Fiat Lingua

Down with Morphemes:
The Pitfalls of Concatenative Morphology

David J. Peterson

October 1, 2009

Abstract[1]

This paper explores the relationship between an adherence to the main tenets of concatenative theories of morphology and the creation of less than realistic languages.

1   Introduction: Morphology 101

In any introductory linguistics class, students will be introduced to the concept of morphology. The first lesson usually goes something like this:

> *In English, we have a word like "cat" and a word like "cats". There is a systematic relation between the two in both form and meaning, in that "cat" refers to one feline entity and has no "-s" suffix, and "cats" refers to more than one feline entity and has an "-s" suffix. Therefore, we can say that the phonological sequence* [kʰæt]—*or "cat"—refers to a particular feline entity, and the suffix "-s" means "plural".*

Modern day linguists have taken this rather basic notion of morphology quite a bit further in recent years[2], but that research, by and large, hasn't informed the practices and habits of modern day conlangers.

In this paper, I shall examine the notion of the morpheme and discuss its impact on modern conlanging. I will then propose an alternative to traditional morpheme-

---

[1] This paper is based on a talk I gave at the First Language Creation Conference in 2006 ("Down with Morphemes!: What Word and Paradigm Morphology Can Teach Us about Language Creation"). While that talk focused a good deal on Bochner's version of Word and Paradigm Morphology, this paper will instead take a closer look at the notion of the morpheme and its role in modern day conlanging.

[2] For two rather sophisticated (though divergent) theoretical approaches, see Ackerman, et al (2009) and Halle and Marantz (1993).

based approaches to language creation, and finish with some sample derivations and evolutions.

1.1   A Note to Non-Artlangers

It's true that, for the most part, the target audience of this paper is those who aim to create languages that are more or less naturalistic (and human). Nevertheless, those who are interested in creating auxlangs or non-human languages may profit by reading this paper. For example, many alien languages focus on semantics and/or grammatical categories while still employing grammatical strategies that are far from alien. Here's an artificial example:

(1) ðəeoikpɫɛʃtɯm-ymfstik ɣʟhʏəjrvzd ɯɯiiil ʍɵvbʏχq-ɑwɢɪtʰajoki.
    /title-DMLR.AND. ɣʟhʏəjrvzd ɯɯiiil
        separate.from.two.boxes-NON-PAST.NON-FUT.NON-PUNC./
    "ðəeoikpɫɛʃtɯmymfstik ɣʟhʏəjrvzd is separating an ɯɯiiil from two boxes."

This is a fake alien language, and it looks pretty strange. It's nearly unpronounceable, it distinguishes ninety different titles (this one is for a Dimlarian android), ɣʟhʏəjrvzd is a pretty crazy name, it has affixes which only tell you what something isn't (non-past, non-future, non-punctual), and one can't even approximate in English what an ɯɯiiil is. Then you have the verb ʍɵvbʏχq, which is a basic verb (as basic as "to eat" in English) which means "to separate something from two boxes". Can't get more alien than that, right?

Well, perhaps one can. The sample above, in fact, is all but equivalent structurally to the following sentence of Turkish:

(2) Bay-an Gül mektup yaz-ıyor.
    /title-FEM[3] Gül letter write-PRES.CONT/
    "Miss Gül is writing a letter."

The number and strangeness of the titles in this alien language, focusing on one example, doesn't make the structure "title base + modifying suffix" any less ordinary. Similarly, natural languages of all stripes have "verb + tense/aspect suffix". Increasing the number of suffixes and varying what they encode doesn't make a language any less human: it just makes encountering its system in the wild increasingly unlikely, and makes the language look, for lack of a better word, artificial.

---

[3] I make no theoretical claims about the status of the *-an* suffix in Turkish. Suffice it to say that *Bay* is the equivalent of English "Mr." and *Bayan* is the equivalent of English "Miss".

When creating an alien language, bearing in mind how human languages work can help to create something that's truly non-human. Additionally, if one is trying to create an ideal language, one needs to know what to avoid, and what actually isn't too much trouble for language users.

2   The Morpheme

The notion of a morpheme can be traced back to Panini, but a good definition for how it is understood today can be found in any elementary linguistics textbook. Here, for example, is how morphemes are introduced in the fourth edition of *Contemporary Linguistics*:

> *The most important component of word structure is the **morpheme**—the smallest unit of language that carries information about meaning or function.*

Thus, the meaning of a word like "bats" can be cut up as follows:

Orthographic Form	Phonetic Form	Meaning
bat	[bæt]	flying mammal
-s	[s]	plural number

Figure 1: Morphemic analysis of English "bats".

Crucial here is the relationship between the singular and plural forms for the flying mammal "bat". Specifically, in a morpheme-based theory, we know that the [s], for example, carries information about the word's meaning because it is the sole difference between the singular and plural form of "bat". No such systematicity can be seen with, for example, the [æ] in "bat". Exchanging that sound for [ɛ] will, indeed, produce a new word, but as there is no identifiable systematic relationship between gambling and flying mammals, there is no evidence for [æ] being a morpheme in "bat", just as there is no evidence for [ɛ] being a morpheme in "bet"[4].

This type of analysis (whether of necessity or not is a matter worthy of discussion) had a profound effect on the understanding of meaning. As morphemes are

---

[4] An interesting chicken and egg argument arises just from examining this account of the relationship between words. Specifically, in order for a difference between two words to be morphological (and in order to determine whether or not a given bit of phonological material is a morpheme), a categorical or paradigmatic distinction must already exist—even though it's the morphemes themselves which presumably create these categories. That is, a singular and plural paradigmatic distinction in nouns must exist in order to prove that the "-s" suffix is a plural morpheme.

the storeholders of meaning, larger meaning is combined in precisely the same way that morphemes themselves are combined. Thus, a word like "bats" has a kind of mathematical spellout as shown below:

bat	+	*-s*	=	*bats*
[bæt]	+	[s]	=	[bæts]
flying mammal	+	plural	=	many flying mammals

Figure 2: The concatenation of "bat" and "-s".

