Notes on Language Creation and Ergativity - http://dedalvs.free.fr/notes.html

Notes on Language Creation

There is no "How To" book for language creation. Everyone has their own opinions; everyone has good ideas. These are a few of mine (opinions, not good ideas--that's for you to decide).

Preface

Awhile back, Jeffrey Henning (the man behind Langmaker.com) suggested I create a kind of "How To" page for my website, perhaps something like his excellent Model Languages newsletters, which you can find here. He suggested I could start off with some of my better CONLANG posts, probably like this one, which someone posted as a resource to Langmaker.com.

My first thought upon hearing (well, reading, I guess) this suggestion was: Why? Not because I didn't think that at least some of my CONLANG posts were useful/helpful--I hope they were. My posts on CONLANG, though, are all spur of the moment, and are certainly not meant to be authoritative in any way. I felt like if you added something like a "How To" page, it would presuppose that you were an authority on the subject, and that there's some special reason why people should believe what you say. I don't feel like an authority on conlanging, and I certainly don't want to make it sound as if I think I am. So that was one reason I was hesitant.

Another reason was that there are plenty of "How To" sites out there already, such as Pablo David Flores's essay How to Create a Language (warning: that link is to a .pdf), as well as probably the best guide out there, Mark Rosenfelder's Language Construction Kit. And, of course, there are always the essays of Rick Morneau (if you're a language creator, or interested in language creation, and are *not* familiar with these essays, you probably should become familiar with them). Anyway, the point, I suppose, is this: If we've already got the steel-belted radial, why re-invent the stone wheel?

Wow, it's kind of hard to argue with the logic of that analogy... But anyway, I did, in fact, decide to create a *kind of* "How To" page, of which this page is evidence. The main reason I did so is this: There is no authority on language creation. There's barely even a literature. Sure, there are plenty of books that have created languages in them (go here for an ever-growing, yet inexhaustive, list), but there are very few (one?) that actually *discuss* the creation of language in any depth. Thus, if we, the language creation community, don't discuss our art ourselves, who will? Chances are it'll be an outsider--

someone like Marina Yaguello, who wrote a book whose title is *Lunatic Lovers of Languages* (thanks for the vote of confidence, Ms. Yaguello). I, for one, don't want that.

Additionally, there are as many ways to create a language as there are people to create them. And since chances are that much of our work will be lost if we don't put it somewhere public, the need to at very least catalog your ideas online is vital. I'm continually amazed at not only the ideas of well-established conlangers, but also of those new to the game who've never even had the privilege of being able to discuss their conlangs with a sympathetic audience. Without fresh ideas, new blood, the communal aspect of the artform can't survive, *po-moemu*.

The purpose of this preface is threefold. First, I wanted to explain why this page is on my site, of course. Second, though, is that I'd like to urge the conlangers reading this (well, the conlangers who have webpages) to put up not only their language sketches, cultural descriptions, scripts, artwork, etc., but also their ideas, their thoughts about language creation; what they've learned. Your experience is invaluable: Let us know about it.

Oh, there's also a third reason for the preface. I wanted to explain how this notebook will be structured. Unlike an actual "How To" guide, this notebook will *not* be in sections that build off one another and gradually increase in complexity. In fact, the first content section (not including this preface or the introduction) is on ergativity--a notoriously sticky subject. So what you should do is just go to the table of contents and see if there's a subject that interests you. If so, click on it, and dive on in. If not, hey, that's life. Try back again some time. I plan to add to this page periodically.

Oh, one more thing. There are two types of links on this page. Those that show up in pink but are *not* italicized go pretty much wherever they say they go. Those that are in pink and *are* italicized, however, go directly to a linguistic definition of the given term which is hosted on SIL's Glossary of Linguistic Terms. It's a helpful site, and I've made use of it liberally not only on this page, but on all my pages.

All right, that's enough of a preface. I bid you a good day, and hope you can find something useful on this page.

Introduction

Let me pause while I figure out what this is an introduction to... Ah, yes.

This is intended to be a general introduction to me as a language creator, so you can know where I'm coming from.

I was never really interested in language the way I am now (and the way most language creators have always been) until my junior year of high school. Before then, I came from a house where English was the first language and Spanish the second, but I never fully

learned Spanish. So, when I got to high school, I took Spanish, because everyone had to take a language. It was in my junior year, though, that I woke up one morning with a startling thought: Millions of people on Earth could speak French fluently, and I wasn't one of them. This greatly disturbed me. I was more embarrassed, than anything else. Like I'd walked into a black tie social event in my pajamas (and little-kid footy pajamas, at that). From that day forward, I was determined to learn every language on Earth, living or dead. (Note: It wasn't until much later that I learned that there were thousands of these things, and that I would have to revise my self-imposed goal, if I hoped to live anything that even resembled a normal life.)

Shortly after my revelation, I started to pick up different language books here and there. And so, I started to teach myself Latin and French. In my senior year, I added a German class, though I was thwarted in my attempt to take French 2 without having taken French 1. I also started to try to learn Arabic. Then when I go to college at UC Berkeley, I took, in my first year, a year of Arabic, a semester of Russian, and a semester of Esperanto. Esperanto was my official introduction to created languages, though at the time, I never imagined that one even *could* create a language for fun. That thought didn't dawn on me until my next semester, when I (finally) took a French class, and took my very first linguistics class: Linguistics 5, introductory linguistics. Some time during the lesson on the IPA, I thought to myself, "Hey, what if I came up with my own IPA, so that I could write English in an Arabic-style script?" I'd become enamored of Arabic, and especially its script, you see. And then I had a startingly thought. "What if I actually created a *language* that was *like* Arabic, but simple and regular, like Esperanto?" And that was the end of it for me. Ever since that day, just about all my free time has been spent creating languages.

That first language was a language called Megdevi, named after myself and my girlfriend at the time. My idea was to create a language that we could speak between ourselves. (What a laugh!) When I realized that wasn't going to pan out, I just started to expand it on my own, adding sounds that I liked, not having to worry about how others could pronounce them any longer. Pretty soon I got some font making software and started creating a font. This led to creating more fonts and more languages.

It wasn't until March of 2001, it turns out (I could've sworn it was November...), that I came across the CONLANG list. It looks like my first message was on March 8, 2001, and it was rather argumentative. An ill omen. Oh well. One thing that's important to understand about me and language creation is that I really thought I had come up with a novel idea. I new that Esperanto had been created back in the 19th century, and that a few others had been created around that time (Ido, SolReSol, Novial, Volapük, etc.), but I didn't know that anyone had actually created a language for fun. Ever. I never read Tolkien as a child (I almost got three fourths of the way through *The Hobbit* once), and still am not fond of him. And even though I knew of him, certainly, I never knew that he created languages. I grouped him together with C.S. Lewis and George Orwell (other writers I read in fourth/fifth grade) as a set of sci-fi/fantasy-type authors, and never dreamed that he, as a member of that group, did anything but write. I'd certainly never heard of the actual Klingon language, or any other type of conlang, for that matter. I

honestly and truly believed that I was the first. I continued to believe for a few months until I came upon Pablo David Flores's page on the internet, and was crushed. After all, if I was one of many, what was the point? So for the few months when I found out about language creation on the web and found out about CONLANG, I was in a bad mood. It's not surprising that I was so arrogant and rude, though it remains, nevertheless, unforgivable (especially since I was probably one of the reasons that David Bell abandoned CONLANG. I still feel very bad about that, and if he ever reads this, I want him to know that I'm sorry).

Anyway, during this time, I started to develop Megdevi. I got to a point where all I had to do was add triconsonantal roots. Thus, the vocabulary began to grow by leaps and bounds. At the same time, there was discussion on CONLANG about vocabulary size. Someone posed (I believe) about how their vocabulary had finally grown to 300 words. I looked at Megdevi and estimated the number of words, and it was well over 5,000. As a result, I got the idea that I was really a lot better at language creation than everyone on the list. What I didn't know, though, was that quite the opposite was true.

The language Megdevi itself (and I won't ever put anything up about it. The Babel Text is here if you want to get an idea for what the language was like) was really a very clever code for English. Its triconsonantal roots encoded semantic categories from which nouns, adjectives and verbs could be made. Any time I came across a construction my language couldn't handle, or learned about something new in one of my linguistics classes, I merely added an affix. And Megdevi had prefixes, suffixes, infixes and circumfixes--every kind of affix I'd heard of at that point. Thus, when it came to translation, its power was unlimited. Any time I came across something it couldn't handle, I'd either add another triconsonantal root, or add a new affix.

