The Language Construction Kit

Models

Natural and unnatural languages

[back to outline]

I personally like naturalistic languages, so my invented languages are full of irregularities, quirky lexical derivations, and interesting idioms.

It's easier, no doubt, to create a "logical" language, and desirable if you want to create an auxiliary interlanguage, à la Esperanto. The danger here is a) creating a system so pristine, so abstract, that it's also impossible to learn; or b) not noticing when you reproduce some illogicality present in the models you're using. Ask me about the irregularities of Esperanto sometime.

Non-Western (or at least non-English) models

[back to outline]

Looking at some non-Indo-European languages, such as Quechua [see my intro to Quechua here in Metaverse], Chinese, Turkish, Arabic, or Swahili, can be eye-opening.

Learn other languages, if you can. If languages are difficult for you, just skim a grammar for nice ideas to steal. Bernard Comrie's *The World's Major Languages* contains meaty descriptions of fifty languages. Anatole Lyovin's *An Introduction to the Languages of the World* readably surveys all the world's language families, pointing out touristic highlights, and gives more detailed sketches of some important languages Comrie skips.

If you don't know another language well, you're pretty much doomed to produce ciphers of English. Checking out grammars (or this html file) can help you avoid duplicating English grammar, and give you some neat ideas to try out; but the real difficulty is in the lexicon. If all you know is English, you'll tend to duplicate the structure and idioms of the English vocabulary. Below I'll give you some hints on minimizing this problem.

Sounds

Non-linguists will often start with the alphabet and add a few apostrophes and diacritical marks. The results are likely to be something that looks too much like English, has many more sounds than necessary, and which even the author doesn't know how to pronounce.

You'll get better results the more you know about **phonetics** (the study of the possible sounds of language) and **phonology** (how sounds are actually used in language). Useful references are J.C. Catford, *A Practical Introduction to Phonetics* (excellent for home study), and Roger Lass, *Phonology*. Below is a quick overview.

Types of consonants

Consonants are formed by obstructing the flow of air from the lungs. As a first approximation, consonants vary in these dimensions:

- **Place of articulation**-- where the obstruction occurs:
    - *labial*: lips (w), lips + teeth (f)
    - *dental*: teeth (th, French or Spanish t)
    - *alveolar*: behind the teeth (s, English t, Spanish r)
    - *palato-alveolar*: further back from the teeth (sh, American r)
    - *palatal*: top of palate (Russian ch)
    - *velar*: back of the mouth (k, ng)
    - *uvular*: way back in the mouth (Arabic q, French r)
    - *glottal*: back in the throat (h, glottal stop as in John Lennon saying bottle).
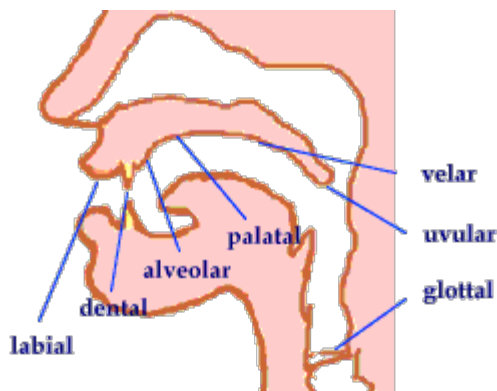
- **Degree of closure**. This proceeds in steps
    - from *stops* (stopping the airflow entirely: p t k)
    - to *fricatives* (impeding it enough to cause audible friction: f s sh kh)
    - to *approximants* (barely impeding it: r l w y).
    - An *affricate* is a stop plus a fricative, which must occur at the same place of articulation: t + sh = ch, d + zh = j.
- **Voicing**: whether the vocal cords are vibrating or not. That's the difference between f and v, t and

d, k and g, sh and zh.
- **Nasalization**: whether air travels through the nose as well as the mouth. For instance, m, n, and ng are stops like b, d, g, but only the oral airflow is stopped.
- **Aspiration**: whether stops are released lightly, or with a noticeable puff of air. In Chinese, Hindi, or Quechua, there are series of aspirated and non-aspirated stops.
- **Palatalization**: whether the tongue is raised toward the top of the mouth while pronouncing the consonant. In Russian and Gaelic, there are distinct series of palatalized and non-palatalized consonants.

