to a Nash equilibrium.

3. *{Yellow}/{Red, Green, Blue}* For similar reasons as in the previous case, the best response of the receiver would be to map the first signal to Yellow and the second to Red (or rather the similarity maximum in the neighborhood of Red). But since Green is closer to Yellow than to Red, the sender strategy is not a best response to the receiver strategy, and the partition thus does not correspond to a Nash equilibrium.
4. *{Green}/{Red, Yellow, Blue}* Since Blue is closer to Green than to Red, this partition does not correspond to an equilibrium for analogous reasons.
5. *{Blue}/{Red, Yellow, Green}* This case is analogous because Green is closer to Blue than to Red.
6. *{Red, Yellow}/{Green, Blue}* The best response of the receiver here is to map the first form to Red and the second to Green. The best response of the sender to this strategy in turn is to map Red and Yellow to the first form, and Green and Blue to the second. So this partition creates a Nash equilibrium. If we identify all speaker strategies that give rise to the same partition of the meaning space, this is even a strict Nash equilibrium.[27]
7. *{Red, Blue}/{Yellow, Green}* The best response of the receiver here is to map the first form to Red and the second to Green. The best response of the sender in turn would be to to map Red and Yellow to the first form, and Green and Blue to the second. Hence this partition does not induce a Nash equilibrium.

So it turns out that with two forms, there are just two parititions that correspond to Nash equilibria: the trivial bipartition and {Red, Yellow}/{Green, Blue}. Only the latter one corresponds to a strict Nash equilibrium. In Selten (1980) it is shown that in asymmetric games (like the one we are discussing here), all and only the strict Nash equilibria correspond to evolutionarily stable states. So only the bipartition {Red, Yellow}/{Green, Blue} is evolutionarily stable.

Let us turn to the analogous game with three forms. Each sender strategy in this game creates a tripartition of the meaning space. Since we are only interested in partitions that correspond to Nash equilibria, we only have to consider convex tripartitions. All convex bipartitions are trivially also tripartitions, with an empty third cell. It is also immediately obvious that such a partially trivial partition cannot give rise to a strict Nash equilibrium. Besides, there are four more, non-trivial convex tripartitions:

1. *{Red}/{Yellow}/{Green, Blue}* The best response of the receiver is to map the first signal to Red, the second to Yellow, and the third to Green. The best response of the sender to this strategy is to use the above-mentioned partition, so this leads to a strict Nash equilibrium.
2. *{Yellow}/{Green}/{Blue, Red}* This does not correspond to a Nash equilibrium because the best response of the receiver is to map the third form to red, and since Blue is closer to Green than to Red, the best response of the sender would be to switch to the previous partition.
3. *{Green}/{Blue}/{Red, Yellow}* The best response of the receiver is to map the three forms to Green, Blue, and Red, respectively, and the best response of the sender in turn is to use the Voronoi tessellation that is induced by these three points. This is exactly the partition in question, so it does lead to a strict Nash equilibrium.

[27] In a strict Nash equilibrium, every player's strategy is the *unique* best response to the other player's strategy.

4. *{Red}/{Blue}/{Yellow, Green}* Since Yellow is closer to Red than to Green, this does not lead to a Nash equilibrium for reasons analogous to case 2.

So in this game, we have two partitions that are evolutionarily stable, namely {Red}/{Yellow}/{Green, Blue} and {Green}/{Blue}/{Red, Yellow}. There is a certain asymmetry between them because the former is *Pareto-dominant*. This means that the average utility of both players is higher if they base their code on the first partition. However, the difference is negligible for large σ. Furthermore, such a global consideration is of minor relevance in an evolutionary setting where rationality considerations, i.e., conscious utility maximization, is assumed to play no role.