The linear order of [bæt] and [s] in "bats" can be verified rather simply. The linear order of the meanings of "flying mammal" and "plural" are a different matter, but we'll leave that aside for the moment. Let's call this State Zero for morpheme-based analysis. With this understanding in mind, I'll go on to examine various problems encountered by morpheme-based analyses, and what was done to resolve them.

3   Problems with Morpheme-Based Analyses

What follows is by no means either a chronological or an exhaustive account of the various problems morpheme-based analyses have faced and resolved. Instead, what I hope to do is illustrate the types of problems morpheme-based approaches face, and the types of solutions proposed. After discussing several examples in this section, I'll discuss what is gained and what lost by the various resolutions that have been utilized to rescue morpheme-based approaches to grammatical analysis in section 4.

3.1 Allomorphy

To get us started, I'll examine a widespread and fairly recognizable phenomenon: allomorphy. Continuing with our "bat~bats" example, the English plural is complicated slightly when other nouns are added to the list:

(3)
   a.   [bæt] "bat (sg.)" ~ [bæts] "bats (pl.)"
   b.   [dɔɹ] "door (sg.)" ~ [dɔɹz] "doors (pl.)"

c.   [wɪʃ] "wish (sg.)" ~ [wɪʃɨz] "wishes (pl.)"[5]

No longer can one say that [s] is the English plural suffix, as now we see both [z] and [ɨz] also acting as plural suffixes.

Though the suffixes differ, they seem to be related, as [s] and [z] differ only in voicing. Further analysis reveals that [s] occurs only after voiceless non-sibilant consonants, with [z] occurring elsewhere (with a little something extra for words ending in [s], [z], [ʃ] or [ʒ]). And so, borrowing some terminology from phonology, [s], [z] and [ɨz] are said to be allomorphs of an underlying morpheme we'll call /z/[6].

With allomorphy, the linguist's task is clear: to locate and define all the morphemes present in a given language, and then to describe what rules govern the distribution of their surface forms.

3.2 Categorical Alternations

Immediate (but by no means insurmountable) obstacles to morpheme-based analyses of allomorphy arise when one ventures out into the wilds of natural language data. With the regular English plural suffix, one might make an argument that its allomorphs are the simple result of garden variety phonological alternations. Such can't be said of some of the irregular plural suffixes that we see elsewhere. Consider the following:

(4)
   a.   ox ~ oxen
   b.   child ~ children
   c.   schema ~ schemata
   d.   datum ~ data
   e.   corpus ~ corpora
   f.   syllabus ~ syllabi

Taking (4a) as an example, there is no formal way to predict that the plural of "ox" should be "oxen", given regular alternations like "fox~foxes" and "box~boxes". Rather, it appears that the plural of "ox" is something that must be learned by rote and memorized thereafter.

_____

[5] A quick note on transcription. I'm operating under the assumption that English has two reduced vowels: a low and a high one. Others may transcribe reduced vowels differently—and, indeed, may transcribe many other items in this paper differently—but the choices in phonetic transcription don't have any bearing on the thrust of this paper. Suffice it to say that those items which must be distinguished shall be.

[6] Calling the underlying English plural morpheme /z/ is not uncontroversial, but the controversy has no direct bearing on the matter at hand, so I'll leave it aside.

Traditional morpheme-based theories have a way to account for such alternations. Focusing on "oxen", one way to account for the data would be to posit two classes of nouns in English: Class A nouns (those that take some variant of the regular "-s" plural suffix) and Class B nouns (those that form their plural with the "-en" suffix). Once the possibility of classes such as the above is admitted, the problem of what to do with the data in (4) is solved: Each word belongs to its own special class. One might summarize the alternations shown in (4) below:

(5)
    a.   Class B: -ø (sg.)/-en (plu.)
    b.   Class C: -ø (sg.)/-ren (plu.)
    c.   Class D: -ø (sg.)/-ta (plu.)
          etc.

On one level, this is a satisfactory explanation, in that it accounts for the data, and it further codifies the notion that these forms are, indeed, irregular, and they must be learned. I'll put off discussion of further issues arising from an analysis like this until later. For the time being, I'll continue with more English data.

3.3 Word-Internal Alternations

Already, cracks have begun to emerge in the explanation evinced in 3.2. Consider the formulation of Class C in (5) I offered. It fairly characterizes the *orthographic* differences between "child" and "children", but it's missing a crucial distinction all English speakers make. Consider the added phonetic transcriptions of "child" and "children" below, along with a few other relevant singular/plural pairs:

(6)
    a.   [tʃajɫd] "child (sg.)" ~ [tʃɪɫdɹin] "children (pl.)"
    b.   [gus] "goose (sg.)" ~ [gis] "geese (pl.)"
    c.   [maws] "mouse (sg.)" ~ [majs] "mice (pl.)"
    d.   [lif] "leaf (sg.)" ~ [livz] "leaves (pl.)"
    e.   [wʊmin] "woman (sg.)" ~ [wɪmɪn] "women (pl.)"

If one ignores the theoretical notion of "morpheme", the apparatus developed in 3.2 would have no problem accounting for the data in (6). Specifically, Class C would be modified to account for the vowel change, and new classes would be developed from (6b-e). Doing so, however, fails to take seriously the theoretical notion of "morpheme".

To reconcile the English data with a morphemic analysis, one has to expand one's understanding of what a morpheme is, and how it works. Take (6a), "child~children".

It's evident that the plural suffix is "-ren", and that this is where the notion of plurality is housed—that is, "-ren" is the plural morpheme. In addition to being atomic units, though, some have argued that morphemes are *agents*, in a sense[7]. In this case, adding the plural "-ren" morpheme to "child" causes the vowel to change from [aj] to [ɪ].

Another way to rescue a morphemic analysis is to claim that the [ɪ] in "children" is actually *a part* of the plural morpheme. The nomenclature might be a little cumbersome (I imagine something like "-[ɪ]-…-[ɹin]"), but one could, essentially, identify a discontinuous morpheme that acts as both suffix and vowel replacer. Such would help to explain the alternation between "woman" and "women", where the plural morpheme is no longer a suffix, but either an infix, or a "vowel replacer"[8]. And, of course, the same analysis could be extended to "goose~geese" and "mouse~mice".

