How to create a language
By Pablo David Flores

(partly based on Mark Rosenfelder's Language Construction Kit[1])
Edited for PDF-format by Gulliver Methuen-Campbell.

Introduction

The original, online version of this file is available in HTML form at the following address: http://www.angelfire.com/ego/pdf/ng/lng/how/how_toc.html

There are very few differences between the PDF and online versions; all hyperlinks have been moved to footnotes to keep the body of the text relatively clear and coherent.

These pages are intended for people interested in creating languages for fictional purposes (or just for fun) and in linguistics in general. They're not meant to be an online linguistics course, but you sure can learn quite a few things about linguistics by reading them, the same way I, not being a linguist, learned from others. They're also not supposed to be a guide to the creation of auxiliary or international languages such as Esperanto.

The pages are divided into two main fields: phonology and grammar. These in turn cover topics going from phoneme theory and phonotactics to typology, morphology and syntax, with interspersed comments on orthographical representation, diachronically change of both grammar and phonology, and methods of word generation. Technical terms are often used -- correctly and clearly, I hope -- but no piece of jargon is left unexplained.

Before starting, I'd like to give the credit deserved to Mark Rosenfelder, who gave me the first tool to engage myself in serious language development. The structure and main points on these pages are based on his work, although I have tried not to copy everything (which would be quite silly of me), but instead give some advice and go deeper into some details he didn't mention in the Language Construction Kit. Some material has also been drawn from the Model Languages[2] newsletter, run by Jeffrey Henning. Fellow conlangers and helpful readers suggested a lot of corrections and useful additions to the original version of this document. Some explanations have been adapted from posts to the Conlang list[3]. Thank you all!

I've used examples from, or mentioned, a good couple dozens of languages, both natural and fictional, the latter by me or by others. I have tried to be as accurate as I can; it all depends on my sources, which are sometimes books from a library that I took back months or years ago, so I have to cite from memory. This also explains the mentions of "an African language" whose name I can't remember, and the somewhat dubitative nature of some statements. Nevertheless, I have a good memory and I believe every piece of information is correct as far as I know; I haven't included conjectures or guesses that aren't noted as such.

If someone finds anything that seems to be a mistake, or wishes to make a suggestion, or wants a particular topic to be discussed here, please write to me[4].

[1] http://www.zompist.com/kit.html
[2] http://www.langmaker.com
[3] http://www.angelfire.com/ego/pdf/ng/lng/conlanglist.html
[4] http://www.angelfire.com/ego/pdf/feedback.html

Sounds

Sounds are the way a language first becomes real in the physical world, so we'll start talking about them. Some people believe that a letter in their alphabet is the same as a sound, or that all sounds in all languages are the same (as the sounds in their own language), only with different 'accents'. Why this is false can be easily explained and understood by most people. I won't mix sound with representation or transliteration, here, and I'll give examples of sounds in languages that may be familiar to you just in order to simplify things. Other languages need not use the same sounds as one's own, or pronounce them the same way.

However, we'll have to stop at a fairly abstract topic first, in order to move on confidently then. We'll talk about **phones** (real sounds) and **phonemes** (the sounds in a language as seen by a linguist).

Phones and phonemes

The immense (actually infinitely dense) range of possible sounds that a human being can produce are called **phones**. Each particular position of the lips, tongue, and other features in our organs of speech can be thought of a point in a multidimensional continuum. Given two positions of the tongue with respect to the interior of the mouth, there is always a position in the middle, and so on. Remember the real numbers from school?

However, we group sounds into prototypical examples of themselves, to study them better and more easily, and we call each of these a **phone**, a single sound that can be described by certain features (for example: the tongue touches the teeth, vocal chords are vibrating, etc.).

In a particular language, we'll find a lot of phones, but those are not the object of our study. We need to distinguish the sounds that are distinguishable by the speakers of the language, i. e. that they conceptualize as different sounds. These are called **phonemes**. A phoneme can be thought of as a family of related sounds which are regarded as the same phonetic unit by the speakers. The different sounds that are considered part of the same phoneme are called **allophones** or allophonic variants. Each allophone is said to be a **realization** of the given phoneme.

In phonetic symbols, phonemic transcriptions are surrounded by slashes (/X/), while phonetic transcriptions (those who distinguish the different phones that are allophones of the phoneme) are surrounded by square brackets ([X]). The standard phonetic symbols that are used by most people nowadays belong to a set, the **IPA** (International Phonetic Alphabet). They are a lot, and you'd need a special font to see them if I used them here, so I (as most people that have to handle IPA symbols in the Web or e-mail) use a transliteration that allows IPA to be represented by 7-bit ASCII characters. There are several kinds of ASCII-IPA renderings[5]. In this site I tend towards a version of the X-SAMPA scheme, as employed customarily in the CONLANG e-mail list (see a chart). If you want to listen to the sounds in the IPA, try IPAHelp[6].

Back on topic... The allophones of a phoneme need not be similar sounds (from one's own point of view, that is). For example, the Spanish phoneme /b/ has two allophones, [b] (like the

[5] http://www.cs.brown.edu/~dpb/ascii-ipa.html
[6] http://www.sil.org/computing/speechtools/ipahelp.htm

English *b*) and [β] (a bilabial fricative, similar to English *v* but with air blown between the two lips). These are similar, related sounds. On the other hand, Japanese /h/ has three allophones, [h], [ç] (more or less like the sound in 'huge', or the German Ich-Laut), and [φ] (like /f/, but blown between the two lips). These are quite different sounds. What makes them allophones is that Japanese speakers treat them as the same sound (phoneme). Note that in German, for example, [ç] and [h] are allophones of different phonemes, so they *can* distinguish words.

Allophones of a given phoneme are in **complementary distribution**. This means that which allophone appears in a particular position depends on the position, and position determines one and only one allophone to be present, and not any of the others. Coming back to our examples, Spanish /b/ is [β] in all positions except after /m/ and when clearly starting a word (for example, at the beginning of a sentence); it's [b] otherwise. You can't have [mβ] or [ab], because only [mb] and [aβ] are possible.

This all boils down to a fact that defines what phonemes are: they are sounds that can make words different. If two sounds are allophones, you can't produce two words exchanging them, because they are in fact the same; if you pronounce one where the other should be, it'll sound bad to native speakers, but they won't hear a *different* word.

You'll see more of this afterwards, in other sections, since I'll keep repeating myself. If you don't understand the concept of phoneme, you'd better keep trying.

Vowels vs. consonants

The sounds used in any language can be divided (generally) into consonants and vowels. This division is not necessarily universal; in many languages some "consonants" like *r, m, n, l,* are actually vowels (this is, they are treated as syllable nuclei, can be stressed, or lengthened, etc.). For example, Sanskrit has syllabic *l* and *r* (as in *Rgveda*); and Japanese syllable-final *n* is syllabic (actually "moraic", but that's a distinction I won't explain here). The division between vowels and consonants is a matter of closure: the more closed the air passages are, the more consonantic a sound is. We will examine the different kinds of sounds using this scale.

Consonants

Sounds vary along **dimensions**. These represent ranges of possible features, or yes-no features. Each language has a phonology with one or more dimensions within which sounds are placed and recognized. One important dimension is the **degree of closure**. According to this, consonants can be classified into:

- **Stops**: the airflow is completely stopped for a moment, and then released, to produce the sound. The sounds *p, k, b, d* in English *pin, king, ban, dad* are stops.
- **Fricatives**: the airflow is not completely stopped, but it causes an audible friction. For example: English *s*, *sh*, *v*, German *ch* as in *Achtung, Ich, München*.
- **Approximants**: the airflow is barely modified at all. For example: English *w, l, r, y*.

Also an **affricate** is a stop plus a fricative occurring in the same place of articulation, like English *ch* (which can be analyzed as *t* + *sh*) or German *z* (pronounced /ʦ/).

A **click** is a sound produced by placing the tongue in position for a stop while there's a second closure somewhere else, accumulating pressure and then releasing the closure (see below).

Then there's the **place of articulation**, this is, where the obstruction or modulation of the airflow occurs. According to this, consonants can be:

- **Labial**: formed by the lips (*w, p*), or by the lips and the tongue (*f*, also called labio-dental)
- **Dental**: between the teeth and the tongue (*th*, French or Spanish *t*)
- **Alveolar**: in the alveola, the place right behind the teeth (*s*, English *t*, Spanish *r*)
- **Alveolo-palatal**: further back from the teeth (*sh*, *ch*), with the body of the tongue retracted towards the palate.
- **Palatal**: at the top of the palate (Russian *ch*, Spanish *ñ* as in *niño*)
- **Retroflex**: with the tip of tongue curled backwards, its underside touching the border of the hard palate (American *r*, in many dialects; in Sanskrit there's a complete series of retroflex consonants (which are called **cerebral**), which parallels the alveolar series *t, d, n, s*).
- **Velar**: at the back of the mouth (*k*, *ng* as in *sing*)
- **Uvular**: way back in the mouth, at the uvula (Arabic *q*, French *r*) [also called post-velar]
- **Glottal**: back in the throat (*h*, glottal stop as in *uh-oh*).

Some other dimensions are:

- **Voicing**: whether the vocal chords are vibrating (voiced) or not (voiceless or unvoiced). Sounds like *p, t, f* are voiceless, while *b, d, v* are voiced.
- **Nasalization**: whether the air goes through the nose (nasal) or not. The sounds *m, n, ŋ (ng)* are nasals.
- **Aspiration**: (this applies mostly to stops) whether there's a puff of air when releasing the airflow. Initial English *p, t, k* as in *paw, toe, kite* are aspirated (while the same sounds in *spawn, star, sky* are unaspirated).

- **Palatalization**: whether the middle part of the tongue is raised towards the palate (the top of the mouth) when pronouncing the consonants. English doesn't have palatalized consonants (see below), but Russian has a whole series.
- **Glottalization**: whether there's a glottal closure together with the main sound. English doesn't have glottalized consonants (see below), but Georgian has a whole series.

Let's examine these contrasts. I call them contrasts because that's what they are: things that may be distinguished. Linguistics is based on contrasts, on differences. If a language doesn't distinguish one sound from another, then it's the same sound for all practical purposes, and in that way it should be studied.

Voicing is a very usual contrast in Western Indo-European languages, not so in many other language families, where this distinction is not made (so in fact *p* and *b*, or *t* and *d*, are regarded as exactly the same sound). In English you might say that /p/ is a phoneme, with two phonetic realizations or allophones, [p^h] (aspirated, at the beginning of words) and [p] (non-aspirated). In Hindi, where aspirated and non-aspirated stops are regarded as different families, /p/ and /p^h/ are two phonemes.

Nasalization is quite a common contrast in many languages. The most common nasals are voiced stops, but some languages do have voiceless nasals, and a few have nasalized fricatives. If you can't imagine how to pronounce a voiceless nasal, take into account that an *m* is actually a nasalized *b*, so a voiceless *m* is a nasalized *p*: pronounce a *p* while you let air through your nose, and you're done. Many people in fact nasalize consonants (and vowels) after a nasal, although they don't notice it: the distinction is usually not phonemic (it can't be used to distinguish a word from another one).

We have already talked about **aspiration**. A language can have aspirated stops, non-aspirated ones, or both; and it can make the distinction phonemic (like Hindi) or just phonetic (like English).

Palatalization is a common device in languages. A consonant is palatalized by raising the middle part of the tongue towards the top of the mouth. Normally the palatalized consonant should be alveolar in the first place. The result is something that sounds like the original consonant plus a /j/ sound (as in *yet*, *new*, *pure*). Russian has a distinct series of palatalized consonants, transliterated with an apostrophe (*t', l', d'*). Spanish has two palatalized consonants, *ll* (only pronounced this way in Spain, not in Latin America) and *ñ* /ɲ/ (as in *año*), also found in French, written *gn* (as in *baigner*).

Glottalization is performed by closing the glotis, and opening it at the same time you pronounce the sound. The glotis is at the back of the throat. Glottalized sounds are usually stops. You can produce a glottalization by producing a **glottal stop** in the middle of the pronunciation of the original consonant, and then releasing the air in the two closures at the same time. But what's a glottal stop? In English, a glottal stop is usually pronounced as a pause before a word that begins with a vowel, especially when the previous one ends in a vowel too, as in *uh-oh*. German always places a glottal stop before an initial vowel. The glottal stop is not phonemic in American English or German, but it's quite a common phoneme in other languages, like Hawai'an (the apostrophe ' represents /ʔ/, the glottal stop) and in some dialects of British English. Glottalized consonants are also called **glottalic**

egressive or **ejective**. Georgian and Quechua have a complete series of glottalized/ejective voiceless stops.

There are also **glottalic ingressive** consonants, also known as **implosives**. Those are produced by making a sound, but just before opening the mouth also rapidly lowering the glottis to produce a hollow sounding effect. Some African languages, among others, have implosive consonants, which are also voiced stops.

There are also some contrasts I didn't mention before:

A **lateral** consonant is one in which the airflow doesn't go between the tongue and another spot, but instead leaves that space closed and lets air pass through the sides (**lateral release**). Some languages, like Welsh, have a voiceless lateral. The most common lateral we know is *l* (which is usually alveolar and voiced). However, English /l/ has two variants, one alveolar and one velar [L\], the latter occurring in syllable-final position, especially in clusters, as in *milk*. This 'dark L' is an independent phoneme in other languages.

If you use only the two main dimensions (degree of closure and place of articulation), and simplify a bit, you can show the distribution of consonants in English with a grid like this in a common variation of SAMPA[7]:

	labial	lab-dental	dental	alveolar	alveo-palatal	velar	glottal
stop	p b			t d	k g		
fricative		f v	θ ð	s Z			
affricate				tS dZ			h
approximant	w			r l		j	
nasal	m			n	ŋ		

(Where /w/ is actually labiovelar, not just labial; /j/ is palatal, not alveolo-palatal; and /r/ may be alveolar or retroflex according to dialect).