There is another asymmetry between the two equilibria though. Recall that the evolutionary model assumes that the strategy choice of the players is not fully deterministic but subject to some perturbation. Using the biological metaphor, some newborn individuals are mutants that do not faithfully copy the genotype of their parents. Suppose the system is in one of the two evolutionarily stable states. Unlikely though it may be, it is possible that very many mutations occur at once, and all those mutations favor the other equilibrium. This may have the effect of pushing the entire system into the other equilibrium. Such an event is very unlikely in either direction. However, it may be that such a switch from the first to the second equilibrium may by more likely than in the reverse direction. This would have the long term effect that in the long run, the system spends more time in the second than in the first equilibrium. Such an asymmetry grows larger as the mutation rate gets smaller. In the limit, the long term probability of the first equilibrium converges to 0 then, and the probability of the second equilibrium to one. Equilibria which have a non-zero probability for any mutation rate in this sense are called *stochastically stable*.[28]

Computer simulations indicate that for the game in question, the only stochastically stable states are those that are based on the partition {Red}/{Yellow}/{Green, Blue}. In the simulation, the system underwent 20,000 update cycles, starting from a random state. Of these 20,000 "generations", the system spent 18,847 in a {Red}/{Yellow}/{Green, Blue} state, against 1,054 in a {Green}/{Blue}/{Red, Yellow} state. In fact, the system first stabilized in the second equilibrium, mutated into the bipartition {Red, Yellow}/{Green, Blue} after 1,096 cycles, moved on into a state using the first partition after another 16 cycles, and remained there for the rest of the simulation. A switch from the first into the second kind of equilibrium did not occur.

Figure 7 visualizes the stable states for the game with two, three and four different forms. As in Fig. 5, the shade of a point indicates the form to which the sender maps this point, while the black squares indicate the preferred interpretation of the forms according to the dominant receiver strategy. The circles indicate the location of the four focal meanings Red, Yellow, Green and Blue.

Let us summarize the findings of this section. We assumed a signaling game which models the communication about a continuous meaning space by means of finitely many signals. The utility of a strategy pair is inversely related to the distance between the input meaning (Nature's choice) and the output meaning (receiver's interpretation). We furthermore assumed that this kind of utility function corresponds to an evolutionary dynamics: a high utility amounts to a strong self-reinforcement of a pair of strategies. Under these assumptions, it follows directly that all evolutionarily stable states correspond to Voronoi tessellations of the meaning space. If the distance metric

[28] This refinement of the notion of evolutionary stability was developed by Kandori, Mailath, and Rob (1993) and Young (1993).