English consonants can be arranged in a grid like this:

	labial	lab-dnt	dental	alv	alv-pal	velar	glottal
stop	p b			t d		k g	
fricative		f v	th th	s z	sh zh		h
affricate					ch j		
approximant	w			r l	y		
nasal	m			n		ng	

Sometimes the same sound in a language takes different forms based on its position in the word. For instance, English p is aspirated at the beginning of a word, but non-aspirated elsewhere; or, English m is usually labial, but it's labiodental before an f (compare *schematic, emphatic*).

Linguists call the basic sounds of a language, the ones that can distinguish one word from another, **phonemes**, and the actual sounds as pronounced, **phones**. They'd say that English has a phoneme /p/, which has two phonetic realizations or **allophones**, aspirated [ph] and non-aspirated [p].

Inventing consonants

[back to outline]

You'll notice that the grid of consonants for English has gaps in it. Does this mean you can invent new sounds by filling in the grid? Oh, yes.

For instance, English has voiced nasals; your language could have unvoiced nasals. English has a velar stop but no velar fricative. German has one (the ch in Bach); some languages have two, a voiced and an unvoiced one. German also has a labial affricate, pf.

Even more exciting is to add entire series of consonants using contrasts not used in English, such as palatalization or aspiration. Or remove a series English has. Cuzco Quechua, for instance, has three series of stops: aspirated, non-aspirated, and glottalized, but it doesn't distinguish voiced and unvoiced consonants.

The key to a naturalistic language, in fact, is to add (or subtract) entire dimensions. It's conceivable that a language could have a single glottalized consonant, but more likely that it will have a series of them (along the points of articulation: p' t' k'). A language might have just two palatalized consonants (Spanish does: ll, ñ), but one that has a whole series of them is more typical.

You can also add places of articulation. For instance, while English has three series of stops, Hindi has five (labial, dental, retroflex, alveolo-palatal, and velar. Retroflex consonants involve curling the tongue backwards a bit), and Arabic has six (bilabial, dental, 'emphatic' (don't ask), velar, uvular, glottal).

Some consonants are more common than others. For instance, virtually all languages have the simple stops p t k. Lass's book gives examples; see also David Crystal's *The Cambridge Encyclopedia of Language,* p. 165.
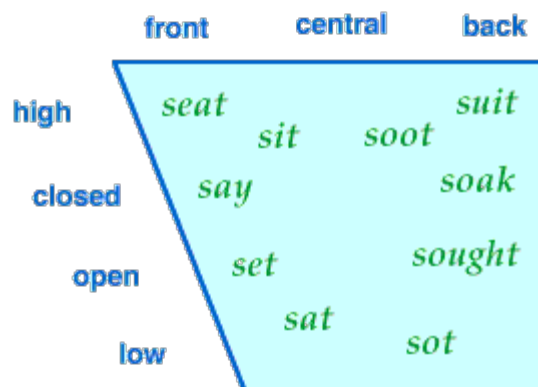

Vowels

[back to outline]

The most important aspects of vowels are **height** and **frontness**.

- **Height**: how open the inside of the mouth is. The usual scale is *high* [i, u], *mid*[e, o], and *low* [a]. There may be two middle steps in the ladder, usually called *closed* [ay, oh] and *open* [eh, aw].
- **Frontness**: how close the tongue is to the front of the mouth. Vowels can be classified into *front* (i, e), *central* (a, or the indistinct vowel in 'of'), or *back* (o, u).

You can arrange the vowels in a grid according to these two dimensions. The bottom of the grid is usually drawn shorter because there isn't as much room for the tongue to maneuver as the mouth opens more.

To get a feel for these distinctions, pronounce the words in the diagram, moving from top to bottom or side to side, and noting where your tongue is and how close it is to the roof of the mouth.

Vowels can vary along other dimensions as well:

- **Roundedness**: whether the lips are rounded (u, o) or not (i, e). English doesn't have front rounded vowels, but French and German do (Fr. u, oe; Ger. ü, ö). We also don't have (say) an unrounded u,

but Russian, Korean, and Japanese do.
- **Length**: vowels may contrast by length, as in Latin, Greek, Sanskrit, and Old English; Estonian has three degrees of length.
- **Nasalization**: like consonants, vowels can be nasalized. French, for instance, has four nasalized vowels.
- **Tenseness**: vowels can be tense or lax-- hard to explain, tho' English is an example; lax vowels are closer to the center of the vowel space-- look at *soot* and *sit* in the diagram.