New consonants

How do you invent new consonants for your language? The first step should be deciding which contrasts you will use. English three places of articulation (POAs) for stops, which are usually the reference frame, and distinguishes voicing for most consonants and nasalization for stops.

The important thing is that the phonology of a language is a system. Consonants which are out of the system (because they use exceptional contrasts, for example) tend to be left out and disappear or are merged with similar consonants. For example, English couldn't possibly have a glottalized consonant, because it would use a contrast not found elsewhere in the language and wouldn't survive long. Exceptions are possible, of course, but try not to abuse them. If you have an exotic sound, you should have others of the same kind. On the other hand, you probably shouldn't invent many strange sounds; you must know how to pronounce

[7] http://www.wikipedia.org/wiki/Sampa.html

each of them, and be able to read your language fluently. (This also involves a careful planning of the transliteration scheme.)

Once you have decided the contrasts you'll be using, set up the grid and fill in the gaps. You'll probably have to invent new symbols or digraphs for some letters (see Writing). If you decide there are too many consonants, delete a series, or just some members. You don't have to occupy all the places in the grid (English, as you may notice, leaves lots of empty spaces). For example, you might have voiced and voiceless stops, but only voiceless fricatives and voiced nasals.

English only has two affricate consonants, voiced *j* and voiceless *ch*, and on the same position. Your language could have affricates in all positions where there's a stop and a fricative; for example *pf* (found in German, as in *Pferd*), *ts* (also in German, written *z* as in *zehn*, and in Japanese, as in *tsukuru*, though it's just an allophonic variant of /t/), *tth* /tθ/ (not in any language that I know, but possible), *tsh* (*ch*), *kkh*, etc.

You can complete a series of consonants, for example the English fricatives: there are no bilabial or velar fricatives (there's no reason why there should be any; but there's no reason why there couldn't, either). An unvoiced bilabial fricative /ɸ/ sounds like an *f* pronounced by letting air out between the lips; and an unvoiced velar fricative /x/ is just the sound represented in Spanish by *j* (as in *Juan*, *viejo*), or the sound of Hebrew *hhet*, sometimes transliterated *kh*. Some languages have both unvoiced /x/ and voiced /ɣ/. Spanish voiced stops between vowels become fricatives, though the distinction is not phonemic, so *b, d, g* in *cabo, cada, soga* are actually a bilabial fricative, a dental fricative (/ð/, English soft *th*), and a velar fricative (/ɣ/).

If you want to go right into it, you can add a contrast not used in English, and create a series of palatalized consonants. Or use aspiration as a phonemic distinction. Or even lateralizing or retroflexing consonants. As Mark Rosenfelder[8] says, the key to a naturalistic language is to add (or subtract) dimensions. Being into the study of Quechua, he mentions that it has not one, but three series of stops: aspirated, non-aspirated, and glottalized; but it doesn't distinguish between voiced and voiceless consonants. So, for a Quechua speaker, the *p* in *pat* and the *b* in *bat* would be the same sound (phoneme), but the *p* in *pat* and the one in *spat* would be clearly different.

Some sounds are more common than others. Most languages have the simple stops /p t k/. From what I've been able to gather, the average language has twice as many consonants as vowels. The simplest systems belong to Hawaiian, with only eight consonants and five vowels, and Rotokas, with six consonants and five vowels. Quechua has a lot of consonants but it's only got three vowels (/a i ʉ/, which are the most common). The most complex systems are those found in the Khoisan linguistic family; the !Xũ language (also written *!Kung*) has 141 phonemes, with 92 consonants, 47 of which are clicks. (!Xũ is pronounced as a glottalized dental click followed by a nasalized /ʉ/).

[8] http://www.zompist.com/

Vowels

Vowels are produced exactly the same way as consonants; they're not different in essential ways from consonants. The main thing is that the airflow is almost not disturbed while passing through the mouth; it's only modulated by the position of the tongue and other parts of the vocal organs. Also, vowels are usually voiced (some languages have voiceless vowels, especially at the end of words; they sound exactly as if you pronounce /h/ with the tongue and lips in position for the vowel).

Vowels can vary along these dimensions:

- **Height**: how open the mouth is. Vowels are usually classified into high (*i, u*), middle (*e, o*) and low (*a*). This scale is of course continuous, not discrete; in some cases you cannot describe a vowel as middle or low, for example, but you have to say it's higher than *a* but not so high as *e*.
- **Frontness**: how close the tongue is to the front of the mouth. Can go from front (*i, e*) to central (*a*), or back (*o, u*). Front vowels are sometimes called **palatal,** and back vowels are also called **velar**. There are also **pharyngealized** vowels (produced with the pharynx), but I can't imagine how they actually sound.
- **Roundedness**: whether the lips are rounded (*o, u,* German *ö*, French *u*) or not (*i, e, a*). (In most languages this covers it all, but Swedish has three degrees of roundedness in a front vowel, from unrouded to semi-rounded to fully-rounded, not just a yes-no choice).
- **Length**: how much you keep pronouncing the vowel, of course. English doesn't distinguish vowels by length, but Latin, Greek, Old English and many other languages do. Estonian has three degrees of length.
- **Nasalization**: like consonants, vowels can be nasalized. In English, a vowel next to a nasal may get nasalized, but this is not distinctive. In French, on the other hand, there are four vowels that can be nasalized or not.
- **Voicing**: vowels are usually voiced, but some languages have voiceless vowels (sounding exactly as /h/ pronounced with the lips and tongue in position for the vowel). In Japanese, /u/ and /i/ are usually voiceless if they aren't high-pitch and stand between voiceless consonants (but they get voiced if for some reason there's need to emphasize them.)
- **Tenseness**: difficult to explain except for examples. In English, the vowels in *pit, put* are said to be **lax**, and the ones in *peat, poot* are called **tense**. I'm sure you understand the difference!
- **Retroflexion**: the same as retroflex consonants. A vowel can be retroflexed by curling the tongue towards the back of the mouth before pronouncing it. An African language (I don't remember the name right now) has three series of three vowels each; the first is of non-retroflex vowels, the second is semi-retroflex, and the third is fully-retroflex! (I assume the neighbouring sounds tend to get retroflexed too.)
- **Constriction**: a constricted vowel sounds as if you were choking. In some languages, this and other ways of pronouncing sounds are phonemic, not just an accident.
- **Others**: there are probably more contrasts for vowels, but I don't know anything about them. Other modifications can be made by stress and tone (in tonal languages like Chinese or Vietnamese; see below).

English has this vowel system:

	lax		tense	
	front	back	front	back
high	pit	put	peat	poot
mid	pet	putt	pate	boat
low	pat	pot	father	bought

If you read a book on linguistics or phonetics, you'll probably find a recurrent diagram for vowels. It uses the two main contrasts (height and frontness) and places vowels in a triangle, like this (corresponding to Spanish or Latin):

HIGH

i u

FRONT e o BACK

a

LOW

Along the *i-u* line are the high vowels, going down to the low vowel *a*, and the front of the mouth is equated to the left side of the triangle. You can place vowels anywhere in the triangle formed by *i-a-u*. The English schwa /@/ (as in a̲live, rode̲nt) is in the middle, right over the *a*; its mid-central. There's a high central vowel *ы* in Russian which would be located in the middle of the line *i-u*. This sound, /ɨ/, is also found in many North American languages and in Guarani (the final *y* in *Paraguay* and *Uruguay* is the Spanish adaptation of this sound, which is a one-phoneme word in Guarani, meaning 'water').

New vowels

As with consonants, you can invent as many vowels as you like. You should take into account that vowels form a system, and one which can't be misbalanced. If you have a tense and a lax version of *i*, then you're using tenseness as a contrast, and it should be present in some other pair of vowels.

Roundedness is not disbalanced in English, or in Spanish. It seems that roundedness is more frequent in back vowels than it is on front vowels. Nevertheless, many languages have rounded front vowels, which English doesn't have (German and French have rounded *i* and *e*, represented *ü, ö* in German). On the other hand, you can have unrounded back vowels (like Japanese *u* or Turkish *ı*).

You can have as many vowels as you want to. The simplest systems have three vowels, generally *i, a, u* (the vertices of the triangle, and not by chance). This means they distinguish three vowel sounds, not that its speakers do not know how to pronounce an *e* or an *o*. A Quechua speaker might say something that sounds *e* to an English speaker, but it's actually an *i*, of which English *e* is just a phonetic, not phonemic, variant. Spanish and Japanese have five vowels, *i e a o u*. Swedish has nine vowels, British RP English has twelve, German has fourteen, and !Xũ (the absolute record) twenty-four. But perhaps you shouldn't go that far.

There are at least three languages with only two vowels: Ubykh, Abkhazian and Abaza, spoken in the Northwest Caucasus (in fact, Ubykh is extinct now, as of 1993). Each of them distinguishes between an open vowel /a/ and a close vowel /@/ (a schwa). Phonemically, that is; it's quite probable that phonetically each of these two is realized in multiple ways according to their position and proximity with different consonants.

Stress and pitch

Stress is of course the strength placed on certain syllable of each word (or of the important words in a complete sentence). Languages can have a regular stress rule, in which case you only have to mention it, or it can be irregularly stressed, in which case you should indicate it. English has an unpredictable stress and it's not marked anywhere; even identical words in writing can have different stress patterns. Spanish has an unpredictable stress too, but it can be read correctly without trouble. In Spanish, an unaccented word receives stress on the penultimate syllable if it ends in a vowel or in *n* or in *s*; if it ends in any other consonant it receives stress in the last syllable; and if it is accented (a vowel is marked with an acute accent as in *álamo*, *adiós*), stress falls in the accented vowel. French words always receive stress in their last syllable. Quechua receives stress in the second to last syllable. Latin stresses the second-to-last syllable if both final syllables are short (short vowels and single consonants, as in *seculus* ['**se**kulus]); else stress falls on the first-to-last syllable (as in *secundus* [se'**kun**dus]).

Pitch is the height of the syllable. Japanese, for example, doesn't use stress, but pitch, to "accent" words. Some syllables are low pitched, and some others are high pitched. The pitch of each syllable is determined by the position of the main pitch drop or accent.

In most languages, some words are not stressed when in a complete sentence. In English, for example, "I'm here for the ad" gets no stress over *I'm, for, the*. (Also, unstressed vowels are reduced to centralized forms, namely a schwa or a weak /I/.)

Tone

Tone is the intonation contour of a syllable. Tone exists in all languages, but it's not phonemic sometimes. In English, you pronounce "What did you do?" (normal) and "What did YOU do?" (emphatic reply) differently, and key words have different tones.

In some languages, tone is phonemic. These languages include Chinese (Mandarin and Cantonese), Vietnamese, and a lot of African languages. Each syllable receives a particular tone, which is as characteristic as the height of the vowels in it, and can distinguish words. Mandarin Chinese, for example, has four tones, called high, rising, low falling, and high falling (you can imagine what they mean). For example: *ma* "mother", *má* "hemp", *mâ* "horse", *mà* "curse". Vietnamese has six tones, two of which include **creaky voice** -- lowering the pitch so much that the individual vibrations of the vocal chords can be heard.

You can try using tones in your language, but I don't recommend it unless your native language is tonal too. It's an interesting device, but it takes quite a lot of self-re-education of the vocal organs. Tone can be a phonemic feature or (rarely in natural languages) a grammatical feature.

There's an interesting short discussion in a work by Marjorie K.M. Chan: "Tone and Melody in Cantonese"[9], positing and answering an interesting question: how do you sing a song in a tonal language?

[9] http://deall.ohio-state.edu/chan.9/articles/bls13.htm

Phonological constraints

Each language has combinations of sounds that are considered difficult, forbidden, or impossible. These are called phonological constraints, and are the moulds into which any word has to be made to fit for the sake of coherence and "familiarity". The rules of syllable- and word-formation are part of what is called **phonotactics** (i. e. which sounds can come in contact with other given sounds).

English is quite free of phonological constraints. Hence the enormous quantity of foreign words it has been able to absorb, like *garage, sombrero, mosquito, ersatz, schmuck*... Some languages do not resist such invasions.

For example, Japanese (one of the most restricted languages) basically allows syllables formed by a (perhaps double) consonant, a vowel (perhaps double), and /n/: (C)V(V)(n). The English word *club* was adapted into Japanese as *kurabu*, to give an extreme example. If you're an anime fan, you know how Japanese anime shows typically employ English (in Sailor Moon, the main character shouted the invocation *muun kurisutaru pawaa akushon* -- that's "moon crystal power action").

Fidjian is almost as much restricted as Japanese: a consonant plus a vowel form a syllable, with an optional consonant at the end of the word.

Finnish didn't tolerate consonants clusters like *pr* or *fl* in not-so-old times. The Elvish language Quenya doesn't tolerate initial or final consonant clusters at all. Greek words can only end in *-s, -n,* or a vowel. Some languages only use certain sounds together with others and never alone.

It's difficult to design a pattern *in abstracto* --but you should have some ideas about it. The main thing is defining whether your language will be vocalic or consonantic, to put it in non-technical and inexact terms. English (and most North European languages) are quite consonantic. Spanish, Japanese and Greek are quite vocalic. Hawai'ian is very vocalic (a word like *Kilauea* is not possible in many languages). The global tendency, according to some theories, is towards the basic consonant-vowel syllabic structure. This is confirmed by the tendency, found in many languages, to simplify the codas -- i. e. to reduce or drop consonants that end a syllable.

A synthetic language with lots of inflections usually prefers a simple structure. (Nevertheless, consider Georgian, a very agglutinating language, where you may find up to six consonants in a row, as in *vprtskvni* "I am peeling it" [*ts* is an affricate, so it counts as one consonant]). An isolating language can have very intricate words, because you won't be adding anything else to them. The best thing is try and try until words begin to look and sound right to your particular taste and mood (just don't change it in midway!).

Sounds tend to influence one another and change. Sound change can ultimately produce a new language, or a distinct dialect.

Sound change

Nobody knows why, but sounds change in all languages. The only languages that don't change are the dead ones.