Datum (6d) is a bit troubling, despite the advances in our theory. We see the usual "-s" plural suffix working as it should (consider [kʰev] "cave" ~ [kʰevz] "caves"), but there's a difference in the voicing of the stem-final consonant between the singular and plural that should be surprising. We can say for certain that it's not a regular phonological alternation (i.e. that the /f/ isn't voicing due to the voiced [z] suffix) because we might otherwise expect a voiced labio-dental fricative in "leaf's" (as in "The leaf's ruddiness"), "leafs" (as in "He leafs through the book") and/or "Leafs" (as in "The Toronto Maple Leafs")[9]. Somehow, then, the facts will have to be accounted for morphologically.

Sticking to our morpheme-based analysis, we can posit a different type of otherwise regular plural suffix. It looks and acts just like the ordinary "-s" plural suffix, but this one causes voicing in the final consonant. Such a suffix would prove quite useful, in that it would also help to explain pairs like [najf] "knife" ~ [najvz] "knives" and even [haws] "house" ~ [hawzɨz] "houses". The mechanism would be roughly akin to the "-ren" suffix which also causes a vowel change.

If that approach is unpalatable, we can borrow the "vowel replacer" idea and apply it to great effect in (6d). To make it work, we simply change the "vowel replacer" into a "feature replacer", such that we have the regular "-s" suffix accompanied by a floating [+voice] feature which replaces the [-voice] feature of a stem-final voiceless

---

[7] In particular, I'm thinking about a rule-based Item and Process approach, but a specific reference doesn't come to mind.

[8] The distinction between these two theoretical notions is absolutely crucial (as will become apparent), even though the idea of a "vowel replacer" is a bit silly.

[9] The argument comes, most famously, from Pinker (1995), who claims that the compound is "Maple Leaf", and that it's the compound that's pluralized, and not the individual word "leaf". As the compound is treated as a regular nominal compound, it takes the regular plural allomorph, giving us "Maple Leafs". One has to wonder, though, why one will quite commonly both hear and see in print "Toronto Maple Leaves" (a quick google search turned up close to 200,000 hits), but one absolutely never hears "saberteeth" as a serious plural of "sabertooth".

fricative of a noun that falls into whichever class "house", "wife", "knife", "leaf" and others like them belong to. Again, the nomenclature will be a bit odd (something like "[+voice]…-[z]"), but nomenclature is merely a representation, after all, of what is *really* happening.

3.4 Zero Allomorphy

So far so good (more or less). We've accounted for all the data and maintained in tact the notion of "morpheme" (though its status grows more and more abstract). Dealing with plurals in English, there is one last bit of evidence which can be summarized as a null alternation:

(7)
    a.   sheep ~ sheep
    b.   fish ~ fish
    c.   deer ~ deer
    d.   bison ~ bison

Very quickly, we can assure ourselves that these are count nouns, and not mass nouns, because sentences like, "I see three deer in the meadow yonder" are perfectly grammatical.

Alternations such as those in (7) would appear to pose a serious challenge to morpheme-based analyses. (Un?)fortunately, such alternations have been accounted for rather simply thanks to the notion of a zero morpheme: a phonologically null suffix which changes the meaning from singular to plural. Thus, the morphological structure of (7a) is as follows:

sheep	+	*-ø*	=	*sheep-ø*
[ʃip]	+	[ ]	=	[ʃip]
bleating, wooled mammal	+	plural	=	many such mammals

Figure 3: The concatenation of "sheep" and "-ø".

I should note that though this null morpheme is conventionally realized as a suffix, its placement, as it is theoretical, is up to interpretation, and may just as well appear as a prefix, circumfix, or anything else one likes.

A further class, then, can be posited whose members all take a phonologically null plural suffix to form their plurals, and all the data in (7) have an explanation. We

were forced to posit a phonologically null suffix, but if one accepts the explanation in 3.3, this new development should prove neither surprising nor troubling.

3.5 Fallout

At this point, we can effectively leave off analyzing natural language data, as pretty much any imaginable morphological phenomenon can be accounted for with the machinery we've constructed. Essentially, the notion of a morpheme can be abstracted to the point where one can take it completely out of the phonological structure of a word and claim that it is responsible for everything going on inside the word. But we needn't go that far to see that the explanation provided up to this point leads to a number of rather drastic theoretical consequences.

To begin with, let us return to (6b): "goose~geese." We've already posited one of two possible explanations: either a zero allomorph of the plural suffix which effects a change in the vowel (turning the morpheme into a function as well as an atomic entity), or a "vowel replacer" morpheme which changes "-oo-" to "-ee-". There is another possibility I didn't mention which should help to elucidate one of the largest problems morpheme-based accounts of language face—namely, that "-ee-" or "-[i]-" is an infix—an *actual* infix.

If we admit, for the moment, that "-ee-" is an infix (a variant of the plural allomorph), then it would mean that the stem to which it applies is "gse" (phonetically, [gs]). It might seem a bit strange to native English speakers, but it would solve a few problems. A new question arises, then: What is the status of "-oo-" or "-[u]-"? Rightly or wrongly, we are forced to accept that "-oo-" is a *singular* infix.

Now we have a "problem" that actually clarifies a number of the ambiguities present in the preceding argument. First, we knew already that English had a "singular" category just as it had a "plural" category. As this is the case, how does one account for the concatenation of meaning in a word like "dogs"? If "dog" means "noun, singular", then "dogs" *must* mean "dog, singular, plural"—that is, a mammalian quadruped that is both singular in number and non-singular in number *at the same time*.