Now, I've no doubt that anybody on the list could've pointed out what was wrong with Megdevi. It would've been like taking candy from a baby who liked to hand out candy to strangers. I think, however, that it was best for me that I discovered it on my own. I believe it was when I was coming up with a new root for "fortify". Thus, the verb meant "to fortify", the verbal noun was "fortification", the utility noun was "(a/the) fortification or fort"... And it was *right then*, right at "fort", that I realized I was doing nothing more than cleverly recreating the vocabulary of English. And it was then that I realized that all the other languages I'd started at the time (languages like Geydr [not mispelled], Sunshine, Dangelis, Color, Mbasa, Zidaan...) were terrible. The more and more I learned in linguistics, the more and more I saw how little I understood about language, and how much my languages had suffered. So, I stopped working on Megdevi, and all the others, and started a new language: Kamakawi. This was the first language I started that I considered somewhat good. It still suffers from some of my old bad habits, as do Sathir, Njaama and Zhyler, but it was a marked improvement. At the same time, I began to appreciate more and more others' languages, and was finally able to really start getting stuff from the CONLANG community.

From that point on, I kind of settled into a groove. I started to learn more languages (Middle Egyptian, Hawai'ian, Turkish...), learn a lot more about linguistics, and to work on the languages that are currently on this site.

Some time near the end of my stay at Berkeley, I started up an experiment with John McWhorter that eventually became the Wasabi experiment. The paper I wrote at the end of this experiment is what I used as my writing sample for my graduate school applications. Additionally, I was able to talk about the talk I gave on language creation at a colloquium that our club at Berkeley (the Society of Linguistics Undergraduates, SLUG) put on, and so, quite literally speaking, I can say that language creation is what got me where I am today: at UCSD as a linguistics graduate student. Language creation has made a great impact on my life thus far, and I hope to be able to do even more with it in the future.

But, for now, it's fun. And that's what matters most. ~:D

Ergativity

Ergativity: The Maltese Falcon of language creation. If you'd like a linguistic definition, you can go *here*, but it probably won't help much. Essentially (and you should take that word with a bucketful of kosher salt), ergativity is this: In English (a nominative-accusative language), the subject of a sentence with a transitive verb and the subject of a sentence with an intransitive verb are treated alike; direct objects of transitive verbs are treated differently. In an ergative-absolutive language, the subject of an intransitive verb is treated the same as the direct object of a transitive verb; subjects of transitive verbs are treated differently. That, however, is only the verytip of the flap on top of the roof on top of the house on top of the iceberg. In fact, that definition is wholly inadequate when it comes to explaining ergativity, but many don't know why. That's fine if you're a doormat salesman; not so fine if you're a conlanger who wants to create an ergative-absolutive conlang.

In this introduction to ergativity, I'll try to explain what *exactly* ergativity is, and how it's manifested in natural languages, as well as how it can be used in created languages. I will be drawing on a number of resources which I'll mention throughout this introduction, and will also list at the end.

So, without further ado, I give you: Ergativity.

1.0 Introducing Terms:

Before jumping into theory and examples, I want to make sure that we've got our terms straight.

a. First of all, there are the terms "nominative-accusative language/system" and "ergative-absolutive language/system". Each of these refer to a language that display either non-ergative or ergative characteristics. This does *not* mean that the language in question will have *cases* with these names. After all, English is a nominative-accusative language, but has no case (except in the pronouns, and those cases work differently than standard nominative-accusative).
b. With that said, the names that are given to these systems do come from somewhere. Specifically, the four words used in the system names are case names. The *nominative* case that identifies the *subject* (regardless of the *valency* of the verb) in nominative-accusative languages. The *accusative* case is a case that (usually) marks the *direct object* of a *transitive verb* in nominative-accusative languages. The *absolutive* case is a case that marks the subject of *intransitive verbs* and the direct object of transitive verbs in ergative-absolutive languages. Finally, the *ergative* case is the name for a case that marks the subject of a transitive verb (not *necessarily* the *agent*) in ergative-absolutive languages.
c. Actually, since I introduced a semantic term up above, it might be useful to go over the relevant ones. An *agent* is, strictly speaking, the initiator of an action. In this section, I'll be referring to the agent of a transitive verb as an A. Now, in a sentence like, "The polar bear's dancing", "the polar bear" is actually an agent--i.e., he's initiating the dancing action. I'll be referring to those types of arguments (i.e., the volitional/agentive subjects of intransitive verbs) as SA. A *patient* is the undergoer of an action. So, for example, in "The polar bear tapped the panda", "the panda" is the one who undergoes the tapping action. I'll be referring to these types of patients as P. Another type of patient would be "the door" in a sentence like "the door swung open". I'll be referring to these types of patients as SP. Three other semantic roles I'll be talking about are recipients (R), *experiencers* (E) and stimuli (ST). I'll explain these when I get to them. The prior four, though, will be important to remember as we go along.
d. Two processes I'll be discussing later on are *passivization* and *antipassivization*. I think it might help just to think of these as a simple valency-decreasing operation, but one typically applies to nominative-accusative languages, and the other typically applies to ergative-absolutive language. Both of these processes affect transitive verbs. The process takes the default argument and turns it into an oblique, and takes the specially marked argument and turns it into the default argument. In a nominative-accusative language, nominative is the default marking; accusative the special marking. In an ergative-absolutive language, the absolutive is the default marking; the ergative the special marking. The resulting verb is a very intransitive-like verb, in both cases. That's all this is.

Okay, those are some terms that we need to make sure we're all on the same page about. (Heh. How's *that* for a sentence ending with a preposition?) If you're not sure how I'm using a term later on, come back here, and it will explain.

1.1 Introducing Some Test Words:

In explaining (and hearing explanations of) ergativity, I've always found it more helpful to look at invented examples than actual examples from natural languages. I will talk about natural languages below, but most of the examples will be shown using the words listed below. The words below will be used to illustrate *all* examples, so that we're not switching languages from example to example, and so that it'll be easier to familiarize yourself with what exactly is going on. Or that's the plan, at least. So below are a list of words from a language that we'll call Ergato:

English	**Ergato**	***English***	**Ergato**
I	ko	*panda*	panilo
you	pe	*fish*	tanaki
she	li	*sheep*	folime
to dance	talu	*man*	hopoko
to sleep	sapu	*woman*	kelina
to pet	lamu	*book*	kitapo
to see	fisu	*wind*	makipo
to give	kanu	*house*	paleni
and	i	*General Preposition*	sa
Valency Reducing Marker	-to	*Oblique Marker*	-k
Past Tense Marker	-ri	*Recipient/Dative Case*	-s
Plural Marker	-ne	*Extra Case Marker*	-m
Default Case Marker	--	*Special Case Marker*	-r

It's important to understand why the markers above do *not* say things like "ergative case marker", or "antipassive marker". These markers are going to be used differently in different contexts in the examples below. Thus, the "special case marker" will show up as both an accusative case marker and as an ergative case marker. Now I'll start in with the examples.

2.0 THE PRISTINE SYSTEM:

There are a lot of conlangs out there that are, essentially, *pristine systems* (note: this is my term). A pristine system, when talking about language, is a system where there are no irregularities, and everything works the same way, no matter the context. This is ideal for an IAL, or a loglang. If your goal is to create a natural language, though, a pristine system is something to be avoided, because *no* natural language is pristine (not even Turkish). Nevertheless, a pristine system (or an attempt at a pristine system) is what many first-time conlangers aim for (most of the time unconciously). I'm now going to show you what a pristine nominative-accusative system and a pristine ergative-absolutive system looks like. I'll start with a nominative-accusative system.

2.1 A PRISTINE NOMINATIVE-ACCUSATIVE SYSTEM:

Before I begin, I want to say that I'm assuming that a pristine system will utilize case marking, because, when it comes to conlangs, that's usually the case. There is such a thing as a pristine language that doesn't use case marking, but I'll get to those later. So now for the pristine nominative-accusative language. To test for pristineness (pristinity?), there are some general sentences you can use. You will want to test:

1.
    a. A sentence with an intransitive verb with a patient-like subject (SP).
    b. A sentence with an intransitive verb with a agent-like subject (SA).
    c. A sentence with a transitive verb with a agentive subject (A).
    d. A sentence with a transitive verb with an experiencer subject (E).
    e. A sentence with a ditransitive verb.

So, let's test those sentences in pristine nominative-accusative Ergato:

2.
    a. *Kelina sapu.* "The woman is sleeping."
    b. *Kelina talu.* "The woman is dancing."
    c. *Kelina lamu panilor.* "The woman is petting the panda."
    d. *Kelina fisu panilor.* "The woman sees the panda."
    e. *Kelina kanu kitapor hopokos.* "The woman's giving the book to the man."