Sounds change into other sounds, sometimes influenced by others. Sound changes can be classified into conditional and unconditional. An unconditional sound change transformed the Old English *sceadu* /'SkæadU/ into *shadow* /'SædOw/, as well as every word beginning with /sk/ into a new one beginning with /S/ (*sh*) . Most modern English words in /sk/ are Scandinavian borrowings, in case you were wondering. A conditional sound change transformed French *marbre* into English *marble*, the second /r/ being dissimulated by the presence of the first one.

The main types of sound changes are:

- **Assimilation**: a sound "gets nearer" to a neighbouring sound, i. e. takes on some of its phonetic features, especially when this eases the pronunciation. For example *assimilate* from Latin *ad-* + *simul-*; /d/ became /s/ because of the neighbouring /s/. Also *cupboard*, pronounced no more as *cup-board* but as *cubbord*. Assimilation can transform two sounds at the same time: *got you* becoming *gotcha*. Italian got a lot of double consonants from old clusters of two different consonants (e. g. *otto* 'eight' from Latin *octo*).
- **Dissimulation**: the reverse of assimilation, two (identical o similar) sounds move away from each other. For example: the changes from (French?) *marbre* to English *marble*, and Latin *arbor* giving Spanish *árbol*, show /r/→/l/ dissimulation. Nasal dissimulation also changed /mn/ to /mr/ in the process that gave Spanish *hombre* from *homre* ← *homne* ← Latin *hominem*.
- **Metathesis**: two sounds exchange places. This generally produces a new combination which is easier to pronounce (although the term "easier" is quite subjective). For example: Old English *thridda* became English *third*. The name of the Turkish city of Iskenderun shows metathesis too (the original form was Alexandretta -- *aleksand(e)r-* → *(al)iskend(e)r-*).
- **Elision**, **syncope**, **apocope**: all these are names for the same phenomenon. They refer to the loss of sounds; elision specifically means loss of unstressed vowels or syllables, while syncope applies to the loss of medial sounds, and apocope is the loss of final sounds. Examples: *elementary* being pronounced /El@'mEntri/ (elision), in French *au revoir* /or'vwa/; *boatswain* /bOws@n/ (syncope); the loss of final *-e* in English is an apocope, as well as the alternative forms of certain words in Spanish (*grande* 'big', *gran casa* 'big house').
- **Haplology**: the loss of a sequence of sounds because of similarity of neighbouring sounds. In Latin *stipendium* should have been **stipipendium*; *haplology* would have been reduced to **haplogy* if it were a common, non-technical word.
- **Liaison**: introduction of a sound between two other sounds, especially between words. Pronounced /li.'zõ/. French, where the word comes from (meaning 'binding'), is the best example: the final consonants of many words are pronounced only when the next word begins in a vowel. For example *C'est moi* /sE'mwa/ vs. *C'est Anne* /sEt'an/.
- **Prosthesis**: an extra initial sound is added to the beginning of certain words, as in Spanish: *e-* before initial cluster *sp-*, Latin *spectrum* > Spanish *espectro* (Spanish

speakers also add /e/ at the beginning of many English loanwords, such as *escáner, estándar* for *scanner, standard*).

- **Epenthesis**: an extra medial sound is inserted between others. In Welsh, an **epenthetic vowel** appears between certain pairs of consonants in final position; for example *llyfr* pronounced as if it were *llyfyr*. In French, *nombre* 'number' got an epenthetic /b/ (into Latin *numerus*) to bridge the gap between /m/ and /r/.

Conditional and unconditional sound changes are not always easy to take apart. If we take the definition as a strict rule, almost all changes are conditional; very few are absolutely unconditional. For example, the change of Latin /k/ (written *c*) in Romance languages is regarded as unconditional, but it was actually produced by the influence of vowels: Latin /k/ changed into /s/ in Spanish and French (although continued to be written *c*) when the next sound was a front vowel (/e/ or /i/).

Sound change most often produces irregularities. In Spanish, the different forms in which the Latin /k/ changed produced the following forms of the verb *decir* 'to say': *digo* 'I say', *dice* 'He says', *dijo* 'He said', *he dicho* 'I've said'. But one specific type of change can be actually regularizing. It's called analogy, and it will be treated to its own section.

Rules of sound change

Sound changes can be of a lot of different types, as we have seen above. But all kinds of sound change obey some rules:

- Sound change is **grammatically unrestricted**. If a certain phoneme changes into another one, it does not matter the word class. A rule of change that transforms one phoneme or set of phonemes into another can have only phonetic restrictions, for example: 'A changes to B whenever it follows C, except in stressed syllables', or 'intervocalic X changes to YZ'. A rule of change cannot be restricted to certain word classes or grammatical constructions, like 'final A and B are dropped, except on adjectives' or 'X changes to Y on inflected nouns'.
- Sound change **has no memory**. This may sound stupid, but it's not. A rule of change that transforms X into Y cannot discriminate between a certain X that the language has had from the beginning and another X that comes from a previous change W → X. Cycles of sound change are cumulative and each one erases the previous one's tracks, so to speak; imagine waves coming to a sand beach one time after another...
- Sound change is **unstoppable**. Some people used to argue that a written language helps to keep the spoken language from changing. This is obviously untrue. What a written language does is to keep the written words looking as they were before the change. If we learned language from books, the argument would probably be true; but we first learn to speak by listening to other people speaking! If a language doesn't change, it's probably dead. This of course doesn't apply to artificial auxiliary languages such as Esperanto, or to artificially resurrected-and-kept-alive languages like Latin. As for Esperanto, I don't know if Esperantists speak the language at home for their children to hear so that they learn it as a (second) native tongue. If they do, the kids will probably be producing changes very slowly over the years (if they do the same with their own children, and so on). This perhaps would horrify doctor Zamenhof and his followers, but it would be a sure sign that the language is indeed used for communication and is alive, a natural(ized) language among peers. As for Latin, everybody pronounces it more or less as they prefer...

These rules have exceptions, but they must be adequately explained. If you write down the history of your language, you may explain them or use 'for some unknown reason...', but don't let this become an excuse for violating linguistic rules.

Exceptions to the rules are mostly caused by analogy or related processes tending to regularize the language. For example, if a sound change makes X become Y and this makes two pronouns sound the same, one of these things will probably happen: 1) nothing, 2) the pronouns will be merged into one, grammatically as they were phonetically, 3) the pronoun to be changed will 'refuse' to change, 4) people will stop using one of the pronouns, replacing it by another construction.

Also, sound change might be slowed down or sped up. Some people have tried to come up with a set of factors that may cause a language to enter a rapid change phase (such as economic and social chaos, wars, a new religious movement, etc.) These theories have proven useless. There are surely social factors that regulate the speed and quality of sound change, but they depend on so many 'social variables' that they are impossible to calculate. Some you can imagine: if an enclosed country (in an island, for example) suddenly gets in contact with a massive and constant amount of foreign visitors, its language will probably begin to change faster, borrowing new words and structures, creating or copying new idioms, and inventing new words for concepts they had no previous knowledge of.

Another cause for exceptions is the fact that some words are less common than others. Words may change if they are said and repeated over and over, thus being "worn out"; strange, rarely used words, are likely to stay unchanged. These rarely used words usually include educated terms, or very formal or specific words. Sometimes they are not exactly preserved, but **re-borrowed** from the ancient language (or another one), like English *foreign*, which comes from Proto-Indo-European **dhwor-*, hence also *door*; or *semaphore*, where *-phore* "carry" has the same origin **bhero-* as the verb *to bear*. Other examples include pairs of related words like night-nocturnal, virile-werewolf, blanch-blank, etc.

Harmony and Mutation

Harmony is a set of sound changes that some languages produce in parts of speech on certain occasions. Although simple, it can be considered a different type of sound change, related to the assimilation process.

One type is called **vowel harmony**. It produces changes on vowels, according to other vowels in the same word. Vowel harmony is present in Turkish, the Finno-Ugric languages (such as Hungarian and Finnish) and some Native American languages. These have in common the fact that they are agglutinating, so the root of the word may be followed by a lot of suffixes or come after a string of prefixes, which are concatenated (agglutinated). The stressed vowel in the root (which is usually the first or the last one, depending on whether you use suffixes or prefixes) is categorized according to a certain contrast, usually the place of articulation. So you may have, for example, vowels divided into front (*i, e*, German *ä, ö, ü*) and back (*a, o, u*). Then you change all the vowels in the agglutinated affixes to match the quality of the root vowel. In this way, each affix has to have two forms, a front form and a back form. (Some languages may have three or four steps in the scale instead of just two.) For example, take a look a some Finnish words with case marks:

autossa 'in the car'
laatikossa 'in the box'
järvessä 'in the lake'

Do you see how the final vowel alternates between **-a** (back) and **-ä** (front)? Some more examples, with the perfect tense of verbs:

on lyönyt 'has beaten'
on ajanut 'has driven'

The perfect tense mark is *-nut* for roots with back vowels, *-nyt* for roots with front vowels (*y* = /y/, like German *ü*).

I have a language with vowel harmony of my own: Knarwaz. Compare the following words: back vowel *gnolpusut* 'in the mountain' vs. front vowel *lempüsüt* 'in the tree'. The first syllables (*gnol-*, *lem-*) are the roots, while the endings show locative case and masculine gender. The form *-pusut* uses the back vowel /u/ because the root vowel /o/ is a back vowel. The form *-püsüt* uses **ü** = /y/ (rounded **i** or front **u**) because the root vowel /e/ is a front vowel.

Vowel harmony can also be extended to other contrasts besides place of articulation; it could include length, nasalization or roundedness, too. Vowel height harmony is also possible, but it isn't found in any known natural language.

Another form of harmony is called **nasal harmony**. It's found on Guarani (the language of a South American native group which inhabited in North-eastern Argentina and Paraguay, where it's still spoken by many people and has formed a pidgin). I don't know of any other language featuring nasal harmony, but again I didn't go researching. Nasal harmony 'turns on' nasalization in certain consonants of the agglutinated affixes (yes, Guarani is also agglutinating) when the root of the word contains nasal consonants. So many affixes have two forms, a nasal one and a non-nasal one. For example, from *hecha* 'see' we can form *jajoechapeve* 'until we see (each other)'. This is non-nasal. But from *hendu* 'hear', we must

say *ñañoendumeve* 'until we hear (from each other)', where **ñ** is the palatalized **n** also found in Spanish (almost like /nj/). See the change? Non-nasal palatal **j** changes to nasal palatal **ñ**, and also non-nasal labial **p** (in *-peve*) changes to nasal labial **m** (*-meve*).

You can have other types of harmony in your language. For example, a kind of '**inverse harmony**' where two consecutive syllables cannot have the same vowel, or cannot begin by a certain consonant cluster. This is closely related to the phenomenon of dissimulation, only that it's systematic, not accidental. Greek provides an example of this: when deriving words from their roots, there can't be two fricative sounds beginning consecutive syllables; it there are, the first one becomes a stop. For example, the root *thrikh-* 'hair' gives *trikhós* (instead of the expected ***thrikhós*). (Greek also produces a lot of assimilation.)

Sandhi or mutation

Sandhi is the name given by the ancient Sanskrit scholars to a regular set of sound changes which are produced on words on certain conditions. It can be also called **mutation**. These changes can be of several forms. I will mention one, the one I'm most familiarized with: lenition.

Lenition or **softening** is a change produced on the initial sounds of words whenever they are used in certain positions, or for certain purposes. These changes affect the beginning of words by removing, adding or changing initial sounds. In that way, words can have two or more forms.

Of the Western languages I know something of, Welsh and Irish have lenition patterns. Welsh, in fact, inspired the phonology of the famous Sindarin language invented by J. R. R. Tolkien for the Grey Elves of Middle-Earth. I don't know much Welsh, but I happen to have some material on Sindarin, which has lenition patterns taken from Welsh. So I'll use Sindarin for the examples.

Sindarin lenition affects the initial consonants of words in certain contexts. A lenited consonant changes this way: the voiceless stops *p, t, k* become voiced *b, d, g*. The voiced stops become fricatives, except for *g*: *b, d, g* change to *v*, *dh* (/ð/), and nothing. Voiceless *lh* and *rh* become voiced *l, r*; *s* gives *h*, and *m* gives *v*.

In Sindarin, a word is lenited when it is ***(a)*** the object of a verb and is next to it, ***(b)*** anything after conjunctions and articles, ***(c)*** an adjective following the noun it describes, and ***(d)*** the second element of a compound. For example: from *certh* 'rune' we have *i gerth* 'the rune'; from *peth* 'word' the magic spell *Lasto beth lammen* 'listen to the word of my tongue'; from *calen* 'green' the name *Tol Galen* 'Green Island'; from *mellyn* 'friends' the name *Elvellyn* 'Elf-Friends'.

Welsh mutation patterns are quite more complicated than that; there are three types of mutation, called soft (lenition), nasal, and spirant mutation. Welsh also features a related phenomenon involving verb conjugation (at least for the verb *bod* 'to be') where interrogative and negative forms, besides changing intonation and/or using particles, produce a change in the initial sounds.

You can use other types of lenition and consonant mutation, and specify when they should be used. In the African language Ful, a personal-class noun is lenited when it's pluralized;

singular *jim* 'mate', plural *yim'be* 'mates', with lenition *j* → *y*. Curiously, thing-class nouns are lenited exactly the opposite way.

Writing your language

Once you have determined which sounds your language will have, you'll need a way to write them down in the Roman alphabet (**transliterate** them), and perhaps an alphabet of its own. We'll talk about alphabets in a minute.