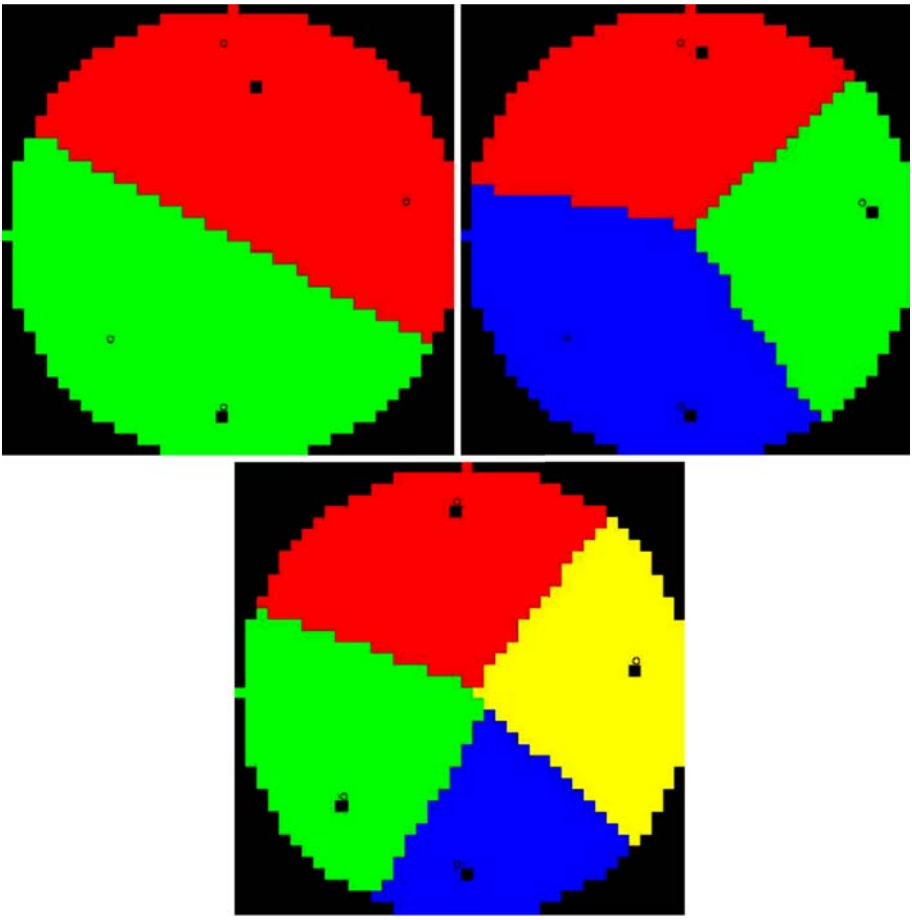

Fig. 7 Evolutionarily stable states of the signaling game with focal points

is Euclidean, this entails that semantic categories correspond to convex regions of the meaning space.

The precise nature of the evolutionarily stable states and of the stochastically stable states depends on the details of the similarity function, Nature's probability distribution over the meaning space, and the number of forms that the sender has at her disposal. The last example sketched a hypothetical scenario which induces the kind of implicative universals that are observed in natural languages with regard to color universals. Our choice of parameters was largely stipulative. We intend the example to be interpreted programmatically: while the evolutionary dynamics of the signaling has the potential to explain color universals, the precise values for the relevant parameters still have to be determined by psycholinguistic and psychological research.

5 Conclusion

In this article we discussed the notion of a linguistic universal, and possible sources of such invariant properties of natural languages. In the first part, we explored the

conceptual issues that arise. Linguistic universals can be due to features of an *innate language faculty* (the nativist view which is connected to the name of Noam Chomsky) or due to cognitive and social constraints on *language use*. The latter perspective is the focus of interest of functional linguistics. Both approaches ultimately connect linguistic universals to evolutionary processes. The nativist view attributes universals to results of *biological* evolution, while the functionalist view considers *cultural* evolution essential in shaping language.[29]

Another issue which is almost as important but less frequently discussed concerns the mode of replication in cultural language evolution. There are actually two replicative processes at work. First, language is perpetuated via *first language acquisition* of infants, which learn the language of their adult environment. Second, even after language acquisition is completed, language usage is shaped by mutual *imitation* of adult speakers. The first process is asymmetric. This kind of replication is sometimes called *vertical* transmission in the literature, and we followed that terminology. Imitation, on the other hand, is symmetric, and one can speak of *horizontal* transmission.[30]

In the second part of the paper, we focussed on the explanatory potential of horizontal evolution in this sense. We particularly focused on two case studies, concerning Zipf's Law and universal properties of color terms, respectively. We showed how computer simulations can be employed to study the large scale, emergent, consequences of psycholinguistically and psychologically motivated assumptions about the working of horizontal language transmission.

Acknowledgements We would like to thank Hartmut Fitz, Samson de Jäger, and Martin Stokhof for critical discussion, and Samson de Jäger for checking our English.

References

Bard, E. G., Anderson, A. H., Sotillo, C., Aylett, M., Doherty-Sneddon, G., & Newlands, A. (2000). Controlling the intelligibility of referring expressions in dialogue. *Journal of Memory and Language, 42*, 1–22.

Berlin, B., & Kay, P. (1969). *Basic color terms: Their universality and evolution*. Chicago: University of California Press.

Berwick, R., & Weinberg, A. (1984). *The grammatical basis of linguistic performance*. Cambridge, Massachusetts: MIT Press.

Bock, J. K. (1986). Syntactic persistence in language production. *Cognitive Psychology, 18*, 355–387.

Branigan, H. P., Pickering, M. J., & Cleland, A. A. (2000). Syntactic coordination in dialogue. *Cognition, 75*, B13–B25.

Brighton, H. (2002). Compositional syntax from cultural transmission. *Artificial Life, 8*, 25–54.

Cangelosi, A., & Parisi, D. (Eds.). (2002). *Simulating the evolution of language*. London: Springer.

Chomsky, N. (1957). *Syntactic structures*. The Hague: Mouton.

Chomsky, N. (1962). Explanatory models in linguistics. In E. Nagel, P. Suppes, & A. Tarski (Eds.), *Logic, methodology, and philosophy of science: Proceedings of the 1960 International Congress*. Stanford: Stanford University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge, Massachusetts: MIT Press.

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

Christiansen, M., & Kirby, S. (2003). *Language evolution.* Oxford: Oxford University Press.