The above is extremely indicative of a pristine nominative-accusative system. The thing that tips you off to its being a nominative-accusative system is that the subject *kelina*, "woman", is in the same case (the default case) in sentences (2a), (2c) and (2e). The thing that lets you know that the system is pristine is that *kelina* is in the same case for sentences (2a) and (2b), and also for sentences (2c) and (2d). English is not a pristine system when it comes to this criterion, though it's not because of case. Take the two translations of sentences (2c) and (2d) above and compare each to its incorrect counterpart in English below:

3.
    a. The woman is petting the panda.
    b. *The woman pets the panda.
    c. The woman sees the panda.
    d. *The woman is seeing the panda.

Sentences (3b) and (3d) above are grammatical, but they don't mean the same thing as sentences (3a) and (3c), respectively. This is because in the present tense English is sensitive to whether the subject is an experiencer (E) or an agent (A). Instead of it being marked as a case, it's marked with the presence or absence of the auxiliary "be".

Now, it's not enough to merely test the sentences in (1) to determine whether or not the system is pristine. I'll explain more about why this is later. Suffice it to say that you should also test:

4.

a. A sentence with a pronoun as the subject of a transitive verb.
b. A sentence with an inanimate noun as the subject of a transitive verb.
c. A sentence in the past tense with a transitive verb.

So, let's test those quickly in pristine nominative-accusative Ergato:

5.

a. *Li lamu palinor.* "She's petting the panda."
b. *Kitapo lamu palinor.* "The book's petting the panda."
c. *Kelina lamuri palinor.* "The woman petted the panda."

Now, with sentence (5b), you're going to have to use your imagination. So let's say a woman has a very clean panda that she doesn't want people petting with their hands (because hands have germs). So, not wanting to offend her (or her panda), you pick up a book and kind of stroke the panda with it. Suddenly, the woman asks, "What are you doing?" You reply, "I'm petting your panda." "With your filthy hands?!" she screams. You reassure her, "No, no. *The book's petting the panda.*" Far-fetched, but it will serve our purposes.

Anyway, the point is that nothing has changed with respect to case marking. The subject of the sentence still gets default marking, and the object still gets special marking.

Based on all this evidence, you can determine that the system is a nominative-accusative system, and that it's pristine. That is, the subject of the sentence will *always* get default marking, no matter what the tense is, or what kind of verb it is, what tense, animacy, etc. It's hardcore nominative-accusative. And that means that you can safely label the *-r* suffix as being an accusative marker.

Now that we've determined what kind of system we have, let's look at the valency-reducing mechanism. This will *only* apply to verbs that have at least two arguments: A subject and object (however they're marked, casewise). So we can ignore intransitive verbs for now. So let's look at a couple sentences:

6.

a. *Kelina lamu palinor.* "The woman's petting the panda."
b. *Palino lamuto (kelinak).* "The panda's being petted (by the woman)."
c. *Kelina kanu kitapor hopokos.* "The woman's giving a book to the man."
d. *Kitapo kanuto hopokos (kelinak).* "The book's being given to the man (by the woman)."

So, a few things to notice. The first and most obvious thing to notice is that what was the object in the transitive sentence (marked with *-r*) is now the subject in the passivized sentence (now given default marking). Second, the verb is marked with *-to*, to let you know the passivization process has occurred. Third, the actual subject of the sentence has been made superfluous. That is, just as you can say "The panda's being petted", so can

you say *Palino lamuto* in this version of Ergato. Expressing the actual subject is *optional*. Finally, with respect to that optional subject, notice that if you *do* express it, it's no longer in subjective case (default marking/nominative), but in an oblique case. This is the case for just about every language that has a passive. What will change is what that oblique case is. So, in English we just have a prepositional phrase headed by "by". In Turkish, you have something similar, only with a postposition. The point is that the noun will be marked in some totally different way, and will be treated a different way by the syntax.

Well, that's about it for pristine nominative-accusative Ergato. So, onto pristine ergative-absolutive Ergato!

2.2 A PRISTINE ERGATIVE-ABSOLUTIVE SYSTEM:

This should go a lot faster. In section 2.1, I wanted to explain why we were doing a lot of the things we were doing. Now that you know, though, we can right to the examples. So, here are our initial batch of test sentences:

7.
   a. *Kelina sapu.* "The woman is sleeping."
   b. *Kelina talu.* "The woman is dancing."
   c. *Palino lamu kelinar.* "The woman is petting the panda."
   d. *Palino fisu kelinar.* "The woman sees the panda."
   e. *Kitapo hopokos kanu kelinar.* "The woman's giving the book to the man."

Immediately, something should jump out at you as being radically different. Aside from the case marking, the subject is appearing in totally different places! This is because this system is *pristine*. A truly pristine system would line up cases on the same side of the verb, no matter what. So the equivalent to the pristine nominative-accusative system is an ergative-absolutive system where the absolutive case (now the default marked case) *always* comes before the verb, the ergative case (now the *-r* case) *always* comes after the verb, regardless of whether it's the subject of the sentence or not. A good many first-time ergative languages are not pristine, but usually it's unconcious, because, since English is a nominative-accusative language with no case marking, it seems natural to always put the subject on the same side of the verb. That's not the way a pristine ergative-absolutive system would work, though.

Now that we've hurdled that...hurdle, we can talk about the other differences. Most notably, the subject of the sentence is being marked differently depending on whether it's in a sentence with a transitive verb or a sentence with an intransitive verb. Notice, though, that this system isn't sensitive to the status of the subject. So in an intransitive sentence, the subject is marked with the absolutive, regardless of whether it's an SA or an SP. Similarly, in a transitive sentence, the subject is marked with the ergative, regardless of whether it's an A or an E.

Let's quickly look at our other test sentences:

8.

a. *Palino lamu lir.* "She's petting the panda."
b. *Palino lamu kitapor.* "The book's petting the panda."
c. *Palino lamuri kelinar.* "The woman petted the panda."

As you can see, there's no change in case marking, or in the placement of the subject.

Now onto antipassives. Antipassives seem to *really* confuse a lot of folks, and I think it's because, to a nominative-accusative speaker, there doesn't seem to exist a conceivable reason to ever use an antipassive. The usual example from English used to try to explain antipassives is the verb "eat". So, you can say "I ate breakfast", or you can say "I ate". Thus, the object is kind of superfluous. This, however, is *not* the same thing, and that's *not* why antipassives are used. I'll do my best to explain here.

To begin with, let's actually see some antipassive sentences. Here goes:

9.

a. *Palino lamu kelinar.* "The woman is petting the panda."
b. *Kelina lamuto (palinok).* "The woman is petting (and what she's petting is the panda)."
c. *Kitapo hopokos kanu kelinar.* "The woman's giving the book to the man."
d. *Kelina hopokos kanuto (kitapok).* "The woman is giving to the man (and what she's giving is a book)."

I used those convoluted translations in (9b) and (9d) to try to show how the optional phrase in an antipassive *feels* to the speaker. It really is extra, unnecessary information.

Anyway, notice what happened. If the absolutive is the default, unmarked case, and the ergative is the special, marked case, what an antipassive did was got rid of the special case. Thus, you might say that there's less mental work involved when it comes to case in antipassives (maybe). Also, an antipassive allows you to focus on one aspect of the action, in this case, the performer of the action. Finally, think about why we use passives in English most of the time. If you think about it, the usual reason to use a passive is if you want to conjoin things in discourse. So, let's say we're talking about an accident where one car is at fault (i.e., it hit the other one). I might say, "I saw the car that was hit". I probably would never say, "I saw the car that the car at fault hit it" (that's probably not even grammatical). The second sentence is how you'd *have* to say it, though, if there were no passive. Why? Because when two sentences are conjoined in English, the subjects go together. So, if you say, "The Toyota hit the Honda and skidded", the car that skidded *has* to be the Toyota, and could *never* be the Honda. The same kind of thing happens in ergative-absolutive languages, but instead of the subject being carried over, it's the absolutive argument. Maybe an example will help explain:

10.

a. *Palino lamuri kelinar i [palino] talu.* "The woman petted the panda and [the panda] danced."

b. **Palino lamuri kelinar i [kelinar] talu.* "*The woman petted the panda and [the woman] danced."