There is, of course, an easy fix to this problem. We've already posited a null plural allomorph. What, then, is wrong with a null *singular* allomorph? We could argue that the null singular is the regular (Class A) singular morpheme, and that others (such as "-oo-") are irregular. Doing so would allow us to account for all the data in a simple and straightforward fashion. Below is a sample of that analysis:

Noun Class	Sample Lexeme	Singular Affix	Plural Affix
Class A	[bæt] "bat"	-[ ]	-[s]
Class B	[ɑks] "ox"	-[ ]	-[in]
Class C	[tʃɫd] "child"	-[aj]-	-[ɪ]- -[ɹin]
Class D	[skimə] "schema"	-[ ]	-[tə]
Class E	[deɾa] "data"	-[m]	-[ ]
Class F	[gs] "goose"	-[u]-	-[i]-
Class G	[ms] "mouse"	-[aw]-	-[aj]-
Class H	[kɔɹp] "corpus"	-[is]	-[ɚɹə]
Class I	[lif] "leaf"	-[ ]	-[+voice]-…-[z]
Class J	[wmn] "woman"	-[ʊ]-…-[ɨ]-	-[ɪ]-…-[ɪ]-
Class K	[ʃip] "sheep"	-[ ]	-[ ]

Figure 4: Morphemic analyses of some of the data presented.

I've made a number of theoretical decisions in the table above (e.g. analyzing "data" as a word with a null plural suffix and a singular -[m] suffix) that are largely unimportant, as reformulating them would produce the same result: a lexeme column, a singular column and a plural column each populated by a bunch of morphemes. What they are makes little theoretical difference.

And this, of course, arises from a simple examination of English pluralization strategies. If one takes the entirety of the English language into account—let alone the languages of the world—the picture gets even muddier.

4   Language Creation and Linguistics

At this point, it's my hope that the reader will grasp his her or head in disgust and cry out, "How unbelievably absurd! Why have I been taking morpheme-based theories of grammatical description seriously all these years?!"[10], but I recognize that that will not be the reaction of most. Indeed, the most powerful thing about a framework is that it can withstand pretty much any assault—sometimes by its proponents simply ignoring

---

[10] Or, even better, "Gee, that's just what I've been thinking all these years!"

any and all counterarguments[11]. There will always be a way to rescue any framework, including morpheme-based analyses of language.

At this point, what I want to do is demonstrate what effect morpheme-based theories have on conlanging, as that question is of central importance to this paper. Before leaving linguistics behind, though, I have a few brief comments and observations to make.

4.1 What Is Lost

The state of affairs in figure 4 above may look bleak to some, but to others the analyses may seem perfectly viable. To others still, the analyses will look like they could use a bit of pruning and polishing—that they're in need of just a bit of tweaking to bring them in line with a modern morpheme-based account of English plural patterns. Unfortunately, no matter how much these analyses are spruced up, the most troubling problems will remain.

Consider the problem of allomorphy. In English, there's a simple bit of allomorphy with the indefinite article "a/an". Basically, you get "a" before words that begin with a consonant phone and "an" before words that begin with a vowel phone, as below:

(8)
    a. a sheep / *an sheep
    b. a conlanger / *an conlanger
    c. an owl / *a owl
    d. an omen / *a omen

The environment is simple, as is the rule, but the form of the indefinite article is not, necessarily. That is, though it might be perceived as being simpler to have a consonant in between two vowels to break up hiatus, there's no special reason that that consonant is [n]. It could have been anything, or a different word entirely. The form, however, just happens to be "an". Stating that "a" and "an" are allomorphs of the same morpheme captures this distinction nicely: They are two morphologically related forms which are *not* related phonologically, even though the allomorphy is phonologically conditioned.

The problem with a morpheme-based analysis is that *all* morphemes bear the same relationship. That is, there is no reason to expect that two morphemes should look anything alike in any circumstance.

---

[11] Cf. Postal (2009) on Noam Chomsky, for example.

Consider the plural morpheme in "leaves" and the plural morpheme in "greaves". According to the analysis presented in figure 4, it's an accident that these morphemes look and sound alike. As the morpheme in "leaves" participates in a different type of alternation (associated also with a voice alternation in the stem), it's a different morpheme (or allomorph), and, as such, may have taken on any form. That is, we might have "greave~greaves" and "leaf~leavork". Formally, the morpheme-based approach fails to capture the simple generalization that the suffix is identical.

In a similar vein, the theoretical notion of a morpheme has nothing to say about seemingly bizarre situations such as that exemplified in the data below:

	"to sip" (regular)	"to hit" (irregular)	"to shit" (archaic)	"to shit" (modern)
Present	*sip*	*hit*	*shit*	*shit*
Past	*sipped*	*hit*	*shat*	*shit*

Figure 5: Present and past forms of "to sip", "to hit" and "to shit".

In English, certain uncommon irregular verbs have gradually fallen into disuse, and their accompanying irregular past tense forms have been forgotten (consider how infrequently English speakers use "throve" as the past tense of "thrive"). The archaic past tense of "to shit" is still around, but it's considered a bit archaic or fanciful (or comical). The more common past tense form is identical to the present tense form. Morphemically, one would have to analyze the above thus:

	"to sip" (regular)	"to hit" (irregular)	"to shit" (archaic)	"to shit" (modern)
Present	sip + -ø	hit + -ø	shit + -ø	shit + -ø
Past	sip + -ed	hit + -ø^x	shit + -ø^y	shit + -ø^x

Figure 6: Morphemic analysis of the forms in Figure 5.

The analysis in Figure 6 is, at best, purely descriptive: It's a simple statement of facts. However, doesn't it seem odd that the former irregular verb "to shit" lost its former irregular pattern and gained *another* irregular pattern? Why isn't "shitted" the common past tense? A morphemic analysis has nothing to say about this—nor anything to say about the fact that the past tense form "shat" is still available to English speakers.

Perhaps more troubling is the status of morphemes themselves. By definition, all morphemes are equal. Morphemes are the "building blocks" of language, and though they differ in distribution (free vs. bound), the "atomic weight" of each morpheme, so to speak, is equal. If one takes the notion of "morpheme" seriously, then the plural morpheme "-s", the third person singular agreement morpheme "-s", and the word "adroit" should all take up the same space in one's head. As all are morphemes, the brain should make no distinction between one or the other, and they should be treated roughly the same.