Comrie, B. (1981). *Language universals and linguistic typology.* Oxford: Basic Blackwell.

[29] It seems obvious but still needs to be stressed time and again that those two views are not contradictory but complementary. Few people will deny that both levels of evolution are important for understanding the nature of language.

[30] This terminology has its origin in evolutionary biology, where one distinguishes "horizontal" and "vertical" transmissions of genetic information.

Croft, W. (2000). *Explaining language change: An evolutionary approach*. Harlow: Princeton University Press.
de Boer, B. (2001). *The origins of vowel systems*. Oxford: Oxford University Press.
Dessalles, J. (1998). Altruism, status, and the origin of relevance. In J. Hurford, M. Studdert-Kennedy, & C. Knight (Eds.), *Approaches to the evolution of language*. Cambridge, UK: Cambridge University Press.
Dudau-Sofrnie, D., Tellier, I., & Tommasi, M. (2003). A learnable class of CCGs from typed examples. In G. Jäger, P. Monachesi, G. Penn, & S. Wintner (Eds.), *Proceedings of formal grammar 2003*, Vienna. ESSLLI 2003.
Fowler, C., & Housum, J. (1987). Talkers' signaling 'new' and 'old' words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language, 26*, 489–504.
Gabaix, X. (1999). Zipf's law for cities: An explanation. *Quarterly Journal of economics, 114*, 739–767.
Gärdenfors, P. (2000). *Conceptual spaces*. Cambridge, Mass: The MIT Press.
Gazdar, G., Pullum, G., Sag, I., & Klein, E. (1985). *Generalized phrase structure grammar*. Cambridge, Massachusetts: Cambridge University Press.
Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*, 447–474.
Greenberg, J. (1963). Some universals of grammar with particular reference to the order of meaningful elements. In J. Greenberg (Ed.), *Universals of language* (pp. 58–90). Cambridge (Mass.): MIT Press.
Greenberg, J. (1966). *Language universals, with special reference to feature hierarchies*. The Hague: Mouton.
Grice, H. P. (1967). Logic and conversation. typescript from the William James Lectures, Harvard University. Published in P. Grice (1989), *Studies in the Way of Words*, Harvard University Press, Cambridge Massachusetts, pp. 22–40.
Hawkins, J. (1994). *A performance theory of order and constituency*. Cambridge, Massachusetts: Cambridge University Press.
Hering, E. (1878). *Zur Lehre vom Lichtsinne*. Vienna: Carl Gerold's Sohn.
Horning, J. (1969). *A study of grammatical inference*. Ph.D. thesis, Stanford University.
Hurford, J. (1989). Biological evolution of the Saussurian sign as a component of the language acquisition device. *Lingua, 77*, 187–222.
Jäger, G. (2004). Evolutionary game theory and typology. A case study. *manuscript*. University of Potsdam and Stanford University.
Jakobson, R. (1968). *Child language, aphasia, and phonological universals*. The Hague: Mouton.
Kanazawa, M. (1998). *Learnable classes of categorial grammars.* Stanford: CLSI Publicatons.
Kandori, M., Mailath, G., & Rob, R. (1993). Learning, mutation, and long-run equilibria in games. *Econometrica, 61*, 29–56.
Kirby, S. (1999). *Function, selection, and innateness. The emergence of language universals*. Oxford: Oxford University Press.
Kirby, S. (2005). The evolution of meaning-space structure through iterated learning. In A. Cangelosi, & C. Nehaniv (Eds.), *Proceedings of the second international symposium on the emergence and evolution of linguistic communication*, pp. 56–63.
Kirby, S., & Christiansen, M. (Eds.). (2003). *Language evolution*. Oxford: Oxford University Press.
Kirby, S., & Hurford, J. (2001). The emergence of linguistic structure: An overview of the iterated learning model. In D. Paresi, & A. Cangelosi (Eds.), *Simulating the evolution of language*. Berlin: Springer Verlag.
Komarova, N. L., & Nowak, M. A. (2003a). Language dynamics in finite populations. *Journal of Theoretical Biology, 221*, 445–457.
Komarova, N. L. and Nowak, M. A. (2003b). Language, learning and evolution. In M. H. Christiansen, & S. Kirby (Eds.), *Language evolution*. Oxford: Oxford University Press.
Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. Chicago: University of Chicago Press.
Levelt, W. (1983). Monitoring and self-repair in speech. *Cognition, 14*, 41–104.
Lindblom, B. (1990). Explaining variation: A sketch of the H and H theory. In W. Hardcastle, & A. Marchal (Eds.), *Speech production and speech modelling* (pp. 403–439). Dordrecht: Kluwer.
Maddieson, I. (1982). *Patterns of sounds*. Cambridge (UK): Cambridge University Press.
Mandelbrot, B. B. (1954). Structure formelle des textes et cummunication. *Word, 10*, 1–27.
Maynard Smith, J. (1982). *Evolution and the theory of games.* Cambridge (UK): Cambridge University Press.
Miller, G. (1957). Some effects of intermittent silence. *American Journal of Psychology, 70*, 311–314.
Nettle, D. (1999a). Is the rate of linguistic change constant? *Lingua, 108*, 119–136.