That is, in my opinion, probably *the* reason why valency-reduction systems exist. If you don't have them, everything you say becomes extremely roundabout. For example: "Yesterday, there was an accident that I saw. A Toyota came and smacked a Honda and the Honda skidded along the street. Later on, I saw the car, such that the Toyota hit it. The Toyota had banged it up pretty badly. The Toyota made it such that its trunk wouldn't close, and also made it such that one couldn't see out of its rear window." If you allow for valency-reduction (in this case, passivization), the whole thing becomes much shorter and easier to understand. In this way, antipassivization is no different from passivization. Think of it as a kind of luxury. After all, not all languages have valency-reduction systems. You best thank your lucky stars that your language does! (Or, well, that the language you're reading right now does.)

3.0 Syntactic Ergativity:

You know, I think it'd be easier to explain syntactic ergativity before going on to split-ergativity. So I'll do that. I'm going to explain how pristine syntactic nominative-accusative and ergative-absolutive languages work, because, basically, it's identical to what's above, but without the case-marking.

3.1 A Pristine Syntactic Nominative-Accusative System:

English is just about a pristine syntactic nominative-accusative system. Almost. Its sensitivity to experiencer verbs in the present and its pronouns are the only thing standing in the way. Close, though.

I'm just going to list the sentences. Note that when I say *syntactically* nominative-accusative or ergative-absolutive, it means that relations are determined by word order. So here's pristine syntactic nominative-accusative Ergato:

11.
a. *Kelina sapu.* "The woman is sleeping."
b. *Kelina talu.* "The woman is dancing."
c. *Kelina lamu palino.* "The woman is petting the panda."
d. *Kelina fisu palino.* "The woman sees the panda."
e. *Kelina kanu kitapo hopoko.* "The woman's giving the book to the man."

In the examples above, the object comes after the verb, and the subject before, in all cases. In the case of an indirect object, it's put after the direct object (remember: this is a *pristine* system. If the direct object is going to come after the verb, it should *always* come *directly* after the verb). Aside from sentence (11e), this should look a lot like English. Now for the next set:

12.

a. *Li lamu palino.* "She's petting the panda."
b. *Kitapo lamu palino.* "The book's petting the panda."
c. *Kelina lamuri palino.* "The woman petted the panda."

Again, not different from English. If this were a purely syntactic language (i.e., isolational), you might expect the past tense suffix to be a past tense word, but that really doesn't have any bearing on what we're doing now. So, now for the last set:

13.
a. *Kelina lamu palino.* "The woman's petting the panda."
b. *Palino lamuto (sa kelina).* "The panda's being petted (by the woman)."
c. *Kelina kanu kitapo hopoko.* "The woman's giving a book to the man."
d. *Kitapo kanuto hopoko (sa kelina).* "The book's being given to the man (by the woman)."

In these examples, the preposition is used to indicate the demoted subject, just like English "by". Notice that the demoted subject comes *after* the indirect object (which now sits next to the verb) in (13d).

Well, that really does it for pristine syntactic nominative-accusative Ergato. The important thing to notice is that what is what is wholly dependent upon word order. We'll see more of the same with pristine syntactic ergative-absolutive Ergato below.

3.2 A PRISTINE SYNTACTIC ERGATIVE-ABSOLUTIVE SYSTEM:

Now we can see the flip-side of the pristine syntactic coin. Here's the first set of examples:

14.
a. *Kelina sapu.* "The woman is sleeping."
b. *Kelina talu.* "The woman is dancing."
c. *Palino lamu kelina.* "The woman is petting the panda."
d. *Palino fisu kelina.* "The woman sees the panda."
e. *Kitapo hopoko kanu kelina.* "The woman's giving the book to the man."

Here the absolutive argument always comes sentence-initially, and the ergative argument always comes directly after the verb. Also, you should know that the placement of arguments (i.e., where the absolutive argument goes, where the verb goes, etc.) is totally arbitrary. As long as those places are honored no matter what happens, the system is considered pristine. Now let's look at our secondary examples:

15.
a. *Palino lamu li.* "She's petting the panda."
b. *Palino lamu kitapo.* "The book's petting the panda."
c. *Palino lamuri kelina.* "The woman petted the panda."

Again, these extra facets don't affect the position of the arguments in the sentence. Now for our antipassive examples:

16.
   a. *Palino lamu kelina.* "The woman is petting the panda."
   b. *Kelina lamuto (sa palino).* "The woman is petting (and what she's petting is the panda)."
   c. *Kitapo hopoko kanu kelina.* "The woman's giving the book to the man."
   d. *Kelina hopoko kanuto (sa kitapo).* "The woman is giving to the man (and what she's giving is a book)."

Here again, in these examples, the absolutive and ergative arguments are switching places, and the demoted absolutive argument (the old one) is optionally expressed as a PP headed by our all-purpose preposition *sa*.

And that's how a syntactically ergative language works. Rather than looking at case marking, you look at word order, and how the different arguments show up in different types of sentences. Admittedly, it's probably *easier* to see this kind of thing when there's case marking, but not all languages mark case overtly. Plus, a syntactically ergative conlang would be a real rarity; quite unique.

Now it's time for the tough stuff.

4.0 Split-Sensitivity:

I'm calling this section "split-sensitivity" because *all* languages show split-sensitivity to *something* to *some* degree. I've already shown an example from English. Even though it's nominative-accusative, it's sensitive to experiencer verbs in certain situations, but not in others (e.g., in the past tense). Split-sensitivity is a blanket term for any language that shows one kind of pattern in one place, and a different kind of pattern in a different place. That's all. The thing that characterizes these languages is: (a) What is split (case marking, for example); and (b) where the split occurs. We'll now delve into split-sensitivity.

4.1 Tense-Based Split-Ergativity:

One of the most common types of ergativity is ergativity that's split based on tense. Hindi and Georgian both display this kind of ergativity. The most common way to split it is so that in the present tense (or nonpast), the language displays a nominative-accusative system, and in the past tense, the language displays an ergative-absolutive system. So let's focus on that kind of split and see what our test sentences look like:

17.
   a. *Kelina sapu.* "The woman is sleeping."
   b. *Kelina talu.* "The woman is dancing."
   c. *Kelina lamu panilor.* "The woman is petting the panda."
   d. *Kelina fisu panilor.* "The woman sees the panda."

e. *Kelina kanu kitapor hopokos.* "The woman's giving the book to the man."

All these sentences are in the present tense, so, unsurprisingly, they look just like the sentences in (1). Now here's where the difference lies:

18.
a. *Li lamu palinor.* "She's petting the panda."
b. *Kitapo lamu palinor.* "The book's petting the panda."
c. *Palino lamuri kelinar.* "The woman petted the panda."

Now let me stop right here to explain some things. What you see above is what you'd expect if you were melding to pristine systems (i.e., where the word order and case marking are just like those in the pristine ergative-absolutive version of Ergato). This is *not* usually the case, though. First off, it's much more likely that the subject of the sentence would be in the same place. Thus:

19.
a. *Kelinar lamuri palino.* "The woman petted the panda."

Second, though it would be economical to use the same case marker to mark the accusative and ergative, the ergative languages I know of (I'm thinking of Georgian in particular) don't. Instead, what you'd see is something like this:

20.
a. *Kelina lamu palinor.* "The woman's petting the panda."
b. *Kelinam lamuri palino.* "The woman petted the panda."

In effect, what you have is three case markers. One case marker (the default marker) marks the nominative in the present and the absolutive in the past. Another, the special marker *-r*, marks the accusative in the present. Then you have a third, the extra case marker *-m*, which marks the ergative in the past. This is exactly the type of system that Georgian has (give or take the lack of an accusative marker that's distinct from the dative, and the inappropriate use of the word "tense").

As you might expect, the valency-reduction mechanism works differently in the present and past. However, here there are further wrinkles. This is how one might imagine the system would work:

21.
a. *Palino lamuto (kelinak).* "The panda's being petted (by the woman)."
b. *Kelina lamurito (palinok).* "The woman's petting (and what she's petting is the panda)."