English has a rather complicated vowel system:

```
                 --lax--                --tense--

              front------back        front------back

high          pit         put        peat        poot

mid           pet         putt        pate        boat

low           pat         pot            father  bought
```

Interesting simple systems include Quechua (three vowels, i u a) and Spanish (five: i e a o u). Simple vowel systems tend to spread out; a Quechua i, for instance, can sound like English *pit, peat,* or *pet.* Spanish e and o have two allophones each: open (as in *pet, caught*) in syllables that end in a consonant, closed (as in *pate, pot*) elsewhere.

Again, for your invented language, don't just add an exotic vowel or two; try to invent a vowel system, using the dimensions listed above. For instance, starting from the English system, you could bag the tense/lax distinction, add roundedness, and then collapse the front and back low vowels (there are often more high than low vowels).

Stress

[back to outline]

Don't forget to give a stress rule. English has unpredictable stress, and if you don't think about it your invented language will tend to work that way too.

French (lightly) stresses the last syllable. Polish and Quechua always stress the second-to-last syllable. Latin has a more complex rule: stress the second-to-last syllable, unless both final syllables are short and aren't separated by two consonants.

If the rule is absolutely regular, you don't need to indicate stress orthographically. If it's irregular, however, consider explicitly indicating it, as in Spanish: *corazón, porqué.*

In English, vowels are **reduced** to more indistinct or centralized forms when unstressed. This is one big reason (tho' not the only one) that English spelling is so difficult.

Tone

Mandarin Chinese syllables have four **tones**, or intonation contours: high level; rising; low falling, and high falling. [For *zhongguórén*: No, I haven't described the third tone wrong. Think about it.] These tones are parts of the word, and can be used to distinguish words of different meanings: *ma* 'mother', *má* 'hemp', *mâ* 'horse', *mà* 'curse'. Cantonese and Vietnamese have six tones. [The first tone should have a straight line over the vowel, and the circumflex over the third tone should be inverted, but this is the best I can do in html, and it beats adding numbers.]

If that seems a bit elaborate, you might consider a pitch-accent system, such as I used in another invented language, Cuêzi: the stress in a word can either be high or low in pitch. Japanese and ancient Greek are pitch-accent languages.

In (standard) Japanese, syllables can be either high or low pitch; each word has a particular 'melody' or sequence of high and low syllables-- e.g. *ikebana* 'flower arrangement' has the melody LHLL; *sashimi* 'sliced raw fish' has LHH; *kokoro* 'heart' has LHL. It rather sounds as if a tone has to be remembered for each syllable; but this turns out not to be the case. All you must learn for each word is the location of the 'accent', the main drop in pitch. Then you simply apply these three rules:

- Assign high pitch to all moras (= syllables, except that a long vowel is two moras, and a final -n or a double consonant takes up a mora too)
- Change the pitch to low for all moras following the accent
- Assign low pitch to the first mora if the second is high.

Thus for *ike'bana* we have HHHH, then HHLL, then LHLL.

Phonological constraints

Every language has a series of constraints on what possible words can occur in the language. For instance, as an English speaker you know somehow that *blick* and *drass* are possible words, though they don't happen to exist, but *vlim* and *mtar* couldn't possibly be English.

Designing the **phonological constraints** in your language will go a long, long way to giving it its own distinctive flavor.

Start with a distinctive syllable pattern. For instance,

- Japanese basically allows only (C)V(V)(n): *Ranma, Akane, Tatewaki Kunoo, Rumiko Takahashi, Gojira, Tookyoo, konkuuru, sushi,* etc.

- Mandarin Chinese allows (C)(i, u)V(w, y, n, ng): *wô, shì, Mêiguó, rén, wényán, chìàn, mànhuà, Wáng, Zhang,* etc.

- Quechua allows (C)V(C): *Wallpakuna sarata mikuchkanku, achka allin hatun mosoq puka wasikuna,* etc.

