Fiat Lingua

Title: Designing an Artificial Language: Vocabulary Design

Author: Rick Morneau

MS Date: 07-29-1994

FL Date: 06-01-2021

FL Number: FL-000075-00

Citation: Morneau, Rick. 1994. "Designing an Artificial Language: Vocabulary Design" FL-000075-00, *Fiat Lingua*, <http://fiatlingua.org>. Web. 01 June 2021.

Fiat Lingua is produced and maintained by the Language Creation Society (LCS). For more information about the LCS, visit http://www.conlang.org/

Designing an Artificial Language:

Vocabulary Design

by Rick Morneau

February, 1994
Revised July 29, 1994

[The following is a heavily edited compilation of several articles I posted to the Conlang email list in February, 1994. The Conlang mailing list is dedicated to the discussion of the construction of artificial languages. To subscribe, send an email message with the single line:

SUBSCRIBE CONLANG *your name*
to LISTSERV@BROWNVM.BROWN.EDU. Many thanks to Rick Harrison for kicking off the discussion and for his constructive criticism. I have also written a much longer monograph on the subject of Lexical Semantics for artificial languages that covers all of what follows in much greater detail, along with many other topics related to word design.]

Rick Harrison again provides some interesting food for thought, this time in an area that is closely related to one of my favorite topics - lexical semantics. Here are my comments on some of the points he raised:

Concerning vocabulary size, compounding and derivation:

First, we should make a distinction between WORDS and ROOT MORPHEMES. If compounding and/or derivation is allowed, as is true with every language I'm familiar with, then vocabulary size can be essentially infinite. Even in a system with little or no derivation (such as Chinese and Vietnamese), you can create zillions of words from compounding, even though the number of root morphemes is limited. The problem here, though, as Rick pointed out, is that you often have to metaphorically or idiomatically stretch the meanings of the component morphemes

to achieve the desired result. How, for example, should we analyze English compounds such as "blueprint", "cathouse", "skyscraper" and "billboard"?

Another problem surfaces if you want your compounds to be semantically precise. (By "precise" I mean "as precise as the inherent precision of the basic components will allow".) This will often mean that additional morphemes must be added to a word to indicate how the component morphemes relate to each other. For example, what is the relationship between "house" and "boat" in the word "houseboat"? What is the relationship between "house" and "maid" in the word "housemaid"? Obviously, the relationships are different.

Some languages juxtapose complete words, but keep them separate, as is almost always done in Indonesian, and often done in English. Some English examples are "stock exchange", "money order" and "pony express". However, there is still ambiguity about the relationships between the words. To remove these ambiguities, you will need additional morphemes, which could take the form of linking morphemes such as English prepositions. Swahili uses this approach for all of its compounds, and French uses it for most (French examples: "salle a manger", "eau de toilette", "film en couleurs", etc. Note, though, that the French prepositions are very vague and their use is often idiosyncratic.) If you wish to use this approach, though, make sure that you have enough linking morphemes to deal with all possible semantic distinctions.

Unfortunately, if you don't have a very large and expandable set of root morphemes, you'll definitely run into trouble if your goal is semantic precision. Personally, I don't like artificial languages (henceforth ALs) that limit the number of possible root morphemes - you never know what you're going to run into in the future. An AL should not only give itself lots of room for expansion, but it should make it as easy as possible to implement.

Another thing that should be considered is how easy it will be to learn the vocabulary. This can be best achieved by limiting the number of root morphemes. But if we limit the number of root morphemes, we run into the problems mentioned above!

Actually, there is a solution to this problem. You must design your vocabulary in two steps, as follows:

First, your AL must have a powerful classificational and derivational morphology for verbs. (Other state words, such as adjectives and adverbs, will be directly derived from these verbs.) This morphology will be semantically precise.

Second, root morphemes should be RE-USED with unrelated NOUN classifiers in ways that are mnemonic rather than semantically precise. I.e., the noun classifiers themselves will be

semantically precise, but the root morphemes used with them (and which will be borrowed from verbs) will be mnemonic rather than semantic.