That would be a nice way for it to work. And maybe there are some that do. However, there are theories about the evolution of *some* ergative-absolutive systems that suggest

that ergativity in the past tense arose from present tense passive constructions. So what you might get would look something like this:

22.

a. *Kelina lamu palinor.* "The woman's petting the panda." (Present Tense, Active)
b. *Kelinak lamuto palino.* "The woman petted the panda." (Past Tense, Active)
c. *Palino ke lamu (sa kelina).* "The panda's being petted (by the woman)." (Present Tense, Passive)
d. *Palino ke lamuto (sa kelina).* "The panda was being petted (by the woman)." (Past Tense, Passive)

So remember what those markers mean. The first sentence is standard issue. The *second* sentence, however, might look like a passive. According to some theories (I've heard this about Hindi, but it is just a theory), what happened was that the passive was used so often that it *became* the past tense, and so the valence-reducing marker *-to* now function as (and, well, is) the past tense marker. But since it was a passive, the subject is marked with the oblique case (that's what the *-k* is). And, of course, in a standard passive, the promoted object is marked with the subjective case. When this construction becomes the normal past tense, though, the word order falls in line (subject first; object last), and so you get what looks like an ergative-absolutive system only in the past tense. Then what I wanted to show with sentence (22c) is that some new construction would arise to fulfill the role of the present tense passive. So, *ke* in that example would be some kind of auxiliary, and the reintroduced subject would be reintroduced by a "by" phrase, like English, rather than being expressed with the oblique (now ergative) case marker. Then, in the past tense...who knows? (22d) is my guess as to what *could* happen to create an antipassive. It might be advisable to see what Hindi does. (I'll check on that.)

Now, this subsection is devoted to ergativity split by *tense*, not just *past* tense. The thing is, I've never heard of a split-ergative language that splits it (based on tense) any other way. This could partly be because of the theory I mentioned above. That theory aside, though, this split could work the opposite way: Ergative-absolutive in the present; nominative-accusative in the past. Or maybe even the future. It could be an aspectual split: perfective vs. imperfective. It's perfectly possible. This is just the most common. Georgian does something that really isn't best described as a split system based on tense. This is because what constitutes "tense" in Georgian is incredibly complex. Each verb can be conjugated in 12 or 13 different ways, and these ways are divided into three series: present, aorist and perfect. If I remember right (I'll check my notes and get it straight later), it's the perfect series that displays an ergative-absolutive pattern, whereas the present and aorist series display a nominative-accusative pattern. Anyway, in the case of Georgian, I'd argue that the split isn't based on tense, but on morphological category. The Georgian system is a *fascinating* system for many reasons. You might go here for more information, or look up Stephen R. Anderson's paper on case in Georgian (though don't take it too seriously).

4.2 Pronominally-Based Split-Ergativity:

Another common way to have a split system is to have one kind of system that's used with overt nominals, and to have a different system used with pronouns. A prime conlang example of this kind of system is the masterful David Bell's ámman îar (click here to go directly to the part that explains the ergativity of ámmar îar). A lot of ergative languages do this, but often it's mixed with an animacy (or, as Payne calls it, "agency-worthiness") system, which I'll describe later.

The basic concept behind a system where the split is based on whether you have a pronominal argument or an overt NP isn't that hard to imagine. For this example, let's say that Ergato displays an ergative-absolutive pattern for overt nominals, and a nominative-accusative pattern for pronouns. Here are our example sentences:

23.

a. *Kelina sapu.* "The woman is sleeping."
b. *Kelina talu.* "The woman is dancing."
c. *Kelinam palino lamu.* "The woman is petting the panda."
d. *Kelinam palino fisu.* "The woman sees the panda."
e. *Kelinam hopokos kitapo kanu.* "The woman's giving the book to the man."

I changed the word order to a (in my mind) more natural word order for an ergative-absolutive language. So now there's a dominant SOV word order, but the case marking on the subject changes, so that you get an *-m* when the subject is an A. Other than the word order, though, the sentences in (23) are identical to those in (7). [Note: I'm going to go ahead and continue using *-m* as the default ergative marker when A's and P's are marked separately.] Now let's look at our secondary test sentences:

24.

a. *Li palino lamu.* "She's petting the panda."
b. *Kitapom palino lamu.* "The book's petting the panda."
c. *Kelinam palino lamuri.* "The woman petted the panda."

Check out sentence (24a). The only way you know which is the subject and which the object is the word order. But that's not the whole story. So far we've sentences with two overt NP's and one with a subject pronoun and object NP. Now let's look at an intransitive sentence with a subject pronoun, and two transitive sentences, one with a subject NP and an object pronoun, and the other with two pronouns:

25.

a. *Li sapu.* "She's sleeping."
b. *Palinom kor lamu.* "The panda's petting me."
c. *Li kor lamu.* "She's petting me."

In (25), you can see the fully fleshed out version of a pronominally split-ergative language. A and S pronouns are marked just like S and P NP's, and P pronouns have a special accusative marker.

So now we come to valency-reduction. I have no information at hand that addresses what I want to know (e.g., what happens with split-ergative systems and passivization/antipassivization). The only examples that Payne lists of antipassivization in his otherwise fantastic book *Describing Morphosyntax* are from languages that are entirely ergative-absolutive. Thus, I'll list what a language *might* do, or could conceivably do:

26.

a. *Li (kelinak) lamuto.* "She's being petted (by the woman)."
b. *Kelina (lik) lamuto.* "The woman's petting (her)."

What I've shown in (26) is, essentially, a subject controlled valency-reduction system. In other words, depending on what the subject of the sentence is, that determines whether the result is interpretted as a passive (in the case of a pronominal subject) or as an antipassive (in the case of an overt NP subject). It's also possible that you might have two different kinds of systems. So, maybe you have a normal antipassive system for NP's, and then a different kind of antipassive system for pronouns. Either way could work. (Note: David Bell's pronominally split-ergative language ámman îar appears to have taken a semantic approach to valence functions, as opposed to morphological. In other words, you can make any transitive sentence into a passive sentence or an antipassive sentence regardless of case marking. Go here for a thorough account.)

The example I showed above featured an ergative-absolutive system for overt NP's, and a nominative-accusative system for pronouns, but it could easily go the other way. Additionally, you could have different systems for different pronouns, but I'll discuss that in more depth when we get to the section on animacy.

One last thing I want to mention (something that doesn't deserve its own section) is person marking on verbs. Person marking on verbs can work exactly the same way as separate pronouns. My language Sathir is a language that works this way (the language is ergative, but pronominal subjects are marked on verbs, whether they're A's or S's). If we wanted to use Ergato as an example, we could pretend that the pronouns were pronominal suffixes (for one type), and suffixes and prefixes (for a different type). Here's an example where subjects are marked on verbs if they're not overtly specified. The case marking system is ergative-absolutive. This yields:

27.

a. *Kelina sapu.* "The woman's sleeping."
b. *Kelinar palino lamu.* "The woman's petting the panda."
c. *Sapuko.* "I'm sleeping."
d. *Palino lamuko.* "I'm petting the panda."

In the above example, the NP's show normal ergative-absolutive case marking (S and P get default marking; A special), but subjects are marked the same way regardless of their status. That's one way it could work. Now imagine a language where NP's are marked in

a nominative-accusative way, and verbs inflect for both subject *and* object. Here's what that could look like:

28.
   a. *Kelina sapu.* "The woman's sleeping."
   b. *Kelina palinor lamu.* "The woman's petting the panda."
   c. *Sapuko.* "I'm sleeping."
   d. *Palinor kolamu.* "I'm petting the panda."
   e. *Kolamupe.* "I'm petting you."

The sentences in (28) are essentially a variant on the word order model. The point is that, in transitive sentences, subjects are inflected with a prefix and objects are inflected with a suffix. In intransitive sentences, subjects are marked with a suffix, just like objects in transitive sentences. At the same time, overt NP's are marked in a traditional nominative-accusative way. This same effect could be achieved (and often is) by having different forms of pronominal inflection for the different roles. Here, though, I wanted to keep it simple.

I think that about does it for pronouns. We'll revisit pronouns when we discuss animacy.

4.3 Semantically-Based Split-Ergativity:

This type of split is *extremely* common in *all* the world's languages, though usually in small doses. Essentially, this type of split is a split that causes similar arguments with different semantic roles to be marked differently. The example of this I already discussed is English's sensitivity to verbs of experience in the present tense. But that's not the whole story. Not by a long shot.

Let's start off with something simple. This is what English's pattern might look like in a case-marking language:

29.
   a. *Kelina sapu.* "The woman is sleeping."
   b. *Kelina talu.* "The woman is dancing."
   c. *Kelina lamu panilor.* "The woman is petting the panda."
   d. *Kelinas fisu panilo.* "The woman sees the panda."
   e. *Kelina kanu hopokos kitapor.* "The woman's giving the man a book."