- English goes as far as (s) + (C) + (r, l, w, y) + (V) + V + (C) + (C) + (C): *sprite, thinks.*

Try to generalize your constraints. For instance, m + t is illegal at the beginning of a word in English. We could generalize this to [nasal] + [stop]. The rule against v + l generalizes at least to [voiced fricative] + [approximant].

Another process to be aware of is **assimilation**. Adjoining consonants tend to assimilate to the same place of articulation. That's why Latin *in- + -port = import, ad + simil- = assimil-*. It's why the plural -s sounds like z after a voiced stop, as in *dogs* or *moms*. It's also why Larry Niven's *klomter*, from *The Integral Trees*, rings so false. m + t (though not impossible) is difficult, since each sound occurs at a different place of articulation; both sounds are likely either to shift to the dental position (*klonder*) or the labial (*klomper*). Another possible outcome is the insertion of a phonetically intermediate sound: *klompter.*

Alien mouths

[back to outline]

If you're inventing a language for aliens, you'll probably want to give them *really different* sounds (if they have speech at all, of course). The Marvel Comics solution is to throw in a bunch of apostrophes: "This is Empress Nx'id"ar' of the planet Bla'no'no!" Larry Niven just violates English phonological constraints: *tnuctipun*. We can do better.

Think about the shape of the mouth of your aliens. Is it really long? That suggests adding a few more places of articulation. Perhaps the airstream itself works differently: perhaps they have no nose, and therefore can't produce nasals; or they can't stop breathing as they talk, so that all their vowels are nasal; or the airstream is at a higher velocity, producing higher-pitched sounds and perhaps more emphatic consonants. Or perhaps their anatomy allows quite odd clicks, snaps, and thuds that have become phonemes in their languages.

Several writers have come up with creatures with two vocal tracts, allowing them to pronounce two sounds at once, or accompany themselves in two-part harmony.

Or, how about sounds or syllables that vary in *tonal color*? Meanings might be distinguished by whether the voice sounds like a trombone, a violin, a trumpet, or a guitar.

Suggesting additional sounds is difficult and perhaps tiresome to the reader; an alien ambience can also be created by removing entire phonetic dimensions. An alien might be unable to produced voiced sounds (so he sounts a pit like a Cherman), or, lacking lips, might skip over labials (you nust do this to de a thentrilocooist, as ooell).

Alphabets

Orthography

[back to outline]

Once you have the sounds of your language down, you'll want to create an orthography-- that is, a standard way of representing those sounds in the Roman alphabet.

I don't recommend trying to be very creative here. For instance, you could represent **a e i o u** as **ö é ee aw ù**, with the accents reversed at the end of the word. An outlandish orthography is probably an attempt to jazz up a phonetic system that didn't turn out to be interestingly different from English. Work on the sounds, then find a way to spell them in a straightforward fashion.

If you're inventing a language for a fantasy world, it's wise to take account of how English-speaking readers will mangle your beautiful words. Tolkien is the model here: he spelled **Quenya** as if it were Latin, didn't introduce any really vile spellings, and kindly indicated final e's that must be pronounced. Still, he couldn't resist demanding that c and g always be hard (I couldn't either, for Verdurian), which probably means that a lot of his names (e.g. Celeborn) are commonly mispronounced.

Marc Okrand, inventing **Klingon**, had the clever idea of using upper and lowercase letters with different phonetic values. This has the advantage of doubling the letters available without using diacritics, but it's not very aesthetic and it sure is a tax on memory.

Or you may go for neatness, as I did in inventing Verdurian. I don't like digraphs, so I adapted Czech orthography-- č for ch, š for sh, etc. This ultimately involved creating a special Macintosh font, so I was probably crazy. (Note however that fonts for non-Western-European languages are plentiful by now.)

A sense of variation among the nations of your world can be achieved by using **different transliteration styles** for each. In my fantasy world, for instance, Verdurian Ďarcaln and Barakhinei Dhârkalen are not pronounced that much differently, but the differing orthographies give each a different feeling. Surely you'd rather visit civilized Ďarcaln than dark and brooding Dhârkalen? (Tricked you. It's the same place.)