The notion seems to run counter to the basic facts of language. Language users interact with the morphology of a language—the collection of so-called derivational and inflectional morphemes—quite differently from the lexicon—i.e. (in this instance) the collection of lexical items that comprise one's vocabulary. As a quick example, language users will frequently forget words, yet it would be quite odd for a speaker to say something like, "Yeah, I saw Tommy. He's eat…eat…dang it, what's the suffix? Oh yeah: -ing. He's eating in the backyard." There may be other explanations for this phenomenon, but the fact remains that morphology and lexical items are treated quite differently by language users. Conflating the two seems to be setting up a kind of psychological fantasy that treats human brains much like computers: machines that can amass data, but can't do anything with it unless told specifically how to evaluate it. To the extent that one ought to be concerned about the psychological reality of a given theory, the foundation of morpheme-based approaches should, at the very least, be reconsidered.

4.2 The Proper Place for Linguistics

It has been discussed at length at various times on various online listservs and fora just what the relationship between linguistics and language creation is, and what it ought to be. Certainly, many conlangers have had disheartening experiences within academia, and discouraging encounters with academics—both within linguistics and without. And though there have been strong opinions about what role language creation should play—if any—in linguistics both within the community and within academia, time appears to be resolving the matter for us. I personally have had wonderful experiences in linguistics, and received a lot of encouragement from a number of linguists—several of whom are conlangers themselves. There may still exist some antagonism, but I believe the instances are growing fewer in number, and that the situation will improve in the future regardless of any action taken by any party in the future.

Much less attention has been paid to what role linguistics and linguistic knowledge should play in language creation. The online community of conlangers consists of a large number of young people who either are taking linguistics courses or have in the past, along with several professional linguists or former linguists, and

countless self-taught amateur linguists. A kind of knowledge separate from academic linguistics altogether has grown and expanded with the online community, and can be considered the collective common knowledge of the online conlanging community. This knowledge, shared piecemeal in forum postings, blog posts, private and public e-mail, and even on Twitter and Facebook status updates, has helped to enrich the community and has advanced the art of language creation considerably in the past decade or so.

The extent to which this collective knowledge overlaps with general academic linguistic knowledge is uncertain, but the mapping surely isn't perfect[12]. The inconsistency may be a result of the method of delivery and the consumption of the knowledge that comes to the community from academic linguistics. As nearly as I can tell, there are three major sources of linguistic knowledge:

1. Linguistics students and professors.
2. Published papers and books.
3. Online sources (usually second-hand, at best).

From these sources, generally conlangers make use of the following: reported structures from natural languages; linguistic universals; methods of analysis; and various "rules" derived from specific analyses. Here I would like to focus on the last two.

Ignoring errors of transmission, there is a grave danger in making use of linguistic frameworks used for analysis and the results of such analysis in constructing a language. The job of a theoretical linguist is clear: To analyze language data. That is, a linguist starts with data, and ends up with an analysis. Many (if not most) conlangers have implicitly defined the job of a conlanger as being the *exact opposite* of a linguist's— that is, to begin with an analysis, and to work backwards to create data.

There are a number of very serious problems with the "Reverse Linguistics" approach to conlanging. I'll briefly examine three of them.

First, linguists make mistakes, and frameworks change—as do languages, for that matter. Considering languages of the Philippine type (Tagalog, Chamorro, Ilokano, Cebuano, etc.), I've read Relational Grammar and Minimalist Program analyses of the same language, and the two are night and day. Often different linguists from different traditions will disagree on the basic facts of a given language, and their frameworks differ so greatly that they can't even have a conversation about the data. It's more than likely that a paper written now on a given language will be inaccurate (or perhaps "inaccurate") in no more than ten years' time.

_____

[12] Consider the Conlang Trigger System: An entirely novel system based on a misunderstanding of an old analysis of languages like Tagalog. The system is one that doesn't exist in any natural language, but it's a workable system that's been put to good use by a number of conlangers (myself included).

Second, one of the driving assumptions behind much of the work of modern linguists is that all languages are the same (this, in particular, is a problem within the generative tradition). As such, there's an external pressure amongst many theoretical linguists to make languages "fit". Challenging a generally accepted universal is a serious thing. It's more felicitous to discover that a construction that looks completely bizarre and other-worldly actually works just like something else that has already been explained in another paper written in the same framework. And, unsurprisingly, this happens quite a bit—which is used as evidence that the approach was correct to begin with, and the vicious circle continues. The fault, though, lies not with linguists, by any means. The problem is that many of the frameworks developed by linguists are indestructible. A simple Item-and-Arrangement morphemic framework can be used to analyze *any conceivable* language; there is literally no phenomenon it can't handle. Whether these analyses are worthwhile is usually a question left for later.

Finally, there is a pressure amongst many phonologists, morphologists and syntacticians to ignore the history of a language. The reasoning is sound: Children are born into this world without any knowledge of the history of any language on Earth—and yet, they all learn to speak. This means that they work only with the data presented them—the synchronic state of the language. As such, the history of a given language *must* be ignored by linguists trying to analyze the synchronic phenomena present in that language. Whether one agrees or disagrees with this theoretical assumption as it applies to linguistics, I would argue that one *cannot* ignore the history of a constructed language. Linguists try to come up with an analysis for how a child will make sense of a given system; a conlanger is trying to *create* that system—and, perhaps, the series of linguistic events that gave birth to that system. The two are entirely different enterprises, and require different methodologies.

Ultimately, I would argue that within the realm of language construction, linguistic knowledge should be *subordinate* to one's own sense of how to construct a language. As it is today, often linguistic knowledge is taken as a kind of law to be obeyed. Realistically, the accumulation of linguistic knowledge is simply not sophisticated enough to act as a measure of any kind for a constructed language. The reason is simple: Linguists and conlangers are working at cross purposes. It seems unlikely that there will ever be a time that linguistic knowledge can be used in the way conlangers presently would like to use it simply because the answers linguists seek are not of immediate use to a conlanger seeking to create a naturalistic language. In fact, using "Reverse Linguistics" will likely lead one to creating a rather clunky and unnatural conlang, as we will see in the next section.