Above, the word order doesn't change, but notice that the case marking on the subject of (29d) is dative case marking, just like the case marking on the indirect object of (29e). This is a common occurrence in the world's languages, where an experiencer subject gets marked as a recipient of some kind. Additionally, the object of (29d) is marked with the nominative, or default case. Now, the above system, like English, makes sure to line up the subject. A different language, though, might make sure to line up the case, instead, yielding the following:

30.

a. *Kelina sapu.* "The woman is sleeping."
b. *Kelina talu.* "The woman is dancing."
c. *Kelina lamu panilor.* "The woman is petting the panda."
d. *Panilo fisu kelinas.* "The woman sees the panda."
e. *Kelina kanu hopokos kitapor.* "The woman's giving the man a book."

The reason for the above would be that, grammatically (or morphologically), *panilo* in sentence (30d) *is* the subject, and, therefore, should line up with the other subjects. It really depends on how the language defines the notion of subject.

Now how about this. We've seen three different case markers employed in one system: Default, *-r* and *-m*. Thus far, though, we haven't seen them all in the same tense. Can it happen? You bet it can. This is what it would look like:

31.

a. *Kelina sapu.* "The woman is sleeping."
b. *Kelina talu.* "The woman is dancing."
c. *Kelinam lamu panilo.* "The woman is petting the panda."
d. *Kelina fisu panilor.* "The woman sees the panda."
e. *Kelinam kanu hopokos kitapo.* "The woman's giving the man a book."

In this admittedly bizarre system, S's are marked the same way as P's (default marking), and A's are marked with *-m*. Then, possibly for semantic reasons, E's are marked the same as S's and P's, and ST's (stimuli) are marked with a third case, *-r*. That's really a bizarre system. Here's a more normal one that a large number of natural languages have:

32.

a. *Kelina sapu.* "The woman is sleeping."
b. *Kelinam talu.* "The woman is dancing."
c. *Kelinam lamu panilo.* "The woman is petting the panda."
d. *Kelina fisu palinor.* "The woman sees the panda."
e. *Kelinam kanu hopokos kitapo.* "The woman's giving the man a book."

Here's a system where's there's a distinction drawn between SA's (agent-like subjects) and SP's (patient-like subjects). In (32a) and (32d), the subjects of those verbs are more like patients than agents, so they get default marking, as do normal P arguments. The subjects of (32b), (32c) and (32e), though, are more agent-like (after all, one hopefully doesn't dance by accident). Thus, they're marked with *-m*. Finally, ST's are marked with *-r*. (Note: For what it's worth, I think this marking may be optional. Stimuli could very well be marked with the default case--or even with *-m*, possibly.)

Since we brought up SA's and SP's, I'd like to mention a little fact that can pop up in many different systems. Let's say volitionality is important to a given language. Thus, SA's are marked with an ergative marker (say, *-m*), and SP's are marked with an

absolutive marker (default marking). This could be a hard-and-fast rule, *or* the language can use the volitionality generalization to its advantage. Consider this possibility:

33.

a. *Kelina sapu.* "The woman is sleeping."
b. *Kelinam sapu.* "The woman is sleeping on purpose."
c. *Kelinam talu.* "The woman is dancing."
d. *Kelina talu.* "The woman is dancing on accident."

I could use other verbs that would make more sense here, but I'd rather not use *too* many different made-up words. Instead, I'll make up different contexts. So, for (33b), let's say the woman isn't so much a woman, but a young girl. It's Sunday morning, and she's woken up, but she knows tomorrow is Monday, and she remembers how nice it is to just laze about in bed. But she hears that her mother has awakened... And her mother wants to make her go to church, thereby ruining her lazy morning. As if on cue, in walks her mother to say, "Get up, Hildegarde: It's time for church." Oh, but young Hilde's concocted a fiendish plan. "Perhaps if I *pretend* I'm asleep," she thinks, "my mother will leave without me, not wanting to be late." And thus, Hildegarde attempts to sleep *on purpose*, as to fool her mother. That's context number 1 for sentence (33b). [Incidentally, this rarely works. I've heard.]

Now, for (33d). Imagine a dance at a high school gym--let's say, Pacifica High School's gym, located in sunny Garden Grove, CA. Now imagine that there's a woman (or girl) there who doesn't want to dance because she's afraid she won't be that good and doesn't want to embarrass herself. She's by no means unpopular. Several boys (yes, and even a girl or two) have asked her to dance, but she's systematically declined each one, citing the weather, an obscure religion, uncomfortable heels, a full bladder, etc. Unbeknownst to her, though, the ants that live beneath Pacifica High School in the Realm of the Ant have plotted against her. "Foolish human!" squeaks the queen of the ants. "She thinks she can attend a dance and *not* dance!? We'll see about that. My minions!" The queen's armies snap to attention, "Yes, your highness!" "This night we shall teach that wallflower a lesson. If I'm not mistaken, I spotted a cookie crumb that somehow fell onto that young girl's dress. Your queen desires a late night snack. If you have any love left for your queen at all, you'll bring me that crumb, do you hear!" "Right away, your highness!" And with that, the ants go marching one by one. Hurrah! Hur--"*AHHHHH!*" screams the young girl, as she spies the benighted trail moving slowly yet persistently up her calf. To get them off, she jumps; she twists; she flails wildly, and...as if *by accident*, the young girl is dancing! Young and sweet; only seventeen...

So there's your context. Languages that work this way are rather neat, because you can handle something so common, yet so rarely encoded morphologically, simply by changing the case of the subject.

This is by no means the end, though. After all, if there are different names for each of these types of semantic arguments (SA, SP, P, A, E, ST...), couldn't there be a language that marks each one separately? Yes, there certainly can. I'll show you two different

examples. In natural languages, this is rare, but attested. The most common of those types attested looks something like this:

34.
- a. *Kelina sapu.* "The woman is sleeping."
- b. *Kelinam talu.* "The woman is dancing."
- c. *Kelinam lamu panilor.* "The woman is petting the panda."
- d. *Kelina fisu palinor.* "The woman sees the panda."
- e. *Kelinam kanu hopokos kitapor.* "The woman's giving the man a book."

In the example above, SP's are marked with default case marking, SA's with *-m*, and objects (regardless of status) are marked with *-r*. This is a common enough pattern. But we can go further. Though I don't *believe* it's attested among natlangs, you can imagine a language like the following:

35.
- a. *Kelina sapu.* "The woman is sleeping."
- b. *Kelinak talu.* "The woman is dancing."
- c. *Kelinam lamu panilor.* "The woman is petting the panda."
- d. *Kelinap fisu palinol.* "The woman sees the panda."
- e. *Kelinam kanu hopokos kitapor.* "The woman's giving the man a book."

I had to make up some case markers on the fly in this one. Okay. Above, SA's are marked with default marking. SP's are marked with *-k*. A's are marked with *-m* (there are two. No language marks the agent of a transitive verb differently from the agent of a ditransitive verb. But one can imagine...). P's are marked with *-r*. Indirect objects are marked with *-s*. E's are marked with *-p*. And, last but not least, ST's are marked with *-l*. Now that's a very precise language. I'd like to point out that though this *type* of thing is attested, it's generally meted out differently than either of the two examples above (more on that when we get to animacy).

We're *almost* done with this section, but there's one bit left. We've talked about SA's and SP's, but consider the following English sentences:

36.
- a. "The woman's petting the panda."
- b. "The book's petting the panda."
- c. "The wind's petting the panda."
- d. "The panda's being petted (by the woman)."

Those four sentences have four different types of subjects--two of which we haven't really talked about before. The first in (36a) is simply an agent. The last in (36d) is a subject that is, in fact, a patient (i.e., the subject of a passive). The second subject in (36b) is something we've talked about, but not directly. Remember the story about the woman with the clean panda? The woman is still the one initiating the petting action, but the book is the instrument used to perform the action. Thus, the subject is an instrument (SI).

In (36c), unless the wind is some kind of sentient being, the wind is neither an instrument nor an agent, but simply a force of nature: a non-volitional subject (I'll call it SN). One could imagine a language where all four of these are marked differently, as in these sentences below:

37.

a. *Kelinam lamu palino.* "The woman's petting the panda."
b. *Kitapok lamu palino.* "The book's petting the panda."
c. *Makipos lamu palino.* "The wind's petting the panda."
d. *Palino lamuto (sa kelinak).* "The panda's being petted (by the woman)."

I'm fairly certain that such a language as that in (37) doesn't exist, but it could. For that reason, I wanted to bring it up. And that, unless I think of something else later on, will finally conclude this section on semantically-based split ergativity.

4.4 Animacy-Based Split-Ergativity:

It's been alluded to several times in the text above, so here it is: The section on animacy. Animacy really interested me for a long time because I didn't understand it. I don't claim to be a master on the subject now, but I do understand what people say about it. I've also intended Sheli to be a language that's sensitive to the animacy of its subjects and objects.