Transliteration can be a nightmare. The ideal thing would be having one symbol for each sound, but the Roman alphabet doesn't have symbols to represent some very common sounds. Here you have your first choice: will you invent or use one symbol for each sound, or use some other devices? If you want one symbol for each sound, then you'll probably have to use either non-letter symbols (such as *′ @ ?*) or resort to **diacritic marks**, i. e. modify letter symbols by using little signs on top of (or below) them. The accents and diereses over vowels are diacritic marks: *á è î ÿ*. English doesn't use any diacritic marks. Spanish shows some stressed vowels with an acute accent: *acá éramos ínfimos órganos súbitos*, and writes the palatalized nasal sound as *ñ* (as in *año*). French uses accents to show that a written *e* should be pronounced and for the sake of tradition in many words: *été âme à mère*; and it has a letter *ç* for /s/ before *a, o, u*. Portuguese shows nasalized vowels with a tilde (~) over them (as in *são*). German shows front versions of back vowels with a diereses over them (*ö ü*). Danish writes a kind of rounded *a* with *å*, and a fronted *o* with *ø*. Many languages have non-standard letters for certain sounds, and unless you speak those languages and your keyboard is configured for them, you won't be able to easily access to them when writing your language in your computer.

If you don't want to use so many strange symbols, you'll probably have to use two or more symbols to represent some sounds, like English uses *sh* and *th* for single sounds. These are called **digraphs** (trigraphs are possible but to be avoided for the sake of length). The letter *h* is very good for digraphs. But you have to take something into account: two symbols should never be used to form a digraph if they can appear on their own to represent two different sounds. English can use *th* because the cluster /t/+/h/ does not appear in English, but couldn't use *sn* to represent a nasal fricative, because some words have *sn* with the value of /sn/.

Transliteration has no rules on which symbols you use to represent which sound, but you should try to make the language readable: it's OK to use *zh* to represent /f/, but most people will surely read something completely different from /f/ when they find it, and besides, you already have a more familiar *f* to fill that place, right?

Transliteration should be as phonemic as possible. English is a bad example; words are written the way they were pronounced centuries ago, so the written and spoken forms of a word are usually inconsistent. French is even worse (in a word like *oiseau*, pronounced /wa'zo/, there's not one sound corresponding to its 'proper' letter). Written Spanish and Italian are quite phonemic, and almost as much important, the sounds can be guessed from the written form, although inaccurate. Some languages are remarkably consistent in their written forms.

Alphabets and other scripts

An **alphabet** is a collection of symbols representing sounds. You can invent an alphabet for your language if you want to. If you do, and your romanized spelling is phonemic, then your alphabet should be too: one symbol for one sound. You can use digraphs and add diacritics to your own alphabet. If your language derives from another language for which you already

had an alphabet, then probably the newest language will use the old alphabet, but some letters will have changed sound. For example, Spanish uses the Latin alphabet, but the letter *c* now represents /s/ before *e, i*. This is not phonemic spelling, but the change is completely regular.

When inventing letters, play around with them and write them quickly one after another. People write carelessly in most cases, and elaborate letters are likely to be simplified. Also try to make each letter different from all others, so that they are not confused. When two symbols look very similar, people find ways to distinguish them. The dot over the *i* appeared when the little stick of the lowercase *i* began to be confused with the vertical lines of *m*'s and *n*'s in Gothic handwriting. Computer fonts and programmers distinguish *0* (zero) and *O* (the letter *o*) by writing a slash over the zero.

You have to decide how you will read and write. Will it be from left to right, like the Roman and Cyrillic alphabets are usually written? Hebrew and Arabic are written from right to left, and vowels are not written except in children's books and (Arabic) in the Koran. Japanese is usually written from top to bottom and from right to left, but it's written from left to right in certain books, like mathematics ones.

Alphabets are not the only kind of writing. Chinese uses **ideograms**, or characters which used to represent a picture of an object. Each character represents a concept and is read as a syllable; but words that sound the same and are not related are written as different characters. Chinese characters have two parts, the **radical** and the **phonetic**. The radical gives an idea of the meaning, while the phonetic gives an idea of the sound; a radical can sometimes act as a phonetic and vice versa.

Japanese uses a mixed system of ***kanji*** (ideograms) and ***kana*** (phonetic syllabic characters). In general, the main content of what you're trying to say is written in *kanji*, while particles, conjunctions and inflectional endings are written in *kana*. There are about 90 *kana* divided into two sets (***hiragana*** and ***katakana***). *Hiragana* are most often used for original Japanese words; *katakana* are preferred for borrowed words, and also to add emphasis, just like italics in the Roman alphabet. Also, when an unusual *kanji* is used, it can be clarified by spelling it phonetically in *hiragana*, which are called ***furigana*** ('handicap *kana*'). You can change the quality of the consonant in a *kana* by using some diacritic marks. There are 1945 'standard' *kanji*, of which 1006 are taught in elementary school, and each *kanji* can be read according to its Japanese pronunciation (*kun-yomi*) or its original Chinese pronunciation (*on-yomi*). As if it weren't confusing already, each *kanji* can have several readings of each of the two forms.

Korean uses an alphabet called Hangul (or Hangeul), which is a **featural code**, a system in which similar sounds are represented by similar symbols. I don't know when this was originated, but it requires a remarkable phonetic analysis. In Hangul, symbols are grouped in syllables, making the writing look as if it was composed of many ideograms or syllabic characters, which is not the case.

Arabic uses a **cursive alphabet**, which is unusual because most peoples in history have started out with block letters, due to the nature of the material support for writing. Arabic was written with fine brushes on some kind of smooth surface from the beginning, I guess; cursive letters are completely inadequate for (quick) stone carving or clay.

Thai, while a syllabic language, uses a **phonetic alphabet** of single letters, which often have little curls and twists at the ends. Some other scripts of peoples in that area of the globe use that kind of characters which seem a bit too much elaborate. The reason is that they were first written using materials which required lines to be 'closed' in some way.

This all boils down to a principle: to invent an alphabet, you must know where it's going to be written and by what means.

Inventing an alphabet is simple, but a syllabary (or ideograms) can be a headache, so you should think of it carefully before. Ideograms are probably the worst kind of writing, and you should probably refrain from using them unless you have a photographic memory. Syllabaries are fine, but they work best on very restricted languages; English has an enormous number of possible syllables, and inventing a sign for each one would be impossible.

Take a look at some natural language scripts in Ancient Scripts[10], a page with examples from all around the world.

Ordering your script

We're used to have our letters in order. This is very useful for dictionaries and phone books, and for indexes in general. How are you going to order your symbols?

Western alphabets derived from the Roman alphabet usually follow a predictable order. English uses a relatively small set of symbols, and digraphs aren't considered independent symbols, but this is not so in other languages. For example:

- The Spanish alphabet consists of all the letters in the English alphabet, plus the following: *ch* (which goes after *c*), *ll* (after *l*), and *ñ* (after *n*). So you won't find a word like *chico* under the *C* chapter. Does your language use a Latin-derived script? What extra symbols do you have, and which of them are given their own place in the ordered alphabet.
- Finnish alphabetizes the umlauted vowels *ä* and *ö* after the letter *y*.
- In Dutch, the digraph *ij* is sometimes still considered one symbol. (Older typewriters have a key for it!)
- In Swedish, *v* and *w* are considered two versions of the same letter, so they fall into the *V* chapter of alphabetic lists. This causes great trouble given the many English and German words with *w* that have been borrowed into Swedish (which only uses *v* for native words).

Some other languages, using non-Latin scripts, order their characters in different fashion. Some of them use the phonetic features of sounds to order the letters; for example, first the labials (*p, b, m, f*), then the alveolars (*t, d, n, s*) and so on.

As for syllabaries, there's usually also a fixed order. In Japanese, both types of *kana* are arranged like this: first the vowels, *a i u e o*, then the syllables beginning with *k* (*ka, ki, ku, ke, ko*), then *t-, n-, h-, m-, y-, r-, w-,* and finally the symbol for syllabic *n*. Another order, more traditional, was used in former times (and is still used in indexes and tables, as opposed

[10] http://rabbitmoon.home.mindspring.com/asw

to the modern order, which is used in dictionaries). This order follows a poem by Buddhist monk Kuukai, which uses each character of hiragana exactly once:

Iro ha nihohe to
chirinuru wo
waka yo tare so.
Tsune naramu
uwi no okuyama
kefu koete
asaki yume
mishi wehi mo sesu.

(Note: this is probably not good modern Japanese, nor is this the correct pronunciation. The kana for *ha* is pronounced *wa*, and the *kana* for *wi* and *we* are obsolete. The *kana* for *wo* is pronounced *o*.)

As for ideograms, Japanese *kanji* (and Chinese *hanzi*) are ordered by the radical number and, within the same radical, by the number of strokes needed to write the character (there's a method to count them properly).

It would be a nice idea to have letters with names that mean something, or that can be recited in order. Latin letters have meaningless names in all languages that use them, and their names are often too similar to one another, hence the need for codes like 'Alpha, Bravo, Charlie'... Other languages and scripts don't have such problems.

Grammar

This section will take some grammar issues and develop them, showing with examples, when possible, how natural languages manage them, and what can you do about them. You can't have a language without a grammar; if you don't think about it, you'll probably copy the structures of your own language, and the whole thing will be an exercise of translation of single words.

Morphological typology

The classic categorisation is that languages can be **inflecting**, **agglutinating**, or **isolating**. This categorisation has proven to be too limited, but I'll explain it, because it's a good starting point to understand the differences.

Inflection

An inflecting language uses inflections, which may be affixes used, for example, to conjugate verbs, decline nouns and other tasks. Some languages use suffixes for this purposes, while others use prefixes; most use both, though there's usually a preference. A few languages employ infixes or circumfixes. Examples of inflection in English are the *-s* used for pluralizing names and the *-ed* used to form the past of regular verbs.

Another type of inflection (and "purer", if you like) is the change of the root forms of words. Examples are the inflection of strong verbs of English, like *sing/sang/sung*, which are inflected forms of a root concept "sing". Inflection by vowel change (called **ablaut**) is quite usual in certain languages. Consonant change does exist, but it's rarer. Curious examples in English are the pairs *breath/breathe* (changes voiceless to voiced *th*, besides vowel change), *house* (noun) vs. *to house* (verb) (same change).

Inflection includes some other devices like changing suprasegmental features like tone, stress or pitch; lengthening a vowel or geminating a consonant; and repeating a part of the root (reduplication). The main thing about inflections, however, is that an inflection can carry more than one meaning at the same time. For example, in Spanish *viví* "I lived", the inflection *-í* shows that the verb is in the past tense, first person singular, indicative mood. Examples of inflecting languages are English, Spanish, German, Latin, Greek, and in general all Indo-European languages.

Agglutination

An agglutinating language uses suffixes or prefixes whose meaning is unique, and which are concatenated one after another without overlap. Some known agglutinating languages are Quechua and many other American languages, Turkish, Finnish, and Hungarian. For example, in the Quechua word *wasikunapi* "in the houses", the plural suffix *-kuna* is separate from the locative case suffix *-pi*. In Finnish, *huoneissansakaan* means "(not) even in their rooms", and it consists of five agglutinated morphemes, "room-s-in-their-even".

Isolation

An isolating language doesn't use affixes or root modifications at all. Each word is invariable, and meanings have to be modified by inserting additional words, or understood by

context. The best known example of isolating language is Chinese. In Chinese, a noun by itself is not singular, nor plural; and a verb has no tense or person; these distinctions are made by adding quantifiers, adverbs, or pronouns. In effect you say "books" by saying "several book".

Analysis and synthesis

The modern classification of language grammars is a continuous scale which goes from **analytic** to **synthetic**. The more analytic a language, the more meaningless the words by themselves, so as to say, and the more important is context and word order (analysis is thus roughly equivalent to isolation). The more synthetic a language, the more self-contained the words (synthesis involves inflection or agglutination).

The scale is meant to be taken as a reference; there are no extreme points, but you can compare two languages and say that one is more synthetic than the other. Chinese is very analytic; a Chinese word by itself can mean a lot of different things, because no distinctions are made in it: you don't know if it's a verb, a noun, an adjective, or if it's past tense or future, or plural, or singular, or anything, you only have the root concept. Some Native American languages like Nootka or Chinook are the other end, so synthetic that indeed they were called **polysynthetic**, inflecting words in such ways that a single word can mean "the many little fires been lit in the house in the past" (I'm not making this up; the word is *inikwihl'minih'isit*, and by the way, it's not properly a verb or a noun; it needs verbal or noun prefixes...). In the middle, we have Japanese (quite analytic except for verbs), English (quite analytic too, as it barely distinguishes noun case or verbal person), Spanish, French and Italian (of the ones I know a bit of), German (already with many inflections) and all the agglutinating languages, which are in fact a subset of inflecting languages, Latin, Greek, Sanskrit...

So you'll have to pick up a point in the scale and stay there. This is probably the most important decision in the process. Each kind of grammar has its own pros and cons.

- An **isolating** language avoids a lot of work on difficult fields like deciding how to pluralize nouns and conjugating verbs. But it requires that you plan a rigid word order for sentences, and respect it at whatever cost, after assuring that it can't lead to ambiguities (serious ones at least). And a totally isolating language is difficult to devise, because you have to eliminate all traces of inflection, even ones that you'd never suspect about.
- An **agglutinating** language means a careful planning of affixes (dozens of them) which must have unique meanings. Also, you must decide in which order they will appear after or before a word. Finally, agglutinating languages may tend to produce very long words, or ones that are very difficult to pronounce (consider Georgian, where many affixes are formed by just one or two consonants; sometimes they have to be joined to other affixes of the same kind, so you might end up with six consonants in a row).
- An **inflecting** language produces shorter words and compact sentences (the more inflecting the language, the more compact the sentences), but it requires that you plan all inflections and combinations of inflections, because sometimes you won't be able to place two or more of them in a row (agglutinated). You can take inflection to its simplest expression (as in English) or produce a polysynthetic language which inflects words for almost every conceivable purpose. The more inflected a language, the more you'll have to care about concordance (the agreement of adjectives and nouns, and nouns and verbs).

Sapir's Classification

There's another classification of languages, which is far more complex, and was created by Edward Sapir in the 1920s. This divides concepts into four classes:

Group I. Basic (concrete) concepts (objects, actions, qualities): normally expressed by independent words or radical elements; they don't include any kind of relationship with other words. For example, English nouns and adjectives like *dog, party, ugly, strange*.