Let me clarify the first step somewhat:

1. Design a derivational morphology for your AL that that is as productive as you can possibly make it. This will almost certainly require that you mark words for part-of-speech, mark nouns for class, and mark verbs for argument structure (i.e., valency and case requirements) and grammatical voice.

2. Start with a common verb (or adjective) and decompose it into its component concepts using the above system. For example, the verb "to know" has a valency of two, the subject is a semantic patient and the object is a semantic theme. (The theme provides a focus for the state "knowledgeable". Unfocused, the state "knowledgeable" would be closer in meaning to the English words "intelligent" or "smart".)

3. The root morpheme meaning "knowledgeable/intelligent" can now undergo all the morphological derivations that are available for verbs. Some of these derivations will not have counterparts in your natural language. Many others will. For example, this SINGLE root morpheme could undergo derivation to produce the following English words: "know", "intelligent", "teach", "study", "learn", "review", "instruct", plus words derived from these words, such as "student", "intelligence", "education", etc. You will also be able to derive words to represent concepts for which English requires metaphor or periphrasis, such as "to broaden one's mind", "to keep up-to-date", etc. It is important to emphasize that ALL of these words can be derived from a SINGLE root morpheme.

In other words, use a back door approach - start with a powerful derivational system, and iteratively decompose words from a natural language and apply all derivations to the resulting root morphemes. In doing so, many additional useful words will be automatically created, making it unnecessary to decompose a large fraction of the remaining natural language vocabulary.

Now, let me clarify the second step:

Root morphemes that were used to create verbs can then be re-used with unrelated NOUN classificational morphemes in a way that is semantically IMPRECISE, intentionally, but which is mnemonically useful. For example, a single root morpheme would be used to create the verbs "see", "look at", "notice", etc. by attaching it to appropriate classificational affixes for verbs. These derivations would be semantically precise. The SAME root morpheme can then be used to

create nouns such as "diamond" (natural substance classifier), "glass" (man-made substance classifier), "window" (man-made artifact classifier), "eye" (body-part classifier), "light" (energy classifier), and so forth.

Thus, verb derivation will be semantically precise. Noun derivation, however, cannot be semantically precise without incredible complication. (Try to derive words for "window" or "hyena" from basic primitives in a manner that is semantically precise. It CAN be done, but the result will be unacceptably long.) So, why not re-use the verb roots (which define states and actions) with noun classifiers in ways that are mnemonically significant? Finally, if you combine these two approaches with the compounding scheme mentioned earlier (using linking morphemes), you will be able to lexify any concept while absolutely minimizing the number of root morphemes in the language. Incidentally, this approach also makes it trivially easy to create a language with a self-segregating morphology.

Concerning concept mapping:

First, let me repeat a paragraph I wrote above and then expand upon it:

> In other words, use a back door approach - start with a powerful derivational system, and iteratively decompose words from a natural language and apply all derivations to the resulting root morphemes. In doing so, many additional useful words will be automatically created, making it unnecessary to decompose a large fraction of the remaining natural language vocabulary.

This approach won't guarantee that concept space will be perfectly subdivided, but it will be as close as you can get. If anyone knows of a better system, please tell us about it.

Another fairly obvious advantage is that your AL will be easier to learn, since you'll be able to create many words from a small number of basic morphemes. Ad hoc borrowings from natural languages will be minimized.

Also, such a rigorous approach to word design has some interesting consequences that may not be immediately obvious. If you use this kind of approach, you'll find that many of the words you create have close (but not quite exact) counterparts in your native language. However, this lack of precise overlap is exactly what you ALWAYS experience whenever you study a different language.

In fact, it is this aspect of vocabulary design that seems to frustrate so many AL designers, who feel that they must capture all of the subtleties of their native language. In doing so, they merely

end up creating a clone of the vocabulary of their natural language. The result is inherently biased, semantically imprecise, and difficult to learn for speakers of other natural languages. It is extremely important to keep in mind that words from different languages that are essentially equivalent in meaning RARELY overlap completely.

Fortunately, all of this does NOT mean that your AL will lack subtlety. In fact, with a powerful and semantically precise derivational morphology, your AL can capture a great deal of subtlety, and can go considerably beyond any natural language. The only difference is that, unlike a natural language, the subtleties will be predictable rather than idiosyncratic, and the results will be eminently neutral.