5   Language Evolution

The problem with linguistic universals is that they're notoriously malleable. They prescribe precisely what natural languages can and can't do—until a natural language does something universals say it oughtn't, in which case the list of universals is emended (or the language data is reanalyzed). In other words, linguistic universals allow linguists to predict what natural languages can and can't do—except for the stuff they can't predict. For one attempting to construct a naturalistic language, this information is not terribly useful.

All natural languages do have one thing in common, though: They were all created by humans. That fact is by no means trivial. The human brain has been pretty much the same for centuries. More importantly, the way in which a human community will interact with and manipulate data has also remained static. Examining language data, then, is fine, but what we see are the *results* of humans' interaction with and manipulation of language. By simply examining the results, a conlanger can only produce variations thereon. What the naturalistic conlanger *needs* to do is emulate the method whereby human communities *evolve* language.

In this section, I'll illustrate how Reverse Linguistics can lead to unnatural results, and discuss language evolution.

5.1 Problems with Reverse Linguistics

Below are some sample nominative and accusative forms in four different languages:

	Latin	Russian	Turkish	Spanish
Nominative	*eqvvs*	*kniga*	*kitap*	*gato*
Accusative	*eqvom*	*knigu*	*kitabı*	*gato*

Figure 7: The nominative and accusative forms of Latin "horse", Russian "book", Turkish "book" and Spanish "cat".

A strict morphemic analysis would render each nominative and each accusative form identical: STEM + -NOM.SG for the nominatives and STEM + -ACC.SG for the accusatives. The simplest way to create a new system based on that analysis is to create a bunch of nouns and a bunch of suffixes. And by and large, many purportedly naturalistic conlangs end up as, essentially, long lists of stems and affixes.

The problem here isn't necessarily with the analysis: The problem is with how that analysis is used. A simple examination of the languages in question will reveal that

the systems are far, far more complex than the analysis above would lead one to believe. In Latin, there are different declension classes and countless exceptions. In Russian, the genitive plural of *kniga* is *knig*. In Turkish, the accusative is used only with definite direct objects. In Spanish, certain direct objects actually take a kind of special case marker depending on whether the object is animate and specific. These details aren't revealed with a simple morphemic analysis, but when conlangers start with that kind of an analysis in mind, they often are unable to effectively (or realistically) reproduce what is left out by the analysis. In effect, many conlangs are actually phonological realizations of morphemic linguistic analyses.

5.2 Starting Over

To begin a naturalistic conlang is to face a great mystery—one that may never be fully understood: The origin of language. As far back as we can go, there has always been language—and the further back we go, the spottier the records get. No one knows how language emerged from nothing, or what it looked like at its earliest stages. Even our best opportunities to study this process—modern pidginization and the creation of signed languages—have been lost, as linguists are usually alerted to the presence of a new pidgin or sign languages some fifty or so years after the fact—if they're lucky[13].

    As such, a naturalistic conlanger has to make an executive decision: Where to start, and what to start with. It would be mere speculation to even *guess* what a brand new language will look like ten years after its inception—especially a language created *ex nihilo*. It's up to the conlanger to decide where the language will begin, what system it will have, what lexemes, and so forth, and it's at this point that the creator will have to put their foot down and say, "That's moment zero." That is, before that point, the language has no history; its lexemes are unanalyzable; its systems came from nowhere. As unsatisfying as that may be, there's simply no other alternative (aside from those supplied by imagination[14]).

    Once a conlanger determines the initial state of the language, its evolution can begin.

---

[13] It's worth noting, I feel, that this is, in large part, due to prejudice. Most pidgin languages are regarded by both their speakers and their overhearers as ungrammatical, unimportant—even shameful—and not "real" languages. The same can be said of many sign languages—even well-established ones. We can only hope that humanity has learned its lesson with respect to these wonders of innovation, and that the next time a language is born, linguists will be there at ground zero as events unfold.

[14] For example, a good number of artlangs have imagined a kind of being or society that created a language (perhaps a "logical" or "ideal" language), and that this construction became the proto-form of the modern language.

5.3 Some Examples

The languages we have today are the result of the tension between two opposing forces. Whatever they're called—innovation and conservatism; dynamism and stasis; faithfulness and markedness—their opposition fuels language growth and language change. In order to create a naturalistic conlang, this tension must be replicated—or fabricated, at least.

Take English irregular plurals as an example. By looking at the endpoint, a conlanger trying to replicate what they see in English might create something like the following:

(9)
    a. mate "bear" ~ matek "bears"
    b. virt "sword" ~ virtak "swords"
    c. borl "mouse" ~ berl "mice"
    d. wug "pond" ~ wugop "ponds"
    e. romp "otter" ~ ramp "otters"
    f. toli "card" ~ tuliar "cards"

And a nice description might accompany this, such as, "The regular plural suffix is [-k] after vowels and [-ak] after consonants. There are a number of irregular classes, too, which you just have to memorize." That latter bit may be true of an English speaker, to some extent, but from the point of view of the language, the irregular classes we have are not random: They were arrived at systematically. Taking two lexemes completely out of context may make the system look strange (say "house~houses" vs. "mouse~mice"), but looking at the system as a whole, one can see, even without knowing the history of the language, that the irregularity is principled. In fact, broadly speaking, principled irregularity is the only kind of irregularity there is, and it can't be faked.

Consider how we arrived at "house~houses" and "mouse~mice". There was a stage of English where the alternation in "mouse" was regular. Through a series of regular sound changes, shown below, it came to be irregular in the modern form of the language (note: GVS below stands for "Great Vowel Shift"):

(10)

Initial State:	muːs	~	muːsiz
Loss of [-z]:	muːs	~	muːsi
Umlaut:	muːs	~	myːsi
Loss of [-i]:	muːs	~	myːs
Unrounding:	muːs	~	miːs

GVS: maws ~   majs

Though the type of alternation in the "mouse" paradigm is rare (seen elsewhere only in "louse~lice"), the sounds themselves are familiar enough; English is chock-full of "-ouse" and "-ice" words. Trying to engineer something similar by going backwards, one is much more likely to end up with random results (like that dreadful [-op] suffix in 9a).