Group II. Derivative concepts (generally less concrete than those in group I): normally expressed by affixation of non-radical elements to radicals, o by internal modification inside these. They denote ideas that don't have to do with the proposition (sentence) itself, but give the radical element a certain particular twist of meaning and are therefore intimately related to it in a concrete fashion. For example, English prefixes *pre-, for-, un-* and suffixes *-less, -ly*.

Group III. Concrete relationship concepts (yet more abstract): normally expressed by affixation or internal modification, but commonly in a less intimate fashion than group-II elements. They indicate relationships that go beyond the word itself. For example, English *-s* for plural nouns.

Group IV. Pure relationship concepts (totally abstract): expressed by affixation or internal modification of radical elements, or by independent words, or by word order within the sentence. They connect the concrete elements of the proposition, giving them a definite syntactic form. For example, the modifications of English *him, her* from *he, she* indicating accusative case; the prepositions *to, for*; the position of *the dog* in *I see the dog* indicating that it's the object of the verb, etc.

The classification of languages according to these classes is as follows:

Type A. Languages which only express concepts of groups I and IV, so that they have no means of modifying the meaning of the radical element by means of affixes or internal changes. For example, Chinese.

Type B. Languages which express concepts of groups I, II and IV, preserving pure syntactic relationships and being able to modify the meaning of radical elements by affixation or internal change.

Type C. Languages which express concepts of groups I and III, where syntactic relationships are expressed in necessary connection to barely concrete concepts, but they can't change the radical elements by affixation or internal change.

Type D. Languages which express concepts of groups I, II and III, i. e. where syntactic relationships are expressed in mixed ways, like in Type C, and can also modify the meaning of radical elements by affixation or internal change. In this group belong most of the "flexive" (inflectional) languages with which we are familiar, as well as many "agglutinating" languages.

Each one of the types A, B, C, D can be subdivided into **agglutinating**, **fusional** and **symbolic**. Agglutination means the things added to the radical element are just juxtaposed

(put together); fusional means they are sometimes merged; symbolism roughly means internal change. Type A also has an **isolating** subtype.

The method (agglutinating, fusional, or symbolic) for a certain group of concepts needn't be identical to the method for a different group. The classification uses a compound term, the first part referring to the method for group II concepts, and the second part to concepts in groups III and IV. These methods are sometimes not alone; English uses them all. For example, *goodness* from *good* is agglutination; *books* from *book* is regular fusion, *depth* from *deep* is irregular fusion, and *geese* from *goose* is symbolic fusion or symbolism.

All this rant is just about one thing: you don't have to expect everything must be in its "proper" place in your language (the proper place being that of English). English number (singular vs. plural) is a Group III concept, quite abstract and forming part of the very core of words; we can't conceive an English noun without number. In Tibetan, number is an optional feature and it's not grammatisized as in English; it's not an abstract thing that belongs into the word, but a concrete thing: the idea of plurality, "several" or "many", is expressed by a radical element which is a separate full-fledged word, a Group I concept. It's not syntactic and can therefore be omitted when not needed.

Think hard about this! After you place your language on the scale, you have to decide which word classes you'll use, and how they'll link to one another.

Nouns

Number

Number is not restricted to singular vs. plural; many languages have forms for pairs of things (dual) and some for groups of three things (trial). Others have a paucal number (from the same root as *paucity*, meaning 'few'), that is used for items up to a certain approximate quantity (such as three or four), resorting to the plural for higher quantities.

You can have a singular number which refers to a unique object, or two plurals distinguishing the things at view ('these men') and all the things of the stated kind ('men')... Your imagination is the only limit.

You can however simply leave number out of your system. This is what Mandarin Chinese and Japanese do. You can have a particle or an adjective with the meaning of 'several' or 'many' to express the idea of plurality when needed, if context is not enough to make it clear.

If you use an inflection for plural number, be aware that it doesn't have to be a short suffix; it can be quite long (like the two-syllable Quechua *-kuna*) or be a prefix, or an infix, or it can appear as vowel change (e. g. umlaut or ablaut). Many languages show plurals of some kinds of items by reduplication, which means repeating the whole word, or the first syllable, or the last syllable, etc. In Bahasa Indonesia you have *baterei-baterei* 'batteries' (this is from the multilingual manual of a calculator!); in Japanese you have *hitobito* 'people' from a slightly modified reduplication of *hito* 'person'.

English irregular plurals of the kind *man/men*, *goose/geese*, *mouse/mice* are examples of vowel gradation, which resulted from umlaut, in turn produced by a suffixed inflection that was lost. Other languages are much more regular, like Spanish (which always marks plural with *-s, -es*).

Gender

Gender is the common term for the more general concept of class. Gender need not be feminine vs. masculine. German, Greek and Latin have the genders feminine/masculine/neuter. Swahili has noun classes ('genders') for animals, for human beings, for abstract nouns, etc. Many languages make a distinction based on animacy, between animate and inanimate objects (people and animals vs. plants and non-living objects, or the like). You can invent new distinctions.

Noun classes can be more or less arbitrary. In Indo-European languages there is usually no relationship between the gender and the actual object. While the Spanish noun *mesa* 'tabla' belongs to the feminine gender, not only is it unrelated to femininity, but also has nothing in common with most other feminine nouns, like *comadreja* 'weasel' or *crisis* 'crisis'... The animate/inanimate distinction tends to be less arbitrary, but there are always borderline cases and particular cultural influences (for example, some languages may take 'fire' to be an animate noun). When there are many classes with semantic content (as in Bantu languages) it may happen that some nouns change meanings but stay in the same class (suppose you have a class for round objects and another for square things, and the word for 'ring' comes to mean 'boxing playfield', as in English...).

Case

In a broader sense, grammatical case is the role of the noun in the sentence (for example, subject, object, complement of place, etc.). In the restricted sense which we'll refer to from now on, a case is some morphological mark of that role, usually shown by inflection or agglutination.

There is no fixed set of cases; each language distinguishes one or more morphologically-marked cases and uses them for given purposes. However, some common cases found in many languages are always given the same names.

Latin has the following inflected cases: nominative, accusative, genitive, ablative, dative, and vocative. A noun is in the nominative case when it's the subject of a sentence; accusative when it's a direct object; dative when it's an indirect object; genitive when it's a possessive; ablative when it's part of a verbal complement; and vocative when it shows a call (plus many, many special cases). English actually has a genitive case, marked by the possessive ending *-'s*, and distinguishes nominative and accusative forms of pronouns (*we-us, I-me, they-them*, etc.).

Certain cases are used after certain prepositions (the preposition is said to **govern** the case). My language Terbian has a core case (used for subjects and objects, which are further distinguished by other marks) and an oblique case (used as a genitive or compounding case, and with all postpositions). Romance languages have mostly lost the Latin case system altogether, and resort to prepositions and word order to show syntactic roles. Your language can have many cases; Estonian has 14 cases, and Finnish even more (18, according to some analyses). There are many syntactic roles that can be codified by a case, but these tend to overlap, and the majority are local cases (used to convey relationships of position and movement -- on, over, under, around, inside, outside, at a side, from, towards, into, out of, etc.).

Adjectives

With adjectives, we enter the land of possibilities. You can choose to have adjectives (as a separate word class), or not. Adjectives can be an entirely different word class, as in English; or they can be a subset of nouns (considering morphology and behaviour), as in Spanish or Latin; or they can behave like verbs (as some do in Japanese). Let's examine these alternatives.

If adjectives are a completely different word class, then they don't have to behave like anything else; they can have their own rules of inflection, or not inflect at all. English adjectives are an example of this: they are invariable words (except for the comparative and superlative forms).

If adjectives are like nouns, or a subset of nouns, then they behave like nouns. In Spanish, where nouns have gender and number, adjectives have them too, and they must agree with their head noun. Sometimes they can become nouns without any change; *rojas* means both 'red' (feminine and plural) and 'red ones' (when preceded by an article). Curiously, nouns can become adjectives, in colloquial sentences like *¡Es tan payaso!* 'He's so (much of a) clown!'. In Latin, adjectives agree with their head noun even in case. But the distinction between nouns and adjectives is usually well-defined in these languages; some other languages may choose not to make it.

In Japanese, adjectives of a particular class (*na*-adjectives) behave like nouns; they are placed before the noun they modify, followed by *na*, which is the relative form of the copula 'to be'. For example: *kirei na kimono* 'beautiful kimono' -- the nominal adjective (or qualitative noun, as some people call it) *kirei* means 'beauty' or 'beautiful', and the phrase could be translated as 'kimono which is beautiful / which has beauty'. You can add tense to the adjective by marking tense on the copula: *kirei datta kimono* 'kimono which was beautiful'.

If adjectives are like verbs, then they conjugate like verbs. Another class of Japanese adjectives (*i*-adjectives, because they end in *-i*) work this way; adjectives are usually a kind of participial form of verbs, or a single-word relative clause (relative clauses in Japanese come before the noun phrase they modify, the same as adjectives and demonstratives do). You can think of Japanese adjectives as a combination of an English adjective + the copula 'to be', though Japanese adjectives can and do take the copula sometimes. But the tense is still on the adjective, not on the copula. For example: *Kakkoii desu* 'He is cute' (polite form); *Kakkoikatta desu* 'He was cute'. Here *kakkoi-* is the root, while *-i* is the suffix for adjectives in present tense, *-katta* is for past tense, and *desu* is the polite *present* tense form of the copula. As you see, the tense in this class goes directly on the adjective, not on the copula, which can be omitted sometimes.

In my own language Draseléq[11], adjectives do not exist as such. There are verbs that mean 'to be big', 'to be yellow', and even 'to be four'. You say 'a tall tree' by saying 'talling/talled tree', using a short participle. You say 'the tree is tall' by using the third person singular present tense of the verb 'to be tall' with 'the tree' as the subject: 'the tree *talls'. The best thing about this is that you merge two word classes into one, and you can use whatever devices you invented for one on the other. In Draseléq, you can express the equivalent of 'make/cause to be four' in one word.

[11] http://www.angelfire.com/ego/pdf/ng/lng/draseleq/index.html

Many adjectives may not exist at all in any form (although every language has some words that act like adjectives). The ideas of qualifying can be expressed in other ways. Tibetan uses abstract nouns instead of adjectives; you don't have the adjective 'large', but the noun 'magnitude, largeness', and you can express 'a large room' by saying 'a room of magnitude'. This is not ridiculous in English. 'A room of magnitude' is rare but possible, and 'a disaster of biblical proportions' (which follows the same structure) is common.

In some languages, the adjectives form a closed word class (like prepositions in English); there are a certain number of them (pairs like 'big'/'small' and the colours) and others can't be formed.

If you have a morphologically separate word class for adjectives, you should also invent some affixes to colour their meaning, to negate them, and to transform them into other word classes. Also think of comparatives and superlatives. It's not an obligation to have them, but a language should be able to express such ideas as something being taller, or redder, or uglier, than something else.

As an extra, you can read a compilation of a thread in the Conlang list[12], started by a question by Fredrik Ekman: are there languages without adjectives[13]?

[12] http://www.angelfire.com/ego/pdf/ng/lng/conlanglist.html
[13] http://www.angelfire.com/ego/pdf/ng/lng/languages-without-adjectives.txt

Verbs

Person and number

In many languages, the verb agrees with one of its arguments (one of the noun phrases in the sentence); in languages that mark subject vs. object, generally the subject. However, some languages have double agreement (Hungarian verbs agree with both the subject and the object), which is a form of polypersonal agreement (Basque verbs agree with subject, direct object and indirect object when applicable!). The verb usually agress with the noun phrase in one particular case (nominative in nominative/accusative languages, absolutive in ergative/absolutive ones).

In quite a few languages, there's no agreement at all: English barely distinguishes the third person singular from the rest in the present tense; Mandarin Chinese and Japanese don't mark person in the verb in any way.

Tense

The tense system can be anything from a distinction between present and non-present actions to a complex structure. The only universal tense is present. Many languages don't have a real future tense and employ a past/non-past distinction that conflates present and future. English actually doesn't have a morphological future tense, since futurity is modelled by an auxiliary, *will*, not by inflecting the verb. For the sake of generality we'll call this a tense (a **periphrastic** one).

You can have several types of present or past or future. Spanish has two different pasts; one shows actions that took place over a period of time in the past (imperfect), and the other shows that things just happened. That's more or less the difference between English *I lived* and *I used to live*.

Some languages do not distinguish tense, using adverbs of time or suggesting a temporal frame by other means (like aspect marks) when necessary.

Aspect

From Richard Harrison's Invisible Lighthouse[14] (: Aspect refers to the internal temporal constituency of an event, or the manner in which a verb's action is distributed through the time-space continuum. Tense, on the other hand, points out the location of an event in the continuum of events. In many traditional grammar descriptions, tense and aspect (as well as mood) are conflated together; for example, English has what is called 'present perfect tense', which is in fact a present tense with a perfective aspect.

Verbs can inflect to show that the focus is on the ongoing process (progressive), or a single action (punctual), or a habitual action, or a repeated action (iterative), or the beginning of an action (inchoative, inceptive), or the ending of an action (cessative), etc. Some languages have literally dozens of these aspects. An interesting pair is the distinction between **static** and **dynamic**. A static form describes a particular state, while a dynamic form reports a change in state. In Arabic, *rukubun* means 'ride' in its static forms, and 'mount' in its dynamic forms.

[14] http://www.invisiblelighthouse.com/langlab/index.html

Japanese has a conditional aspect: it can inflect verbs to show conditional clauses, so for *taberu* 'eat' there's *tabetara* 'if/once I eat' and *tabereba* 'if I eat'.

Perfectiveness

Perfectiveness is an aspectual distinction. In grammar descriptions, perfect means 'completed' (referring to the verbal action). *I have come* is perfect (or has a perfective aspect) while *I'm coming* is imperfect. The Spanish example above is an aspect opposition.