If you're inventing an **interlanguage**, of course, you shouldn't worry about English conventions; create the most straightforward romanization you can. You're only asking for trouble, however, if you invent new diacritic marks, as the inventor of Esperanto did.

An example

[back to outline]

Here's the alphabet I came up with for Verdurian:

Note that there's a one-to-one correspondence between the Verdurian alphabet and the standard English representation. This is not very naturalistic-- transliteration schemes are not usually this straightforward-- but it's a good place to start. Once you can fluently read your own alphabet, feel free to add complications.

A good alphabet can't be created in a day. This one took shape over a period of weeks, as I played with various letterforms.

Keep the letters looking distinct. The best alphabets spread out over the conceptual graphic space, so that letters can't be confused for one another. Tolkien is a bad example here: the elves must have been tormented by dyslexia. If letters start to approach each other too closely, users find ways to distinguish them, in the way that computer programmers, for instance, write zeroes with a slash. Europeans write 1 with an elaborate introductory swash-- impossible to confuse with I, but looking much like a 7, which has therefore acquired a horizontal slash!

Remember that letters are written over and over again, over the life of an individual or a civilization. Elaborate letters are likely to be simplified. You can simulate this process by writing the letter over and over yourself; the appropriate simplifications will suggest themselves automatically.

Note that I supplied upper and lower case forms, as in the Roman and Greek alphabets. The lowercase forms are all cursive simplifications of the uppercase forms (which are also the ancient forms). In retrospect I probably shouldn't have imitated the mixed-case system, which on our world is basically limited to Western alphabets. I should have kept the 'uppercase' forms for ancient times, the 'lowercase' forms for modern times.

I tried to give the letters individual histories, as with our alphabet. The letter **t**, for instance, derives from a picture of a cup, **touresiu** in Cuêzi; **n** was originally a picture of a foot (**nega**). I have to admit that I did this backwards-- I invented pictograms that could have developed into the letters, which I had devised years before!

Also note that the voiced consonants, in the uppercase forms, are simply the unvoiced forms with a bar over them (this is a bit obscured with d and t), and that the letters for **š č ž** are all transparent variations of each other. This slightly violates my 'maximally distinct' rule, but I think it adds interest to the alphabet.

You'll also notice both **c** and **k** in the alphabet. This is the sort of ethnocentrism it's all too easy to fall into. Why would another language duplicate the convoluted history of our alphabet's c and k? I've reinterpreted these symbols to refer to /k/ and /q/.

Diacritics

[back to outline]

Some advice: never use a diacritical mark without giving it a specific meaning, preferably one which it retains in all uses. I made this mistake in Verdurian: I used ö and ü as in German, but ë somewhat as in

Russian (indicating palatalization of the previous consonant), and ä as a mere doubling of a. I was smarter by the time I got to Cuêzi: the circumflex consistently indicates a low-pitch accent.

Avoid using apostrophes just to make words look foreign or alien. Since apostrophes are used in contradictory ways (they represent the glottal stop in Arabic or Hawai'ian, glottalization in Quechua, palatalization in Russian, aspiration or a syllable boundary in Chinese, and omitted sounds in English, French, and Italian), they end up suggesting nothing at all to the reader.

Fancier writing systems

[back to outline]

What, you say you want to build a syllabary? A cursive form of your alphabet? A logographic system?

Read a good book on how writing systems work. *Writing Systems* by Geoffrey Sampson is a very good book.

If that seems too much, read up on the type of writing system you want to imitate: Chinese characters, the Japanese or Maya syllabary, the Sanskrit syllabic alphabet, the Korean featural code, the all-cursive Arabic alphabet, and so on.

A book like Kenneth Katzer's *Languages of the World* gives examples of a wide variety of scripts. Comrie's *The World's Major Languages* does the same, but gives more detail. Or invest in the 800-pound gorilla of the field, Daniels & Bright's *The World's Writing Systems,* which explains how *every* writing system in the world works.

Note that logographic scripts and syllabaries tend to work best with languages that have a very limited syllabic structure-- Japanese, with (C)V(n), is close to ideal; English is close to pessimal.

Word building

How many words do you need?