Nettle, D. (1999b). Using social impact theory to simulate language change. *Lingua, 108*, 95–117.
Nosofsky, R. M. (1986). Attention, similarity, and the identification-categorization relationship. *Journal of Experimental Psychology: General, 115*, 39–57.
Nowak, M., Komarova, N., & Niyogi, P. (2001). Evolution of universal grammar. *Science, 291*, 114–118.
Okabe, A., Boots, B., & Sugihara, K. (1992). *Spatial tessellations: Concepts and applications of Voronoi diagrams.* Chichester: Wiley.
Oliphant, M., & Batali, J. (1997). Learning and the emergence of coordinated communication. *Center for Research on Language Newsletter, 11*.
Osherson, D., Stob, M., & Weinstein, S. (1986). *Systems that learn*. Cambridge, Massachusetts: MIT Press.
Peters, S., & Ritchie, R. (1973). On the generative power of transformational grammars. *Information Sciences, 6*, 49–83.
Pinker, S., & Bloom, P. (1990). Natural language and natural selection. *Behavioral and Brain Sciences, 13*, 707–784.
Ross, J. R. (1967). *Constraints on variables in syntax*. Ph.D. thesis, MIT.
Selten, R. (1980). A note on evolutionarily stable strategies in asymmetric animal conflicts. *Journal of Theoretical Biology, 84*, 93–101.
Shannon, C., & Weaver, W. (1949). *The mathematical theory of communication*. Urbana: University of Illinois Press.
Skyrms, B. (1996). *Evolution of the social contract*. Cambridge, UK: Cambridge University Press.
Smith, A. R. (1978). Color gamut transform pairs. *Computer Graphics, 12*(3), 12–19.
Stolcke, A., & Omohundro, S. (1994). Inducing probabilistic grammars by Bayesian model merging. In R. Carrasco, & J. Oncina (Eds.), *Grammatical inference and applications*. Verlag, Berlin: Springer.
Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology, 46*, 195–206.
Tullo, C., & Hurford, J. (2003). Modelling Zipfian distributions in language. In S. Kirby (Ed.), *Proceedings of language evolution and computation workshop/course at ESSLLI*, pp. 62–75, Vienna.
van Rooij, R. (2003). Being polite is a handicap. In M. Tennenholtz (Ed.), *Proceedings of TARK 9*, Bloomington.
van Rooij, R. (2004). Evolution of conventional meaning and conversational principles. *Synthese, 139*, 331–366.
Wolff, J. G. (1982). Language acquisition, data compression and generalization. *Language and Communication, 2*, 57–89.
Young, H. P. (1993). The evolution of conventions. *Econometrica, 61*, 57–84.
Zahavi, A., & Zahavi, A. (1997). *The handicap principle*. Oxford, New York: Oxford University Press.
Zipf, G. (1935). *The psycho-biology of language*. Cambridge, Massachusetts: MIT Press.
Zipf, G. (1949). *Human behavior and the principle of least effort*. Cambridge: Addison-Wesley.
Zuidema, W. (2003). How the poverty of the stimulus solves the poverty of the stimulus. In S. Becker, S. Thrun, & K. Obermayer (Eds.), *Advances in neural information processing systems 15 (Proceedings of NIPS'02)* (pp. 51–58). Cambridge (Mass.): MIT Press.