Mood

Mood refers to whether the action is real and certain (**indicative**), or is doubtful or desired (**subjunctive**), or isn't happening at all (**negative**), etc. etc. The indicative mood (it just happens) is the most common.

English doesn't distinguish indicative and subjunctive (it uses past forms of indicative mood to show the subjunctive), and it uses an auxiliary to negate a verb. In Spanish and other Romance languages, the subjunctive mood is used (among other things) for hypothetical actions and for wishing formulae: *si pudieras* 'if you could'; *ojalá pudieras* 'wish you could'.

Japanese inflects verbs to negate them (*keru* 'I kick', *keranai* 'I don't kick'), while Finnish uses inflected forms of an auxiliary (*ei*) before a form of the main verb (much like English auxiliaries *don't, doesn't*).

There's also the **imperative** mood, which is used to give orders or make requests. These moods, of course, are not the only ones. Nenets, a Siberian (Uralic/Samoyedic) language, has a lot of moods (some of which I would've taken as aspects!): indicative, imperative, hortative ('Let me'), optative ('Let him'), conjunctive ('He will' [request]), necessitative ('He must'), interrogative ('Did he?'), probabilitative ('He may'), obligative ('He should'), approximative ('He seems to'), superprobabilitative ('He probably'), hyperprobabilitive ('He must have'), reputative ('He is supposed to'), Habitive ('He is used to').

Evidentiality

Refers to the kind of evidence that the speaker has about what he or she's saying (does he know about the action from personal experience, or just by hearsay, or just believes it likely?). Quechua, Aymara and many other Native American languages distinguish these aspects with different levels of subtlety. You may have heard of it as 'levels of experience', or 'trivalent logic' (i. e. not only consisting of 'true' and 'false' statements but also of 'maybe' statements).

Argument structure

The **arguments** of a verb are the parts of the sentence (generally noun phrases) that it joins and that it has a close grammatical relationship with. In general this means the subject and (if present) a direct object and maybe also an indirect object.

The number of arguments of a verb is called its **valency** of the verb (by analogy with the valency of chemical elements, which is the quantity of atoms of other elements that can be joined to one atom of the element).

Valency	Verb type	Example
0	impersonal	*none in English*
1	intransitive	"he runs"
2	transitive	"she ate lettuce"
3	ditransitive	"we gave presents to them

So-called impersonal verbs (with valency=0) have no arguments, not even a subject. In English all verbs must have at least a dummy 'it' to fill the subject slot (as in 'it rains'), but e. g. in Spanish the equivalent form *llueve* is impersonal (it appears in the third person singular form, but does not and cannot have a explicit subject).

Most languages do not morphologically distinguish transitive and intransitive verbs, but e. g. Hungarian does (transitive verbs have different person/number inflectional endings than intransitive ones, i. e. different paradigms).

Some intransitive verbs are semantically reflexive; i. e. there's an implied object that is identical to the subject. Some languages mark reflexivity in the verb (English does it, but not productively, in verbs like 'self-destruct'), while others use reflexive pronouns ('itself', 'themselves', etc.) in the object position.

In some languages, pronouns acting as objects (and/or subjects) are incorporated in the verb (Spanish tacks clitic object pronouns on the verb, either before or after).

Some languages are more rigid than others with respect to the argument structure of verbs. For example, transitive verbs may always need a explicit object. Compare this to English, where the objects of many transitive verbs can be left out, and many verbs are interchangeably transitive or intransitive (e. g. *burn*, *write*, *see*, etc.).

Voice

Voice can be understood from two points of view: the syntactic and the semantic. The semantic point of view refers to what voice represents for the meaning of the verb and the sentence. In English you can show whether the topic or theme of the proposition is the subject (active voice) or the object (passive voice). *The dog bit me* is active (the topic is *the dog*), while *I was bit by the dog* is passive (the topic is *I*). Since English, like many other languages, tends to equal topic with subject, this is how you topicalize a part of the sentence (in Japanese this is unnecessary, since topic can be explicitly marked in a different way, apart from the subject/object distinction).

From the syntactic point of view, the idea is that voice changes the way in which the arguments are arranged. Voice change is a grammatical operation that shifts arguments from their original places and may increase or decrease the valency of the verb. In English passive voice constructions, the original object becomes the subject (it gets **promoted**), while the original subject becomes an optional complement (it gets **demoted**).

English and other languages use a periphrastic construction with the verb *to be* and a participle for passive voice. Latin verbs, on the other hand, can be inflected by voice: *curare* 'heal', *curantur* 'they are healed'.

Active and passive are not the only voice distinctions. Greek had a **middle voice**, which suggested an action performed by the subject for his/her own sake. From the point of view of meaning, Spanish has a middle (or mediopassive, or pseudo-reflexive) voice shown by the pronoun *se*: *Se vende bien* 'It sells [itself] well', *apartarse* 'set oneself aside'.

In addition to these, there are voices that are more difficult to define from the semantic point of view, but can be understood as syntactic devices. For example, many ergative/absolutive languages have an antipassive voice, that transforms a transitive verb into an intransitive one ('I eat meat' becomes 'I eat'). In these languages, this also means that the subject is demoted from ergative to absolutive, though this doesn't show up in the translation. Changing the case of the subject may be done to allow coordination with other propositions.

One of my languages, Terbian[15], has an applicative voice, which promotes an optional (oblique) complement to the object position, with a special marking on the verb that shows the general function of the original complement (did it refer to a position or place, to a destiny, to a source?). For example (to take one that is easily translatable), 'he swims under the boat' becomes 'he underswims the boat'. In Terbian there is a kind of antipassive voice that also acts on intransitive verbs with complements by promoting one complement to the subject position and demoting the original subject: 'the cat sleeps on the mat' becomes 'the mat *sleeps the cat'.

Deference

Verbs may show the degree of deference (or the need of politeness) between the speaker and the hearer. In certain languages, there are different forms of verbs (and pronouns) to address a subordinate, a master and an equal. Japanese verbs can be inflected to increase politeness: *hanasu* 'speak', polite form *hanashimasu*. Japanese also has hyper-polite verb forms, and several other registers of speech that may be used in different occasions, by and to different people.

Weirdness and trivia

Some very common verbs in English aren't found in other languages, like 'to have'. Many languages rephrase 'I have a book' by 'A book is to me', or 'with me' or something to that effect, either using prepositions or case marking.

The copula 'to be' is in many languages not a verb, but a special word in its own class. In Japanese the copula has a special paradigm that differs from common verbs.

Many languages (such as Arabic, Hebrew and Russian) simply omit the copula in the present tense (this is called **zero copula**), so two noun phrases, or a noun and an adjective, put together, form a valid sentence (A B = A is B).

[15] http://www.angelfire.com/scifi2/nyh/terb/lng/index.html

Some verbs can be used as grammatical words beyond their original status. For example, in Khmer you use the verb 'to give' as the preposition 'to', to mark the indirect object of verbs. I'm guessing that this might correspond to a serial construction: English 'I give the book to her' could be translated as 'I take the book and give her'. This could be common for languages that avoid ditransitive verbs.

In Ainu, the conjugated forms of the verb 'to have' are used as possessive marks. For example:

kukor	kunupe	kunukar	rusuy
1s.have	1s.brother	1s.see	want

'I want to see my brother'

Note the 1st person singular prefix 1s is placed before verbs and nouns. Given this, it's not impossible to think of a language where possessive pronouns don't exist, nor are they formed from personal pronouns, but are instead subordinate clauses, consisting of conjugated forms of 'to have': 'my brother' becomes 'the brother that I have'.

In Japanese, verbs are sometimes used in place of adjectives, taking advantage of the fact that subordinate clauses come before the modified noun. For example: *sabitsuita kokoro* 'rusted heart' (*sabitsuita* 'it rusted'), *takanaru mirai* 'soaring future' (*takanaru* 'it soars').

Conjunctions

Conjunctions are words which put together different parts of a sentence. English common conjunctions are *and, or, if, but,* etc. Conjunctions can be present or not. It's possible to include some distinctions in conjunctions which aren't made in English; for example, the difference between exclusive and inclusive *or*. In Latin, you can say *vel X vel Y* (X or Y, or both) or *aut X aut Y* (X or Y, but not both). Conjunctions can be sometimes transformed into other things; in Latin, while you have *et* 'and', you can also use a postposed particle *-que* to join two nouns: *Senatus Populusque Romae* 'the Senate and the People of Rome'. Some languages do not have conjunctions at all; they simply put things together. 'X Y' (perhaps with a pause between them) means 'X and Y' (or even 'X or Y', depending on intonation and context). You can also use a case ending to join things, saying 'X together-with-Y' for 'X and Y'. Or you can replace conjunctions by adverbs: 'I tried but I couldn't' gives 'I tried, however, I couldn't'.

Articles

Do you have **articles**? English has two, *a* and *the*. Spanish has four, two indefinite and two definite ones; two are feminine and two are masculine. If your language has grammatical gender, then perhaps the articles should agree with their nouns. In Greek, articles agree not only in gender, but also in number and case, with their head noun. Scandinavian languages place the articles at the end of words, attached to them as inflections (for example, in Swedish *en bok* 'a book', *boken* 'the book', *böcker* 'books', *böckerna* 'the books'). Many languages do not have articles. In most cases, you can paraphrase articles by using adjectives, quantifiers (like *some*, *all*), or demonstratives (*that, this*). Articles are often unstressed and joined to the following words, perhaps with elision of vowels and other simplifications. In French, you say *la voiture* 'the car' but *l'avion* 'the plane'. In Italian and Portuguese, the articles are joined to whatever particle is in their way.

Adpositions and particles

The word 'particle' refers to little words, generally invariable, that modify the meaning of other words, or the sentence. Among them we find adpositions (prepositions and postpositions), which are used by most languages to modify the meaning of noun phrases and create complements (of place, time, manner, etc.).

There are also particles that have a wider range of functions, like the many particles of Japanese, some of which function as postpositional case marks, others as part of adverbial phrases, and others to add different twists of meaning to the whole sentence. For example, *anata no* 'your' uses the genitive particle *no*; the particle *wa* signals a new topic (a change of subject of the sentence and the following utterances), which will be omitted and understood in the next sentences. There's even an 'exclamation particle', *yo*, used to add force to statements; and an 'interrogative particle', *ka*, which signals a question (*taberu ka* 'shall we eat?'). In addition, *ka* produces indefinite deictics (*itsu* 'when', *itsuka* 'sometime').

A language can have prepositions or **postpositions**, or neither (I know of no language that has no adpositions at all, though). Whether a language is pre- or postpositional depends mainly on the position of the parts of speech (especially the verb arguments) in a sentence. As a general rule, SOV languages are postpositional, and VSO languages are prepositional; SVO languages can go either way. When you're designing a language, you can go against these general rules, but you'll soon run into certain practical problems that will make it clear why this is so.

The most common adpositions can be adequately replaced by case, and perhaps adverbs. Japanese shows many relationships with postposed particles which don't have a real meaning, but only general functions. In some cases, when it needs to use the equivalent to an adpositional statement, it uses two nouns joined by the genitive particle: *heya no naka* 'room (genitive) in-side', 'the room's inside, inside the room'. So in fact some of our prepositions are rendered by nouns. This is not unheard of in English ('in front of', 'on top of'), and Spanish is full of noun phrases that replace single-word prepositions (*bajo* 'under' vs. *abajo de*, *encima de* lit. 'on-top of').

Syntax

In simplified terms, **syntax** is the order and structure of words and phrases in a grammatical proposition.

The various components of a sentence often appear in a fixed order. The more analytic the language, generally the more fixed the word order is. In Chinese and English, for example, sentences are ordered in such a way that the misplacement of any word can alter the meaning completely. The more synthetic the language, probably the freer the word order, because synthetic, very inflected words, can stand on their own, and they don't depend so much on context. For example, in Latin *Petrus amat Paulum* 'Peter loves Paul', the subject and the object are perfectly determined by case endings, and their place can be changed with no change of the meaning of the phrase: you can say *Paulum Petrus amat* or *amat Petrus Paulum* and it's OK. But in English, 'Peter loves Paul' and 'Paul loves Peter' mean different things, because word order serves the function of distinguishing subject and object; and 'loves Peter Paul' or 'Paul Peter loves' are impossible or ridiculous.

A synthetic language may have a free word order not only by resorting to case endings, since other grammatical devices such as agreement (between verbs and nouns, nouns and adjectives, etc.) may serve this purpose by reducing ambiguity.

Subject, verb, object

The main structure of a complete sentence includes subject, object, and verb. These can of course be ordered in only six different ways: SVO, SOV, VSO, OVS, OSV, VOS. English affirmative sentences usually employ SVO, although sometimes English lets out an OSV (in sentences like 'this I don't know' or 'to thee I will sing'). Spanish is a bit more loose: usually SVO, VSO as an alternative for most verbs, SOV or OVS when the object is a pronoun, etc. Perhaps certain verbs of your language can use one form, and others use a different one; or perhaps you could use one form for short sentences and another one for longer complex sentences.

There is always an **unmarked word order**, that is, a particular order that doesn't convey any extra information (such as emphasis), and is therefore 'neutral' for the hearer. For example, English unmarked word order is SVO. The examples of OVS order I gave are **marked**; they make you focus on the object.

Some orders are more common than others. According to surveys, SVO and SOV languages each comprise about 40% of the world's languages. VSO languages are relatively frequent too, 15%. The other word orders (where the object is before the subject) comprise about 5%. So if your language is intended to be average, use SVO or SOV; if you want it to be exotic and weird, try OVS, OSV or VOS.

Heads and modifiers

Each part of a sentence can be divided into a **head** and zero or more **modifiers**. The head and its modifiers make up the phrase.

A phrase that functions as a noun (and whose head is a noun) is called a noun phrase. In a noun phrase like 'the little red cottage', the head is 'cottage' and the modifiers are the article and the two adjectives. A phrase whose head is a verb is called a verb phrase, and it may be modified by adverbs, negative auxiliaries, etc.