[back to outline]

Where the conlang bug bites, the **Speedtalk** meme is sure to follow. Let Robert Heinlein explain it:

> Long before, Ogden and Richards had shown that eight hundred and fifty words were sufficient vocabulary to express anything that could be expressed by "normal" human vocabularies, with the aid of a handful of special words-- a hundred odd-- for each special field, such as horse racing or ballistics. About the same time phoneticians had analyzed all human tongues into about a hundred-odd sounds, represented by the letters of a general phonetic alphabet.
> ... One phonetic symbol was equivalent to an entire word in a "normal" language, one

Speedtalk word was equal to an entire sentence.
--"Gulf", in *Assignment in Eternity*, 1953

This is a tempting idea, not least because it promises to save us a good deal of work. Why invent thousands of words if a hundred will do?

The unfortunate truth is that **Ogden and Richards cheated**. They were able to reduce the vocabulary of Basic English so much by taking advantage of idioms like *make good* for *succeed.* That may save a word, but it's still a lexical entry that must be learned as a unit, with no help from its component pieces. Plus, the whole process was highly irregular. (*Make bad* doesn't mean *fail*.)

The Speedtalk idea may seem to receive support from such observations as that 80% of English text makes use of only the most frequent 3000 words, and 50% makes use of only 100 words. However (as linguist Henry Kučera points out), there's an **inverse relationship between frequency and information content**: the most frequent words are function words (prepositions, particles, conjunctions, pronouns), which don't contribute much to meaning (and indeed can be left out entirely, as in newspaper headlines), while the least frequent words are important content words. It doesn't do you much good to understand 80% of the words in a sentence if the remaining 20% are the most important for understanding its meaning.

The other problem is that **redundancy isn't a bug, it's a feature**. Claude Shannon showed that the information content of English text was about one bit per letter-- not too high considering that for random text it's about five bits a letter. Sounds inefficient, huh? On the other hand, we don't actually hear every sound (or, if we're accomplished readers, read every letter) in a word. We use the built-in redundancy of language to understand what's said anyway.

To put it another way: y cn ndrstnd Nglsh txt vn wtht th vwls, or shouted into a nor'easter, or over a staticky phone line. Similarly distorted Speedtalk would be impossible to understand, since entire morphemes would be missing or mistaken. Very probably the degree of redundancy of human languages is pretty precisely calibrated to the minimum level of information needed to cope with typical levels of distortion.

However, go ahead and play with the Speedtalk idea. It's good for some hours of fun, working out as minimal a set of primitives as you can; and the habit of paraphrase it gives you is very useful in creating languages. Just don't take it too seriously; if you do, your punishment is to learn 850 words of any actual foreign language and be set down in a city of monolingual speakers of that language.

Alien or *a priori* languages

[back to outline]

If you're making up a language for a different world, you want, of course, words that don't sound like any existing language. For this you simply need to make up words that use the sounds and the syllable structure in your language.

This can fairly quickly get tiresome. I don't advise you to sit down and come up with a hundred words at

once; you're likely to run out of inspiration, or find that all the words are starting to sound the same. You may also be creating new roots where you could more easily derive the word from existing roots.

It's not hard to write computer programs that will randomly generate words for your language (even respecting its syllable structure). If you do, remember that sounds (and syllable structures) are not equiprobably distributed in natural languages. English uses many more t's than f's, more f's than z's.

Resist the temptation to give a meaning for every possible syllable. Real languages don't work like that (unless the number of possibilities is quite low). Even if you're working on a highly structured auxiliary language, you'll want some maneuvering room for future expansion. And the speakers of your language shouldn't have to throw out an old word whenever they want to construct a coinage or an abbreviation.

You will want a mixture of word lengths for variety; but don't invent too many long words. It's better to derive long words by combining shorter words, or adding suffixes. Or, imitating the way English is full of polysyllabic borrowings from Latin and Greek, or Japanese is full of Chinese loanwords, create two languages, and build words in one out of components in the other.

A few half-recognizable borrowings

[back to outline]

I intended Verdurian to look mildly familiar, as if it could be a distant relative of the European languages. For example:

> **Sul Aď e otál mudray dy tü, dalu esë, er ya cečel rho sen e sënul.**
> Only God is as wise as you, my king, and even there I'm not certain.