So, do you want to create a clone of an existing vocabulary? Or do you want to maximize the neutrality and ease-of-learning of the vocabulary of your AL? You can't have it both ways.

Concerning hidden irregularities:

A classificational system automatically solves all count/mass/group problems, since the classification will indicate the basic nature of the entity represented by the noun. Other derivational morphemes (let's call them "class-changing morphemes") can then be used to convert the basic interpretation into one of the others. For example, from the basic substance "sand", we can derive the instance of it, "a grain of sand". From the basic animal "sheep", we can derive its group meaning, "flock", and its mass meaning, "mutton". Each basic classifier would have a default use depending on the nature of the classifier. Further derivation would be used to create non-default forms. With this approach, it would not even be possible to copy the idiosyncratic interpretations from a natural language, since the classificational system would eliminate all such idiosyncrasy.

All of the problems of verbal argument structure are solved in a classificational system. My much longer monograph on Lexical Semantics goes into considerable detail on this point, so I won't say much here. Basically, though, verbs are created by combining a root morpheme that indicates a state or action with a classifier which indicates the verb's argument structure. For example, the following verbs are formed from the same root morpheme, but with different verbal classifiers that indicate the verb's argument structure:

```
to teach (someone): subject is agent, object is patient
to teach (something): subject is agent, object is theme
to learn: subject is patient, object is theme
to study: subject is both agent and patient, object is
          theme
```

As illustration, the semantics of the English verb "to teach someone something" can be paraphrased as: 'agent' causes 'patient' to undergo a change of state from less knowledgeable to more knowledgeable about 'theme'.

You will also need to make distinctions between verbs which indicate steady states and verbs which indicate changes of state. The above examples all indicate changes of state (i.e., the 'patient' gains in knowledge). Some steady-state counterparts, formed from the same root morpheme, would be:

```
to know: subject is patient, object is theme
to be knowledgeable or smart: subject is patient, no
        object
to review (in the sense "keep oneself up-to-date"):
        subject is  both agent and patient, object is
        theme
```

You will also need an action classifier, which would indicate an ATTEMPT to achieve a change of state, but with no indication of success or failure. For example, the root morpheme for the above examples could be combined with an action classifier to create the verb "to instruct".

Thus, the verb classifier indicates the verb's argument structure, and allows creation of related verbs from the same root morpheme, verbs that almost always require separate morphemes in English.

Finally, if your AL has a comprehensive system for grammatical voice, even more words can be derived from the same morpheme. For example, if your language has an inverse voice (English does not), you could derive the verbs "to own" and "to belong to" from the same root morpheme. Ditto for pairs such as "parent/child", "doctor/patient", "employer/employee", "left/right", "above/below", "give/obtain", "send/receive", etc. Note that these are not opposites! They are _inverses_ (also called _converses_). Many other words can also be derived from the same roots if your AL implements other voice transformations such as middle, anti-passive, instrumental, etc. You can save an awful lot of morphemes if you do it right. And even though English doesn't do it this way, there are many other natural languages that do. So there's nothing inherently unnatural about this kind of system. It's almost certain, though, that no SINGLE natural language has such a comprehensive and regular system.

Finally, for those among you who want a Euroclone, I'm sorry, but I have nothing to offer you. Besides, I doubt if any of you even got this far. :-)

In a subsequent post, Rick Harrison chided me for semantic imprecision in my approach towards noun design. I responded with the following (somewhat edited):

Keep in mind that I'm talking about a CLASSIFICATIONAL language where classifying morphemes are used in both verb and noun formation. Since there is no way to use verbal roots with noun classifiers, and vice versa, in a way that is semantically precise, you can either create a completely different set of root morphemes for nouns, or you can re-use the verb roots for their mnemonic value.

Thus, for nouns, the combination of root+classifier becomes a de facto new root, even though it has the morphology of root+classifier. There is nothing "fuzzy" about it as long as you keep in mind that it's just a mnemonic aid. To me, it seems like a great way to re-use roots that would otherwise be underutilized.