All languages have an unmarked order for heads and modifiers in each case, which is sometimes fixed. A language like English, that places modifiers before heads ('red dog', 'terribly hot summer'), is called **head-last**. A language like Spanish, where modifiers come after their heads, is called **head-first**. There are more technical designations for these tendencies, 'left-branching' and 'right-branching'.

Be aware that I speak of tendencies here. While English adjectives tend always to come before nouns, in poetry they are sometimes placed after them. In Spanish the opposite happens: most adjectives follow nouns, but in some cases they come before, especially for emphasis and in poetic speech. There is also variation according to the kind of modifiers used: English places adverbs before verbs, but longer adverbial phrases (such as 'in the park') after the verb. Japanese places everything before the corresponding heads, even subordinate clauses; the subordinate clause acts as an adjective:

K anojo ga dakishimeta otoko wa goshujin deshita.
she NOM embrace-PAST man TOPIC her _husband be-POLITE PAST
"The man (that) she embraced was her husband."

There are general tendencies correlating sentence-level word order (the order of subject, verb and object) and the place of heads and modifiers within phrases.

Sentence order	Phrase order	Adpositions
SOV	head-last	postpositional
VSO	head-first	prepositional
SVO	either way	either way

These are only tendencies and have many exceptions. While SOV languages are almost always head-last and use postpositions (the prototypical example is Japanese), Latin is SO V, yet uses prepositions and moves heads and modifiers around rather freely. SVO languages can go either way (English and Chinese are both prepositional, but Chinese is markedly more head-last than English; and Spanish, French and Italian, also SVO, are head-first). SOV languages usually mark the subject somehow, since it could get confused with the object that follows; SVO languages don't need that marking (though many of them use it), because the verb itself separates subject and object.

Verb-second languages

Some languages (featuring different word orders) are known to have a peculiarity regarding the position of the verb within the sentence. They are called **verb-second languages** (or shorter **V2 languages**, though that may have bad historical connotations). All the Germanic languages (except English) are V2 languages. The verb (or more correctly, the finite verb or auxiliary) has to be the second constituent of the sentence. This is not the same as SVO or OVS order; English is SVO, but in a sentence like 'Yesterday I went to a party', the verb is actually the *third* constituent (the first is the adverb, 'yesterday', and the second is the subject pronoun, 'I'). For our purposes, constituents are noun phrases (i. e. article or demonstrative + adjectives + noun), verb phrases (i.e. conjugated verbs and auxiliaries), adverbs and adverbial complements.

In V2 languages there is room for one and only one constituent before the verb. If something has to be emphasized, it usually comes to the front of the sentence (this is called **focus fronting** and happens in many languages). If the language is V2, however, this means that something else will have to move to the other side of the verb. For example, in German you can say (the verb, or actually the auxiliary, since the complete verb phrase is *hat geschenkt*, is in UPPERCASE):

Zum Geburtstag **hat** sie ihm ein Buch geschenkt.
for (his) birthday has she him a book given
"For his birthday she has given him a book."

Ein Buch **hat** sie ihm zum Geburtstag geschenkt.
a book has she him for (his) birthday given

"She has given him *a book* for his birthday."

Geschenkt **hat** sie ihm zum Geburtstag ein Buch.
given has she him for (his) birthday a book
She has *given* him a book for his birthday.

Of course, German has case, so the subject and objects don't get so confused as in the English literal gloss.

English is a Germanic language too, and though it has lost V2 compulsory order, it has kept some traces. You can see it in the way questions are asked (*'Who you saw?' is 'Who did you see?' because the auxiliary occupies the second position), in the use of auxiliaries in general, in phrases like 'There is', 'Here is', etc., and notably in seemingly 'inverted' sentences like 'Never had I seen such a thing'.

Trigger systems

This topic is a bit outside the scope of this section, but I felt it was worth including. The word order classification of which I've been talking presume that there will be a subject, a verb and an object, and that they'll be differentiable by the word order itself and/or by case marks.

There's a different system, which is used in Malagasy and most Filipino languages, like Tagalog, in which subject, object and other modifiers may appear in different orders, and they're not marked in traditional ways. It's called a trigger system.

The **trigger** is the part of the sentence over which emphasis is placed (I'd call it the topic, but I'm not so sure about this). The trigger can be the 'subject' of the sentence according to our view, but also the object, or a location, or the verb (predicate) itself. The trigger is marked as such (by a particle or inflection, or by word order), but you only state 'this is the trigger', not its function. Other parts of the sentence are marked differently. Then the verb is marked to show the relationship of the action to the trigger. The 'case' of the trigger is not marked on the trigger but on the verb.

In order to illustrate this, I'll just transcribe part of a post to the Conlang list, by Kristian Jensen, who was kind enough to repost it when I asked for an explanation about the subject. Here it is:

In Tagalog, there are only three markings for case: the Trigger, the Genitive, and the Oblique. This is exactly like most (if not all) the Philippine languages. Furthermore, much like many Western Austronesian languages, there are a large inventory of affixes used to create different nuances in the verbs, notably the verbal trigger. When the trigger plays the role of the agent, an agent-trigger affix is used with the verb. When the trigger plays the role of the patient, a patient-trigger affix is used with the verb. When the trigger plays the role of location, then a location-trigger affix is used with the verb. Etc. etc., etc...

A particularly noteworthy feature of this system is that non-triggered (unfocused) core arguments are marked as the genitive. As a result, "I am buying" and "the buying (of something) of mine" (or "my buying (of something)") have identical structures. Verbal constructions appear to be identical with nominal constructions by the use genitives. One theory has it that the verbal affixes are actually nominalizing affixes. Examples always help.

Take the sentence "The man cut some wood in the forest". With three different arguments, three trigger forms are possible. Below are parsing examples of the way a Filipino language would translate the sentence. I have refrained from using real language examples at this point hoping that it would be easier to understand how the _grammatical system_ (_not_ the morphological system) works.:

AGENT Trigger:
AT-cut GEN-wood OBL-forest TRG-man
"[cutting-agent] [of wood] [at forest] = [man]"
lit.: "The wood's cutter in the forest is the man"
transl.: "The man, he cut some wood in the forest"

PATIENT Trigger:
PT-cut GEN-man OBL-forest TRG-wood
"[cutting-patient] [of man] [at forest] = [wood]"
lit.: "The man's cutting-patient in the forest is the wood"
transl.: "The wood, the man cut it in the forest"

LOCATION Trigger:
LT-cut GEN-man GEN-wood TRG-forest
"[cutting-location] [of man] [of wood] = [forest]"
lit.:"The man's cutting-location of wood is the forest"
transl.: "The forest, the man cut some wood in it"

Note how I have nominalized the verbs in the transcription. Thus, the verb for cutting has been nominalized as an agent, a patient, or a location depending on what role the trigger plays. There are other verbal trigger forms too including benefactor and instrument. My own theory is that trigger languages only have one core argument. Such being the case, trigger languages resort to nominalizing verbs. This might also explain why passive constructions do not exist in trigger languages since the valency of the verb is not changed (cannot change) with different triggers.

In a language using a trigger system, it's not useful to talk about subject, object, etc., and word order may greatly vary. In Tagalog, the predicate (the nominalized verb) is the first word in the sentence, and the trigger is last. Other languages might be different. It's equally useless to talk of transitive or intransitive verbs, or of voice (active, passive, middle).

This is just to show you how things can be really different, and still understandable. See if you can imagine something else!

Morphosyntactic typology

When one talks about verb arguments (or syntactic elements in relation to the verb), one usually distinguishes two basic ones, which we will call subject and object. According to the manner in which a language marks those, we have several types thereof:

1. An **accusative** language is one where

- the subject of all verbs (transitive and intransitive) is marked with one grammatical case, conventionally known as 'nominative';

- the object of a transitive verb is marked with another case, which is conventionally named 'accusative'.
- 2. An **ergative** language is one where
- the subject of an intransitive verb and the object of a transitive verb are both marked with one grammatical case, called 'absolutive';
- the subject of a transitive verb is marked with another case, conventionally known as 'ergative'.
- 3. An **active** language is one where
- the subject of a transitive verb is marked with a grammatical case, usually named 'agentive' (A);
- the object of a transitive verb is marked with another case, usually known as 'patientive' (P);
- the subject of an intransitive verb is marked with either one of the two cases mentioned above (A or P) according to semantic considerations.

A different, more formal way of looking at it, is using three syntactical categories, usually labelled S, A, and P, where S is the only argument of an intransitive verb, and A and P are the two arguments of a transitive verb. There is (it seems) no language on Earth that marks these three roles using three different cases; they're usually divided, one marked with one case and the other two with a different case. Thus, a language that groups (treats alike) S and A is an accusative language (P gets the accusative case); a language that groups S and P is an ergative language (A gets the ergative case); and a language that groups S and A or S and P according to the verb is an active language.

There's apparently no language that groups all three roles; something (some morphology or word order) distinguishes between them on most occasions (and context disambiguates if not). Also, almost no language groups A and P and sets S apart (A and P need to be distinguished since they're both arguments of one verb, but S doesn't need marking since an intransitive verb has no other argument).

Accusative languages

Let us recall the definition given above: accusative languages mark the subject of all verbs with one case (nominative, NOM), and the object of transitive verbs with another case (accusative, ACC). That's why they are also called nominative/accusative.

The typical example of an accusative language is Latin.

domin -us veni-t
master-NOM come-3sPRS
"The master comes."

domin -us serv -um audi-t
master-NOM slave-ACC hear-3sPRS
"The master hears the slave."

Most Romance languages have not preserved the morphological case marks of Latin, but the order of the words within the sentence, as well as concord (grammatical agreement) and context, allow us to differentiate the nominative and the accusative roles. Therefore these

languages (Spanish, Italian, French, etc.) show a syntactic accusative quality, rather than a morphological one.

English, while not a Romance language, also derives from a case-inflected language and has also lost most morphological cases, but its syntactic accusativity can be confirmed by observing sentences where an argument is deleted. In the sentence *"the pupil saw the teacher and left"* there are two coordinated propositions with a common argument. The fact that the missing argument is assumed to be *"the pupil"* points to the fact that English is an accusative language, because the nominative role takes precedence to occupy the vacant space, since the verb in the second proposition (*"left"*) requires a nominative subject. In an ergative language (see below) the missing slot would have been occupied by the absolutive case argument (which is the object of the first proposition).

The great majority of Indo-European languages are accusative. However, some present a partial ergative behaviour.

Ergative languages

An ergative language, as we saw, is one that marks the subjects of transitive verbs with one case (ergative, ERG), and the subjects of intransitive verbs and objects of transitive ones with another case (absolutive, ABS).

The ergative language most known in Europe is Euskara (Basque), which is in fact the only European ergative language, and cannot be grouped within any linguistic family, being probably the last remnant of ergativity left behind after the Indo-European occupation.

Georgian (spoken in the nation of Georgia, an ex-Soviet republic and birthplace of Stalin) shows ergative patterns in one of its verb series (the verb system in Georgian is extremely complicated), but is accusative in the rest. In one grammar sketch of Georgian that I have, it is described as having formal ergativity with features more in line with those of active languages of the Split-S type (see below).

The Australian language Dyirbal is also partially ergative (it uses an ergative structure for third-person sentences, but becomes accusative for the first and second persons), with an underlying syntactic structure that is ergative. Hindi is ergative in the perfect tenses, and accusative in the imperfect ones. (These weird cases have been explained in several ways, all of them rather dense...)

An example of ergativity (from Euskara):

umea erori da
ume -a -0 eror-i da
child-the-ABS fall-PRF AUX:PRS+3sS
the child (ABS) fallen is
"The child fell."

emakumeak gizona ikusi du
emakume-a -k gizon -a -0 ikus-i du
woman -the-ERG man -the-ABS see -PRF AUX:PRS+3sS+3sO
the woman (ERG) the man (ABS) seen has

"The woman has seen the man."

In an ergative language, the argument in the absolutive case is the one that is assumed when it is missing. Thus, while in English *"the pupil saw the teacher and left"* is interpreted as *"the pupil saw the teacher"* + *"the pupil left"*, the equivalent in Euskara or another ergative language (with syntactic ergativity) would be interpreted by assuming the absolutive object of the first proposition as the subject of the second verb (which is intransitive):

"the pupil (ERG) saw the teacher (ABS) and left"
is interpreted as
"the pupil (ERG) saw the teacher (ABS)" + "[the teacher (ABS)] left"

A test of this kind with the native speakers of a language (where they are forced to fill in the vacant slots and complete their interpretation) is a way to decide if a language is ergative/absolutive.

Interestingly, ergative languages usually do not have a passive voice, but they do have an antipassive voice, which deletes the direct object and demotes the subject from ergative to absolutive (i. e. it makes the verb intransitive).

Active languages

As explained above, an active language is one where the S-role (the subject of an intransitive verb) can be marked in one of two ways (either as A = agentive or as P = patientive), according to semantic considerations with respect to the verb or its argument.

Active languages are in turn divided into two types:

- Languages with a split S-role (**Split-S**), in which the decission to mark the Subject of a given verb as A or P has been made beforehand, so to speak, in a conventional way, and fixed as part of the syntactic structure;
- b. Languages with a fluid S-role (**Fluid-S**), in which the decision to mark the subject as A or P depends on real-time semantic considerations and must be taken by the speaker according to his/her intention and the context, since the meaning of the expression can be changed.

The semantic considerations mentioned above may have to do with the kind of concept described by the verb (is it an event or action, or is it a state?), as well as the degree of control or will of the subject over the action or state expressed by the verb (is it a voluntary act or an involuntary one?, does the actor perform it directly or through an instrument?). In Fluid-S languages these considerations have to be pondered by the speaker to twist the meaning to one side or the other. In Split-S languages each verb has these connotations (and the way of marking the intransitive subject) already assigned as part of its definition, and all the speaker may do is learning this and employing it in the usual way, modifying it through other means when s/he deems necessary to change the meaning.