Anyway, so a quick question: What do people mean when they discuss animacy as it relates to language? Well, some languages encode animacy into their grammar. It can be done in many different ways, some of which aren't related to ergativity, per se. The essential point is this. Let's say you have a verb and two noun phrases. Let's say they're this: "eat", "sandwich", "man". In English, these can be arranged in two different ways, giving you "The man eats the sandwich", or "The sandwich eats the man". But leaving out cartoonish contexts, which one of these sentences is *really* the more likely to be uttered by a human being? Chances are, it's the first one. This is because (speaking of reality as we know it), it's not only possible, but highly probable, that a human will eat a sandwich. It is impossible, though (or, at the very least, highly improbable), for a sandwich to eat a human. For that reason, is it even necessary to say which is the direct object and which is the subject, in any way (either with cases or word order)? According to a lot of languages, no. (For a fascinating example, see Payne's discussion of the language Sierra Popoluca in his book *Describing Morphosyntax*.)

So, how does this relate to ergativity? Well, some languages use animacy to split up case assignment. Thus, some types of arguments will get one type of marking, and the rest will get the other type of marking. So here's a simple example:

38.

a. *Kelina lamu hopokor.* "The woman's petting the man."
b. *Hopoko lamu kelinar.* "The man's petting the woman."
c. *Kelina lamu palino.* "The woman's petting the panda."
d. *Palinom lamu kelinar.* "The woman's petting the panda."

e. *Palinom lamu kitapo.* "The panda's petting the book."
f. *Kitapom lamu palino.* "The book's petting the panda."

In the example above, human beings are marked with a nominative-accusative system, and everything less animate than a human is marked with an ergative-absolutive system. The result is that in a sentence like (38c), the subject and object are marked with the same case. But this isn't a problem. Why? Because the more likely subject is the most animate one, which is the woman. Thus, it doesn't matter that there seems to be fixed word order in the sentences above. All six sentences below in (39) could *only* mean "The woman's petting the panda":

39.

a. *Kelina lamu palino.* "The woman's petting the panda."
b. *Palino lamu kelina.* "The woman's petting the panda."
c. *Kelina palino lamu.* "The woman's petting the panda."
d. *Palino kelina lamu.* "The woman's petting the panda."
e. *Lamu kelina palino.* "The woman's petting the panda."
f. *Lamu palino kelina.* "The woman's petting the panda."

In fact, a language that uses this system has the advantage of achieving relatively free word order without having heavy-handed case marking like a language like Zhyler (cases *everywhere* in that language! And it doesn't even have free word order!).

That's the basic idea behind an animacy system as it relates to case marking. So, a question: Is this the only way it can be split (i.e., one type of marking for humans, another type for the rest?). Absolutely not. So what are the ways to split it up? Well, there are two answers. The first is: Anyway you can imagine it. If you can dream it up, it's possible. Now, what's common among natural languages? For that there's a different (and rather definite, it seems) answer. According to Payne, there's a grand hierarchy of agent worthiness which I will try my darndest to reproduce here (I think I'm going to need to use a table...):

40.

1 > 2 > 3 > 1 > 2 > 3 > Proper Name s >	Humans > Non-Human Animates > Inanimates
Agreement > Pronouns	Definte > Indefinite

So...as I understand it...the table above is... Hmm... Okay, I get it. Odd he did it that way, though... Okay, the reason that 1, 2 and 3 are up there twice, is because the *first* set of 1, 2 and 3 refer to first, second and third person verbal agreement markers. The *second* set refers to pronouns. I guess it would've been too difficult to repeat everything after "proper names" twice, though, because those only appear once. Essentially, this is how to read that table. Let's take "proper names". Proper names will always be considered to be of higher animacy than humans, non-human animates and inanimates (regardless of definiteness [I guess in this table, proper names are always assumed to be definite--not

necessarily an uncontroversial claim]). However, both pronominal verbal agreement, and personal pronouns will be considered more animate than proper names. For that reason, if you had a proper name and a pronoun as two arguments, the pronoun would be construed as being the subject, and the proper name the object (to indicate otherwise, an inverse marker, or something like it, would be required).

This relates to case marking because of a universal claim that Payne makes. So let's say that in a given language, everything to the left of proper names will be marked one way, and everything that's to the right of the last 3 will be marked a different way. According to Payne, it will *always* be the case that what's to the left of "proper names" will be marked with a nominative-accusative system, and what's to the right of the last 3 will be marked with an ergative-absolutive system. Why? I can't seem to find a good answer. I'm sure something metaphysical can be guessed at, though.

Anyway, I could spend a long time showing you every possible example of where the hierarchy could be split, but instead I'll show you just one interesting example. This is an Ergato version of a language Payne describes called Cashinawa. Cashinawa has a system where first and second person pronouns are marked one way, third person pronouns another way, and full NP's are marked yet another way. Here's what that might look like in Ergato:

41.
   a. *Ko sapu.* "I'm sleeping."
   b. *Ko lamu per.* "I'm petting you."

So those are the first and second person pronouns, and they're marked with a nominative-accusative system. Now here are the third person pronouns:

42.
   a. *Li sapu.* "She's sleeping."
   b. *Lim lamu lir.* "She's petting her."

Above you have a three-way system, where each argument is marked differently. Again, this is only with third person pronouns. Now here's what the NP's look like:

43.
   a. *Kelina sapu.* "The woman's sleeping."
   b. *Kelinam lamu hopoko.* "The woman's petting the man."

And, to round it off, the NP's are marked with an ergative-absolutive system. Now, here's something to notice: To what does the pronoun *li* refer in the sentences in (42)? I guess the default assumption would be a human, but there's no reason why it couldn't be a female panda, or some other female animal. Despite the semantics of its referent, though, the pronoun will *always* be higher up in the hierarchy. This is why Payne objected to the terms "agentivity hierarchy" and "animacy hierarchy". It doesn't *really* depend on the animacy of the referent--or, at least in this system. Rather, it depends on the

morphological status of the argument. In that way, a less-animate third person pronoun will be higher up in the topic-worthiness hierarchy than an animate human NP. Now, it doesn't *have* to work this way for a conlang. You could easily imagine a system like this:

44.
   a. *Li sapu.* "She (human)'s sleeping."
   b. *Li sapu.* "She (animal)'s sleeping."
   c. *Li lamu lir.* "She (human)'s petting her (human)."
   d. *Li lamu li.* "She (human)'s petting her (animal)."
   e. *Lim lamu lir.* "She (animal)'s petting her (human)."
   f. *Lim lamu li.* "She (animal)'s petting her (animal)."

A system like that above would surely help to disambiguate pronouns in certain situations. But, then again, you might have a whole different set of pronouns for different types of NP's. After all, in English we have "he", "she" *and* "it".

Another thing to remember is that these claims of universality are for the natural languages spoken on this planet we live on. One can easily imagine a language spoken by a race of intelligent (yet still quite cleanly) cats. In this language, perhaps there would be a new category: sentient non-humans. And perhaps NP's referring to sentient non-humans would be higher up in the hierarchy than humans. Additionally, there's always androids and robots, or talking trees. Or one can also imagine a highly-sexist matriarchal society where women are seen as more animate (and more worthy of being the topic of discussion) than men, dividing humans into male humans and female humans (and maybe the same is true of animals and pronouns). Thus, maybe a female flea would be considered more animate than a male human. The possibility for flux in the hierarchy is limited only by the reality you want your language to live in. So in that respect, think of the above as a guide, rather than a set of rules to follow.

5.0 MIXING SYSTEMS:

To quote the great linguist Thomas Wier, "every language shows some features of ergativity and some features of accusativity" (click here for that discussion). Thus, a good system will include some elements from *all* the sections discussed above. I've already mentioned (dozens of times) how English makes a distinction between experiencer and non-experiencer verbs in the present tense. Another famous example is the *-ee* suffix, summarized below:

45.
   a. *Escape* (intransitive verb) + *ee* = *escapee*, "one who escapes" (nominalizes intransitive subject)
   b. *Nominate* (transitive verb) + *ee* = *nominee*, "one who is nominated" (nominalizes transitive object)
   c. *Nominate* (transitive verb) + *or* = *nominator*, "one who nominates" (nominalizes transitive subject)

In the example above, you can see a clear ergative-accusative pattern. This only applies to one tiny little corner of English grammar, but, then again, the same can be said of experiencer verbs in the present. This is part of what goes into creating a realistic language. Not everything is perfect, and not every pattern jumps out and draws attention to itself. Another simple pattern from a natural language can be seen with French. In French, there's a distinction in (what is now) the simple past tense between verbs that take an SA and verbs that take an SP. Take a look at this example:

46.

a. *J'ai dormi.* "I slept." (SA)
b. *Je suis arrivé.* "I arrived." (SP)

In the example above, the subject is enacting the sleeping event (to an extent), whereas in the second sentence, the verb is something that happened to the subject. "Appear" is another verb like this.