Being able to manipulate sound changes is just part of the problem when it comes to naturalism, though. One might wonder why the series of changes that gave us "mouse~mice" didn't apply more generally, and why, for example, the plural of "house" isn't "hice". In fact, there's a famous poem by an anonymous author that bemoans the state of English plurals, a section of which I'll reproduce below:

> *If I speak of a foot, and you show me two feet,*
> *And I give you a book, would a pair be a beek?*

I've reproduced this section in particular because the sound changes above *did* apply to "book", giving us a plural that would now be spelled "beek". Why do we have "books" now? An accident of history, really. Sometimes irregularities simply don't take. With verbs in English, by and large, it's the most common irregular verbs that stay irregular, the less common eventually becoming regular, but it's hard to argue that "book" is less common than "mouse"—or "louse" or "ox", for that matter[15]. More to the point, changes—be they phonological, semantic or pragmatic—are often capricious in their application. It's the conlanger's job to figure out not only what changes applied and how they operated, but when they applied, what portion of the vocabulary was affected, what portion of the vocabulary didn't yet exist, and in which words the irregularities failed to survive.

A similar problem arises from what I call morphemic pressure. In a standard verb paradigm (let's say present tense indicative, two numbers, three persons), morpheme-based thinking tends to lead to a pressure to produce either six affixes—one for each cell —or four affixes—one for each person, and one for plural. Consider, though, a modern French paradigm (phonetic forms, not orthographic):

---

[15] Well, nowadays, at least. Perhaps the answer lies in books not being *as* common with the majority of those who used English back when the irregular patterns were forming…

Person/Number	Singular	Plural
First Person	desid	desidõ
Second Person	desid	deside
Third Person	desid	desid

Figure 8: The phonetic forms of the present tense conjugation paradigm for the French verb *décider*[16].

Synchronically, and without the rest of the language in mind, the paradigm looks exceedingly strange. Yet it is this type of strangeness that is natural. Replicating this, though, should not entail, say, taking four out of six cells of a paradigm like this and deciding that the forms will be identical, giving a unique suffix to the remaining two forms. The key lies in the development of the forms and the system.

5.4   Instituting Change

The first step towards producing a naturalistic language on par with the natural languages we see in the world is the complete and total abandonment of the morpheme. It is not useful, and, in fact, does more harm than good. Do not misunderstand me to say that this means one should abandon *affixes*; by no means. All I mean is don't confuse the fact that an [-s] appears with a large number of plural nouns to mean that the [-s] *means* "plural" in any real or useful sense.

This first step is important, because it will help to simplify the process of evolution to follow. Consider a language with no verbal morphology (presumably a very old language) to which one hopes to add inflections for person and number. Where are they going to come from? If they're going to be suffixes, are the suffixes going to appear out of nowhere?

As it turns out, a number of languages develop person morphology from personal pronouns (sometimes possessive, sometimes not), and at the initial stages, they simply *are* pronouns—and are often identical in form to, for example, possessive pronouns. In Middle Egyptian, for example, the exact same suffix for first person possession ([-i] or 𓇌)[17] was used for a first person subject on verbs. At that stage (or for

---

[16] The orthographic forms (going from first person to third and from singular to plural) are as follows: *décide, décides, décide, décidons, décidez, décident*.

[17] There is a standard convention in Egyptology which I have not followed here. The transcription of glyphs is often given in a transcription system unique to Egyptology which employs letters and letter-like symbols rather than IPA. The standard transcription of the double reed is a "j" with a c over it instead of a dot.

some time during the history of Middle Egyptian) the two were identical, because they were, in fact, the same element.

The key point here is that the forms aren't fixed: neither the meanings nor the phonology nor the parts of speech. All of these facets of a lexeme are open to interpretation by speakers. Sometimes this can result in a change in usage, or a change in meaning, a change in pronunciation—and sometimes a lexeme can lose its identity altogether. A linguistic system is susceptible to constant reevaluation by its speakers, and speakers of a given language, like it or not, simply don't recognize morphemes. Consider the wild reanalysis by English speakers shown below:

(11)
    a. laughing > laughin'
    b. running > runnin'
    c. skipping > skippin'
    d. jumping > jumpin'
    e. thumping > thumpin'

This is a fairly well-documented and well-understood change that happened probably over a century ago in English. If the reader is unfamiliar with the orthographic form, the change here is from a velar nasal [ŋ] to an alveolar nasal [n], and the perception is that the latter is somehow "less formal" than the former. This change, though, went beyond its original domain of application into other realms:

(12)
    a. nothing > nothin'
    b. something > somethin'

Despite the fact that "nothing" and "something" don't actually have the "-ing" suffix, the rule still applies quite comfortably. It's not a rule that applies generally (e.g. one can't say "thin'" for "thing" or "flin'" for "fling"—I doubt even if one can even say "weddin'" as a noun [at least, not in all dialects]), and, furthermore, doesn't even apply to all compounds with "thing":

(13)
    a. everything > *everythin'
    b. anything > *anythin'

One might try to capture the generalization above by stating that only disyllabic compounds with "thing" can be reduced, but then there *are* certain dialects that allow both "everythin'" and "anythin'". Why should it be acceptable in one and not in another?

21

What rule can account for such a thing? And what does one call the suffix/non-suffix whose final [ŋ] can change to [n]? It's certainly not a morpheme—in (12) and (13), it's not even a suffix[18].

The nice thing, from the point of view of a conlanger, is one doesn't have to call it anything at all—and, in fact, one is a better off not doing so. A native speaker is often blind to grammatical descriptions, and, as such, will find patterns where there are none —or, perhaps ought not to be. And this is how we get reduced forms like "nothin'" and "somethin'" in English.

5.5 Expansion and Evaluation

Evolving a language is a kind of cyclical two step process that takes place on many linguistic planes at the same time (phonological, morphological, syntactic, semantic, pragmatic…). The two steps one can call expansion and evaluation.

The definitions of "expansion" and "evaluation" are a bit nebulous, because, as of yet, the specifics of linguistic evolution are poorly understood, at best. We know, though, that these stages have to exist based on what we have seen and what documentation we have of older stages of the world's languages.