For example, 'sleep' shows an involuntary state. In a Split-S language, the speaker will mark the subject of 'sleep' as P always. If s/he wishes to make it explicit that an effort was made to sleep, or something like that, s/he will have to resort to auxiliaries ('try to sleep') or other means to convey this meaning. On the other hand, in a Fluid-S language, while the typical

use of 'sleep' will have the subject marked as P, the speaker might actually be allowed to suggest 'go to sleep, make an effort to sleep' by using the same verb 'sleep' with a Subject marked as A. In this way one could also give different meanings to verbs like 'cough' (generally involuntary, but sometimes wilfully performed by the actor) or 'turn around' (active and usually voluntary, but sometimes an unconscious reflex act).

Daniel Andreasson, from the CONLANG list, researched the subject and sent the list a brief explanation. He states that active languages distinguish between A and P Subjects according to several criteria (each language uses primarily one of these):

- event vs. state
- control
- performance, effect and instigation

"Event vs. state" means that if the verb is an event (like 'run', 'dance', 'chat', 'kill'), then the argument is marked like A. If it's a state ('be hungry', 'be tired'), then it's marked like P.

"Control" means that if the argument of the verb is in control of the event (or state), then it's marked as A. If it is not in control, then it is marked as P. 'Go' and 'be careful' are controlled predicates. 'Die' and 'fall' are not.

Then there's "performance, effect and instigation". Some predicates are in some way performed or instigated by the actor. However, they need not be controlled. These are verbs like 'sneeze' and 'vomit'. In languages like Lakota and Georgian, it's enough if the actor in some way performs the action (or state), s/he doesn't need to be in control. Thus the argument of predicates like 'sneeze' and 'hiccup' are marked as A. In languages of group (b) ("control") these would be marked as P.

Analogy

Analogy is the blanket term for various kinds of processes that change the phonetics and the grammar of a word or expression, produced by very special causes. When I speak of analogy I will usually be referring to phonetic change.

Analogy is the creation of a new form of a word by influence of similar, **analogical** forms. Analogy is quite a fruitful device, and it's probably one of the major word-creators in languages. Let's see an example.

Latin derives from Proto-Indo-European (a language or set of dialects of a language that has been reconstructed based on its daughter languages). In PIE, nouns had case, so they changed form according to case. The word for *honour* was reconstructed as having the forms **honos*, **honosem*. As PIE evolved and gave origin to Latin (and also Greek, Germanic, Sanskrit, etc.) some sound change took place. In particular, the /s/ sound between vowels gradually became voiced (/z/) and finally gave an alveolar trill, /r/ (this change is called **rhotacism**). This only happened when the /s/ was intervocalic, and not in any other position.

(Before) (After)

*honos -> honos
*honosem -> honorem

This, as you see, produced an irregularity; the root form of the word split in two forms, *honos-* and *honor-*. All languages have some irregular forms, but this one (and many others of the same kind) probably wasn't accepted by speakers. Now put your hand over the "Before" column and hide it, ignore it. Speakers couldn't know anything about the sound change, which is a subtle and unconscious process (and not studied in those times). What could you do with the irregular pair *honos/honorem*?

The solution came by analogy with the many words which hadn't changed form (I don't know enough Latin to give an example), and with the same root. They had *honorem* and also *honoris*, perhaps even *honorificum* and so on, so they began saying *honor* instead of *honos*. That's analogy.

Of course, no language ever takes analogy so far as to regularize its whole grammar.

A related form of analogy appears when people create words out of elements they had, based on other similar words. English is quite prolific in this respect. Having words like *pulverize* or *finalize*, English speakers have created analogical forms like *idealize*, *nationalize*, *hospitalize* and hundreds more. If you're creating a language, probably analogy will be the best tool to increase your lexicon.

Grammatical devices

This section is a general one which will mention and summarize the main grammatical devices found on languages, i. e. how a grammar is managed at the practical level (on actual words).

We have already seen most of these devices in a way or another. Here's a brief list of them:

- **Affixion**: this includes adding prefixes, suffixes or infixes to words in order to change their meaning or their relationship with other words. These affixes include what we call inflections and also agglutinated affixes.
- **Word order**: it's free in some languages and fixed in some others (see Syntax). In general, the more synthetic the language, the freer the word order. An analytic language such as Chinese relies on word order to clarify the meaning of words, because they are never inflected and therefore don't show their functions on their structure. (Actually Chinese does have some inflections... in fact, according to certain authors, English is more analytic than Chinese.) A synthetic language like Latin can construct a sentence with scattered words (this is called **hyperbathon** [I think] and is used as a poetic device).
- **Stress and pitch**: we've already talked about them. In some languages they are only formal; in many others, two words can have different meanings according to their stress patterns. Compare English *a record* /'rek@rd/ and *to record* /ri'kord/ (and many other pairs).
- **Tone**: the same as for stress and pitch. Sometimes a change in tone distinguishes two completely different words, and sometimes it produces a different form of the same word. In Shilluk, *yít* (high tone) means "ear", and *yìt* (low tone) means "ears"; tone is not a phonetic feature but a grammatical feature.
- **Alternation**: we've seen it with examples. It's the (regular) change of sounds on words. The most common is vowel alternation, which is indeed found in English: compare *sing, sang, sung*, and *man, men*, etc. In some languages this is not irregular but the norm. Consonant alternation is less common but does exist (compare *a house, to house*, voiceless vs. voiced). Consonants can alternate in different ways, not only by voice; they can change stop to fricative, or fricative to affricate, or simple to double, or even in strangest ways. There's an African language where /t/ alternates with /l/ and /p/ alternates with /w/ (this is voice alternation but also involves other contrasts).
- **Reduplication**: (a part of) the root of a word is doubled, repeated before or after it. A reduplicated verb can increase its force, like Hotentot *go* "look" vs. *go-go* "examine with attention" (used by Philip J. Farmer in *Riders of the Purple Wage*, in the Go-go School of Criticism). A reduplicated noun can be taken as plural, like *gyat* "person" vs. *gyigyat* "people" (again an African language), which also shows vowel alternation. Sometimes the reduplication is just put there as part of an inflection. In Greek, the perfect forms of verbs use reduplication and vowel alternation: *līpō* "I leave", *hélipon* "I left", *léloipa* "I have left".

Creating words

Well, now you have everything set up, so you have to begin creating words. Probably you already have some particles, case endings, affixes, etc., but that's only the skeleton.

How many words do you need? If you're creating a full language (which I assume you are, because you wouldn't have come this far if you weren't), then you'll need about 2000 (two thousand) words to communicate with a certain comfort. You can do quite a lot with about 1000 words, if that scares you; but you'll probably be creating new words now and then.

Mark Rosenfelder mentions (and I'm not going to repeat it here) the thesis of Ogden and Richards. These guys showed that the most part of any English text contains a very reduced lexicon. A group of common words cover 80% or 90% of any text. Then they said, "Well then, let's isolate those words and use them and only them, combining them to form complicate concepts instead of using not-so-common words". For example, forget the word "success" and use "make good". All in all, you could do with only 850 common words and perhaps a hundred more for specific fields.

The argument is right, but it has a failure. The most common words which cover so much of the text are also the ones that carry the least information: articles, prepositions, pronouns, etc. In newspaper headlines, those are usually deleted, because they are not so important and the rest can be understood. The not-so-common words cannot be deleted, because they are the ones which convey all the meaning, all the information. In fact, the theoretical basis of modern informatics says that the most unusual signs are the ones that possess the most information. If you understand the 90% of the words in a text, but the 10% remaining is composed of the most critical information, then you're actually getting nothing except a lot of particles connecting unintelligible concepts.

So don't spare your words. You can never have too many.

How do you start? There's no method, but I'll tell some ways I have used:

- You can translate simple texts. When you need a word, you create it; if there's an available related root, you derive it from there, or else create and note a root first. You can't have words coming out of nowhere. Translation is tedious, and it bothers you to stop at each word and invent it, but it's wonderful to create words. What to translate is your decision. I don't recommend James Joyce or Kierkegaard or Borges, of course. The Babel text is quite good. You can go on with the Bible (or the Talmud or the Rig-Veda or whatever sacred scriptures your religion has, if it does and you have a religion). If that seems too dense, use comic books, or The Hobbit. If you dare, try translating from a conlang[16] (a glossed text) into your own.
- Perhaps you can find a list of basic vocabulary. I have an English-English dictionary intended for non-English speakers, with a list of 2000 common words that are used to explain the definitions, and I've taken some words from there and translated them into my own (invented) language. Don't translate dictionary entries. It's boring, it's time-consuming, and it's pointless: you'll be having lots of unusual words, all of whose English glosses will begin with **a**, and nothing else.

[16] http://www.angelfire.com/ego/pdf/ng/lng/draseleq/dc_cl2cl.html

- Find a topic or field and invent words on it. For example, verbs of motion (walk, go, jump, come, rise, raise, drag, spin), or body parts (head, arms, legs, toes, fingers, face, eyes, hair), or colours (you know the colours), or numbers (you'll have to create a numeric system or use the decimal one), or tools, or animals, or domestic appliances.
- This one I haven't used yet, but it just seems interesting: create rhyming words. Take any collection of English concepts you like, and translate the first one with a certain word in your language, and all the others with words that rhyme with it. Or the other way round (English has lots of rhyming words, especially monosyllables). Or you could build alternating series, words which vary only in their first consonant, or in their vowels (of course they should be totally unrelated concepts, unless sound alternation is a valid inflecting mechanism). You can then use these words to make puns if you like.

There's a very interesting list of words (the **Universal Language Dictionary**) which comprises 1600 words divided into topics, and used in some way by the most common languages of the world. You can find it at the Model Languages site: it comes with the *Langmaker* language generator. Very good, at least to check for words (it's not very fun to sit and generate them one after another). For a simpler but still useful way to generate random words, try Wordgen[17]. It lets you specify beginning, medial and final consonants, clusters, vowels and diphthongs, and the number of syllables you want.

[17] http://www5.palmnet.net/~black/prog/wordgen32.zip

Final words

If you want to become a great language creator, read! Read everything that falls into your hands or passes by. The Web is full of material, though a bit scattered. I have already mentioned some of my sources. Here's a full list of sites you should visit:

Model Languages (www.langmaker.com) is a newsletter devoted to language creation, which used to be published bi-monthly. The newsletter is not published any more, but the old issues are still online. You can find lots of online material there; it's quite a lot of reading material and it also features a wonderful list of more than 200 links to pages about invented languages. There's also a word generator that can handle different syllable structures and produce words, and derive them according to simple phonetic changes.

Mark Rosenfelder has made a terrific work in his site, *Metaverse* (www.zompist.com), including the Language Construction Kit, a review on Quechua, a list of numbers from 1 to 10 in 3500 languages, and lots of material about one of his languages, Verdurian.

Then there's the *Human Languages Page* (http://www.june29.com/HLP/), which is a bit scrambled, but helps you find linguistic resources on lots of natural languages.

The folks at SIL have collected an immense amount of definitions having to do with linguistics and the study of language (including rhetorics). Check out the *Glossary of Linguistic Terms* (http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/).

If you're a J. R. R. Tolkien fan, you can find descriptions of the languages he invented in www.ardalambion.com.

For a look at some real world scripts, you can visit Ancient Scripts (http://rabbitmoon.home.mindspring.com/asw/), a very well-made set of pages with examples of writing systems from around the world, including Mesoamerica, Europe, and Middle East.

You shouldn't leave without visiting the pages in the Scattered Tongues webring.

If you want to get into the conlanging community, join the **Conlang list** by sending an e-mail to listserv@listserv.brown.edu with subscribe conlang your_name as the body of your message. Conlang is dedicated to the discussion of constructed languages for fictional purposes. If you belong to Conlang already, or you're simply curious, visit the Conlang FAQ for a lot of topics covered in past threads, or consult the Conlang Archives.

Joshua Shinavier, a fellow member of Conlang, has a quite comprehensive list of constructed languages of which you can find some material in Internet: The Conlang Yellow Pages (http://www.geocities.com/Athens/Crete/5555/conlang.htm). No better way to learn about language construction than seeing how others have managed it.

And then of course there are libraries, those quiet buildings full of books. I've learned a lot from linguistics books. Most often than not, they are dense and sometimes unintelligible (they weren't intended for ordinary people trying to create languages), but they often provide explanations on curious stuff along with examples. The best way to learn how to invent a language is studying natural languages.

Well, so long! If you're creating a language and would like to expose them to the praise and critique of the world, or just need to get some advice or to give some advice, mail me and I'll do my best to correspond to your expectations.

Acknowledgements

I want to give thanks to the following:

- **Mark Rosenfelder**, for his excellent work in the Language Construction Kit, which taught me a lot and inspired me to write this, and for not complaining when I took big chunks of it.
- **Jeffrey Henning**, for his (also terrific) work as the editor of the famous Model Languages newsletter.
- **Nik Taylor**, a fellow member of CONLANG, who was if I recall correctly the first person to write to me re: How to create a language, correcting some gross mistakes and contributing data about the record 92 consonants of !Xu~ and the average proportion of obstruents to sonorants.
- **Kristian Jensen**, who taught me and the rest of the CONLANG list about trigger systems.
- **Markus Miekk-oja**, a.k.a. Miekko, who shared a lot of curious things about languages real and fictional, including the mysteries of the many Finnish cases and the names and uses of verb moods in Nenets.
- **Jarkko Hietaniemi**, for one nice example of agglutination in Finnish.
- **Donald Patrick Michael Goodman III**, for teaching me how to say "He's cute" in Japanese and then make it past tense.
- **Reena D.**, for correcting a typo in Donald's example.
- **Mathias Lasailly**, a fellow CONLANG member, who supplied the example of possession shown by a subordinate clause with the verb "have" in Ainu.
- **Cseri Benedek**, who corrected my mistake of stating that no languages consistently mark transitivity on verbs by showing me how this is done in Hungarian.
- All the members of the CONLANG list that I haven't named above.
- John Ronald Reuel Tolkien, Jorge Luis Borges, and so many others that have made me think about words, their meanings, their beauty and the magic wrought by them, which makes tangible the matter of dreams and thoughts.