Most complex nominals used in natural languages are not semantically precise - they simply provide clues. What I'm suggesting is something akin to "blurry" English words such as "whitefish", "highland", "seahorse", etc. However, the noun classifiers themselves would be more generic, but would have semantically precise definitions. Thus, what I proposed is actually much closer to what is done in Bantu languages such as Swahili, since it is morphological rather than lexical.

In essense, I am suggesting that you use semantic precision only when it is practical. Re-use root morphemes as mnemonic aids when semantic precision is not practical. The alternative is to create many hundreds (perhaps thousands) of additional root morphemes which will have to be learned by the student.

Also, there is nothing typologically unnatural about my scheme. English creates many complex nominals this way (eg. "cutworm", "white water", "red ant", etc.). My approach, though, uses noun classifiers that are slightly more generic than "worm", "water" and "ant". In effect, it is much more similar to Bantu languages of Africa or several aboriginal languages of Australia. These languages, though, are at the opposite extreme from English, since their classifiers are even vaguer than what I propose. Thus, my ideas fit in quite snugly between the opposite poles of classificational possibility.

Rick claimed that my approach to word design would be more difficult to learn. Here's my response:

Difficult??? Adding regularity to word design will make it easier, not more difficult. Is Esperanto more difficult because its inflectional system is perfectly regular? Of course not. Just because perfect regularity in a natural language is extremely rare does not mean that we should avoid it in

the construction of an AL. Or are you saying that it's okay to have regularity in syntax and inflectional morphology, but that it's NOT okay to have regularity in derivational morphology or lexical semantics?

I suggest that most ALs are irregular in derivational morphology and lexical semantics because their designers are not aware that such regularity is even possible.

Also, instead of being forced to learn thousands of unique-but-related verbs, I would rather learn about one-tenth as many, plus a few dozen classifiers and a few perfectly regular rules that apply without exception. As for nouns, mnemonic aids make them easier to learn - their meanings are unpredictable only if you fool yourself into thinking that they SHOULD BE predictable.

I think (hope?) that there are two reasons why you have difficulty with my proposal. First, you raised a topic that I've given a lot of thought to, and I tried to summarize a large quantity of material that I've written on the topic in just a few paragraphs. Misunderstanding was inevitable. Second, a classificational language may not hold much appeal for you. If so, I'm sure you're not alone.

I choose this approach because it has several advantages. First, and least important, it makes word design fast and easy. Second, it makes learning the language easier. Third, it is totally neutral - no one will accuse you of cloning your native language. Yet nothing in my approach is unnatural - every aspect of it has counterparts in some natural languages. Fourth, and most importantly, is that a powerful classificational and derivational system FORCES the AL designer to be systematic. If done properly, it will prevent the adoption of ad hoc solutions to design problems.

Aaaiiieeeyaaah! That fourth point is SO IMPORTANT, that I want to repeat it. But I won't. :-)

I also believe that the result will have more esthetic appeal to a larger number of people of varied backgrounds. An AL with a large contribution from European languages may appeal to Europeans, but it will probably not be as appealing to non-Europeans.

Postscript:

In the discussions that took place on the conlang email list, I only mentioned the possibility of precisely defining verb roots and then re-using them for their mnemonic value in the design of nouns. It is possible, of course, to do the exact opposite by precisely defining the noun roots and re-using them for their mnemonic value in the design of verbs.

I do not feel that this is a wise approach for the following reasons:

1. Precisely defined verb roots will signify states and actions which can provide a very good indication of the meaning of a noun. However, the reverse is NOT true - precisely defined noun roots can NOT provide a very good indication of the meaning of a verb. For example:

```
whale = big + swim + mammal classifier
dolphin = talk + swim + mammal classifier
penguin = swim + bird classifier
```

However, if instead I precisely defined roots for "whale", "dolphin" and "penguin", how would I use them to create verbs? The problem, of course, is that an entity such as a penguin has MANY attributes, and deciding which one is most cogent is difficult, if not impossible. In other words, going from verb to noun will be much more productive and can provide a greater degree of relative semantic precision than going from noun to verb.

2. Basic noun roots will far outnumber basic verb roots, increasing the number of roots that have to be learned.

End of Essay