What I'm calling "expansion" is a period of change in the state of a language. How long this period is, and what it encompasses, I'm not sure anyone can say. But, for example, we know that at one point in time, English didn't use "go" to form any kind of future tense. Then at some point in time, "I go to eat" was reinterpreted as "I will eat", and soon it became "I'm going to eat", and then "I'm gonna eat", and then "I'm'a eat", and then "I'm'a go eat". I'd say an English speaker currently still has access to the original meaning of the phrase, though, so we're not yet at the point of no return—i.e. we still have access to the literal meaning of "I'm going to eat". Expansion can affect the phonological system in the form of sound changes, which, in turn, can affect the morphological system. It can affect the syntax (consider that Latin was SOV, but that Spanish is now SVO, and, in fact, moving towards VSO), the semantics (the original meaning of "silly" was "holy", etc.), and the various periods of expansion can happen simultaneously or independently of each other.

The other side of the coin is what I call "evaluation". At some point in time, a language needs to settle down and allow its speakers to take stock of just what is a part of their linguistic system, and what isn't. For example, at one point in time English had a fairly stable intervocalic voicing rule. If we borrow a new word into the language that ends with [f], though, that [f] isn't going to become [v] if we add a V-initial suffix anymore. The period of expansion that gave rise to intervocalic voicing came to an end,

---

[18] Though it might be amusing to imagine it is. If I'm everything right now, then yesterday I everythed, and my friend habitually everyths through my things.

and the period of stability that followed was left behind by new periods of expansion. It's that in between period of evaluation that defines the language with respect to a certain feature for a period of time. At some point a given expansion is accepted into the system and the system remains stable until a new period of expansion comes along and the older expansion is fossilized.

A good way to characterize it might be to consider the rings on a tree. When one examines a tree stump, one can see concentric circle in the tree radiating out from its center. The rings are visible precisely because of these periods of expansion separated by periods of stability (or vice versa, if you prefer). Furthermore, if one is standing next to a tree at any given time, it's nearly impossible to tell if it's growing or not. It looks stable enough, after all. The same can be said of language. Considering one's own language, it's nearly impossible to tell if it's changing—or how, or where it's headed. Languages outlive us, and grow more slowly than we do. The best we can do is appreciate the changes that we've been able to document, and try to simulate (at high speed) the process ourselves by conlanging.

6   Conclusion

The approach I'm advocating is not new. Indeed, the first modern artlanger, Tolkien, approached conlanging from a diachronic perspective. Today, though, we enjoy an advantage Tolkien didn't have—one he remarked upon himself, in fact[19]. Specifically, we have each other; we have a community: We have an audience. And if we are to engage in conlanging as an art, then we have to move it forward.

It is on us to both define what rigor in naturalistic conlanging is, and to apply it, if not to each other's work, then at least to our own creations. In the past, it was such a tremendous feat to admit to another that one engaged in language creation that the notion of criticism was anathema. Since conlangers came together and began to form the rudiments of what now is a robust online community in 1991, we've come a long way, and I think it's time to move beyond our initial state of trepidation. The vice is secret no longer.

This paper is the first step towards a definition of naturalism in conlanging[20]. It is an attempt to take stock of what we've learned as a community over the years and to give us the language we need to articulate what it is that we do and why it's special to the uninitiated. If we don't take our work seriously, no one else will, and if we can't

---

[19] "You must remember that these things were constructed deliberately to be personal, and give private satisfaction—not for scientific experiment, nor yet in expectation of any audience. A consequent weakness is therefore their tendency, too free as they were from cold exterior criticism, to be 'over-pretty', to be *phonetically and semantically* sentimental—while their bare meaning is probably trivial, not full of red blood or the heat of the world such as critics demand." (From "A Secret Vice.")

[20] Note the article: *a* definition of naturalism in conlanging.

adequately explain why the elements of one conlang are superior to another, then, in the eyes of the uninitiated who are just now beginning to look in on the world of language creation, all conlangs will be equal—or will be defined by characteristics which they themselves supply. In other words, if we don't have a way to define rigor in conlanging, very soon a definition will be thrust upon us, whether we like it or not.

7   References

Ackerman, Farrell, Gregory T. Stump, and Jim Blevins (editors). 2009. *Paradigms and Pariphrasis*. Stanford: Center for the Study of Language and Information.

Bell, Sarah J. 1983. "Advancements and Ascensions in Cebuano." *Studies in Relational Grammar*, ed. David M. Perlmutter. Chicago: University of Chicago Press: 143-218.

Bochner, Harry. 1993. *Simplicity in Generative Morphology*. New York: Mouton de Gruyter.

Bybee, Joan, Revere Perkins and William Pagliuca. 1994. *The Evolution of Grammar: Tense, Aspect, and Modality in the Languages of the World*. Chicago: University of Chicago Press.

Campbell, Lyle. 1998. *Historical Linguistics: An Introduction*. Cambridge, MA: MIT Press.

Halle, Morris and Alec Marantz. 1993. "Distributed Morphology and the Pieces of Inflection." *The View from Building 20*, ed. Kenneth Hale and S. Jay Keyser. Cambridge, MA: MIT Press: 111-176.

Lakoff, George and Mark Johnson. 1980. *Metaphors We Live By*. Chicago: University of Chicago Press.

O'Grady, William, John Archibald, Mark Aronoff, and Janie Rees-Miller. 2001. *Contemporary Linguistics*. 4th edition. Boston: Bedford/St. Martin's.

Perlmutter, David M. and Paul M. Postal. 1983. "Towards a Universal Characterization of Passivization." *Studies in Relational Grammar*, ed. David M. Perlmutter. Chicago: University of Chicago Press: 3-29.

Pinker, Steven. 1995. *The Language Instinct*. New York: HarperPerennial.

Postal, Paul M. 2009. "The Incoherence of Chomsky's 'Biolinguistic' Ontology." *Biolinguistics* 3.1 (*url* http://www.biolinguistics.eu/): 104-123.

Tolkien, J.R.R. 1983. *The Monsters and the Critics and Other Essays*. New York: HarperCollins.