> **So cuon er so ailuro eu druki. Cuon ride še slušir misotém ailurei. So ailuro e arašó rizuec.**
> The dog and the cat are friends. The dog laughs at the cat's jokes. The cat is quite amusing.

To achieve this impression, I borrowed from a number of earthly languages-- e.g. **ailuro** 'cat' and **cuon** 'dog' are adapted from Greek; **sul** 'only' from French; **rizir** 'amuse' and **ya** 'indeed' from Spanish; **druk** 'friend' and **slušir** 'hear' from Russian. The friendly orthography and the simple (C)(C)V(C) syllable structure also help make the language inviting.

By contrast, another language, **Xurnáš**, was intended to look more alien:

> **Ir nevu jadzies mnošudacij. Toc šizen ri tos bunjači šasik rili. Tos denjic šuš bunji dis kezi. Syu šačo cu šuš izraugi.**
> My niece is dating a sculptor. She can see no flaws in him. He hopes one day to govern a province. Myself, I don't envy that province.

Languages based on existing languages

Interlanguages are often based on existing languages; for instance, Esperanto is chiefly based on French, Italian, German, and English. Here the problem of creating words largely reduces to one of acquiring enough good dictionaries.

A few language creators have tried to approach the task systematically-- e.g. Interlingua is based on nine languages, and usually adopts the word found in the most languages.

Lojban uses a wider variety of languages, including some non-Western ones, and uses a statistical algorithm to produce an intermediate form. The intention is to provide some mnemonic assistance to a very wide variety of speakers. It's an intriguing idea, although the execution is so subtle that the language is often mistaken for *a priori*.

Sound symbolism

Some linguists claim to have found some common meaning patterns among human languages. For instance, front vowels (i, e) are said to suggest smallness, softness, or high pitch; low and back vowels (a, u, o) to suggest largeness, loudness, or low pitch. Compare *itty-bitty, whisper, tinkle, twitter, beep, screech, chirp,* with *humongous, shout, gong, clatter, crash, bam, growl, rumble;* or Spanish *mujercita* 'little woman' with *mujerona* 'big woman'. Cecil Adams took advantage of this pattern when he commented, on the subject of penis enlargement surgery, that "if nature has equipped you with a ding rather than a dong, you'll just have to live with it."

Exceptions aren't hard to find, of course-- notably *small* and *big*.

Inventing alien languages, authors also simply make use of what we might call phonetic stereotypes. Tolkien's Orkish, for instance, makes heavy use of guttural sounds and is full of consonants, while his Elvish tongues are more vocalic, and seem to have plenty of pleasant-sounding l's and r's.

Some guidelines for not reinventing the English vocabulary

- If the literal meaning of an expression doesn't make sense (e.g. "make good", "go all out", "have it in for someone", "look lived-in"), you're probably dealing with an idiom. Translate using expressions that make sense literally ("succeed", "work at full capacity", "have a grudge against someone", "seem inhabited"), or create your own idioms ("laugh at hell", "play bee", "circle your eye at someone", "be breathed and worn").
- Look through the foreign-to-English section of a bilingual dictionary. Look at the range of English meanings particular foreign words have: think about what kind of root concept could cover all of them. Look at the foreign words used to translate a single English word: try to see what distinctions the foreign language is making where English uses that one word.

- Derive your lexicon from basic roots using regular derivation processes.
- Look up the etymology of the English word. See if you can come up with an alternative process.
- Consider a whole class of related English words-- verbs of motion, for instance. Design the related class of words in your language, dividing up the conceptual space in your own way.
- Read Lakoff and Johnson, *Metaphors We Live By*. Create your own metaphors and the vocabulary that goes with them.
- Read a text on semantics (Palmer's *Semantics* is short; Takao Suzuki's *Japanese and the Japanese: Words in Culture*, a.k.a. *Words in Context*, is wonderful), for a greater awareness of the structure of the lexicon.
- For a fantasy language, think about the culture that your language serves. What concepts are most important to it? They will likely have many synonyms, or even be reflected directly in the grammar. What's its history or mythology? They will probably generate a number of derived words.

[back to outline] [on to grammar]

The Language Construction Kit is © 1996 by Mark Rosenfelder.