There are many, many ways you could create a mixed system. One way might be to have a nominative-accusative system to mark pronouns in the present tense, and an ergative-absolutive system to mark NP's in the present, while *all* arguments, pronoun and NP alike, are marked with an ergative-absolutive system in the past tense. And then maybe, in all tenses, the cases are flipped for verbs of experience (i.e., nominative marks pronoun stimuli, and accusative marks pronoun experiencers, in the present, and everywhere else, the ergative case marks stimuli, and the absolutive marks experiencers). The theoretical possibilities are endless (though certain possibilities become more difficult to justify linguistically than others).

6.0 Something Else to Consider: Ditransitives:

One thing that often gets ignored in a discussion of ergativity is the marking of secondary objects in ditransitive clauses. As it turns out, it's by no means simple. Below I'll summarize a description of possible types of indirect object marking laid out explicitly in a paper by Matthew S. Dryer entitled "Clause Types" (warning: that link is to a .pdf).

So far in the nominative-accusative ditransitive examples I've shown, the direct object (P) has always been marked with the accusative case *-r*, and the indirect object (R) has always been marked with the dative case *-s*. Does this *necessarily* have to be the (excuse the pun) case, though? As it turns out, no. Actually, there are three different possibilities. First let's detail the common (to us) pattern. This is a pattern like Latin. This is an example where the direct object of a transitive verb is grouped together with the direct object of a ditransitive verb:

47.

a. *Kelina sapu.* "The woman's sleeping."
b. *Kelina lamu palinor.* "The woman's petting the panda."
c. *Kelina kanu kitapor palinos.* "The woman's giving a book to the panda."

The ordering of the indirect object and direct object in (47c) can vary, but nevertheless, this is a very Latinate kind of pattern. Now let's take a look at a different kind:

48.
   a. *Kelina sapu.* "The woman's sleeping."
   b. *Kelina lamu palinor.* "The woman's petting the panda."
   c. *Kelina kanu palinor kitapos.* "The woman's giving a book to the panda."

In the example above, the cases on the objects of *kanu*, "to give", flip-flopped (as did the order, just to keep everything in line). A language that does ditransitives like this will usually mark that last argument with an instrumental, as opposed to a dative, case. Nevertheless, it is a different case, as opposed to an oblique, like in the English "I gave the book to her". In that English example, the "to her" part isn't as much a part of the argument structure as the R is in the counterpart sentence "I gave her the book".

For a final example, we can see a pattern that looks a lot like the last English example I gave:

49.
   a. *Kelina sapu.* "The woman's sleeping."
   b. *Kelina lamu palinor.* "The woman's petting the panda."
   c. *Kelina kanu palinor kitapor.* "The woman's giving a book to the panda."

As you can see, now there's only two cases operating in the (c) sentence. How do you know which is the direct object and which the indirect object? Strict word order. So, in the above example, there'd be some kind of rule that states that the first object in a ditransitive clause would be interpreted as the indirect object, and the second the direct object. This is exactly how it works in English, in a phrase like, "You gave me him" (an odd sentence, I know. And why? Because of animacy!), "me" is *always* interpreted as the indirect object, and never as the direct object. (Note: There are dialects where the opposite is still productive, thus the indirect object in, "Give it me, I say!" is "me", not "it".)

So those are three possibilities for nominative-accusative systems. What about ergative-absolutive systems? Well, there's three possibilities for them, as well, and they match up nicely with the three systems above.

The first ergative-absolutive system is one where the absolutive argument of a transitive clause is marked the same as the direct object of a ditransitive clause. This is what it looks like:

50.
   a. *Kelina sapu.* "The woman's sleeping."
   b. *Kelinar lamu palino.* "The woman's petting the panda."
   c. *Kelinar kanu kitapo palinos.* "The woman's giving a book to the panda."

This should look just like the system in (47), only with *-r*'s flipped around. This would be like ergative Latin, which I call Nital. Pretty straightforward. Next system:

51.
   a. *Kelina sapu.* "The woman's sleeping."
   b. *Kelinar lamu palino.* "The woman's petting the panda."
   c. *Kelinar kanu palino kitapos.* "The woman's giving a book to the panda."

Again, this is like the examples in (48). Perhaps a helpful way to think of the ditransitive verbs in sentences like these is that *kanu* isn't defined as "to give (something)", but rather "to give to (someone)". The extra case, then, specifies what's being given (again, usually something like an instrumental). Now for the last example:

52.
   a. *Kelina sapu.* "The woman's sleeping."
   b. *Kelinar lamu palino.* "The woman's petting the panda."
   c. *Kelinar kanu palino kitapo.* "The woman's giving a book to the panda."

And, again, the way you tell which object is which in (52c) is strict word order.

That wraps up this discussion of ditransitives. There's more to them, to be sure, but this is all that presently concerns us. Again, it's just something to think of. The status of indirect objects is something I certainly didn't think about in many of my languages, and I believe they're the less realistic for it.

7.0 IMPOSSIBILITIES:

There are certain patterns deemed to be impossible, which makes them immediately interesting. I'll just mention them here.

One that I may have mentioned already has to do with split-tense systems. In all the split-tense systems that have been found, the present tense has a nominative-accusative pattern, and the past tense has an ergative-absolutive pattern. Based on this evidence, experts have deemed the opposite impossible. While it may be easier to come up with a historical explanation for the opposite, it's by no means unworkable.

Related to tense, if you read up on this stuff, you'll notice that the only tenses that are mentioned are present and past, or, at the most, past and non-past. The future tense is never discussed. And I'm sure any conlanger can think up more tenses than even past, present and future. As far as I know, there are no universals for what kind of marking you get in the future (well, except maybe that it probably looks like the present). That's something to think about.

Let's say that we are working with just past, present and future (no aspect). That's three tenses. The reason why nominative-accusative and ergative-absolutive works so well with present and past tense is because they line up: Two systems, two tenses. But what do

these terms stand for? In a sentence with three basic arguments, S, A and P, nominative-accusative stands for the system that groups S and A together to the exclusion of P. Ergative-absolutive, on the other hand, stands for a system that groups S and P together to the exclusion of A. Do you see what I see? There's a third pattern not mentioned here, and, coincidentally, a third tense that doesn't get to play. So imagine, if you will, the following: Nominative-accusative in the present; ergative absolutive in the past; and in the future (using *-sa* as an impromptu future marker)...!

53.

a. *Kelinar sapusa.* "The woman's gonna sleep."
b. *Kelina lamusa palino.* "The woman's gonna pet the panda."

Oh, yeah! This is a system that, paradoxically, groups A and P together to the exclusion of S! This kind of system is unattested in natural languages, and judged impossible. Thus (to my knowledge), it hasn't been *officially* named. Therefore, I'm going to name it. What ties together the subject of a transitive verb and the patient of a transitive verb...? Well, how about this: In a transitive clause, there are two arguments; in an intransitive, there's one. Thus, the case assigned to both the subject and object of a transitive verb is the *duative*, and the case assigned to the single argument of an intransitive verb is the *unitive*. Yeah! That sounds good. Thus, I dub the above pattern a duative-unitive system. I named them this way because the pattern seems to be that the case that's assigned to the subject of a transitive verb is the one that goes first. Hee, hee... Now I wish I had a language that used this pattern. I'll have to work on that...

(Quick Note: On the CONLANG list, this pattern was dubbed the "Monster Raving Loony", or MRL, pattern. The case names were called the "intransitive" and "transitive" cases. I don't like this naming strategy, because both "inransitive" and "transitive" already mean something, and confusion could easily ensue. Go here to see the various related posts.)

Some other impossibilities have been touched on in the animacy section. Here's an idea. Referring to the hierarchy mentioned in the animacy section above, why not have *two* splits. And not like the kind I described for the Cashinawa system. This is a system where the section in the middle is marked one way, and the sections on either end are marked another way. So let's say that all pronouns are marked with a nominative-accusative system, as are everything to the right of humans, and then humans and proper names are marked with an ergative-absolutive system. That would be strange, and definitely would violate the universal Payne proposed.

Another impossibility one can imagine is with ditransitives. In all six examples above, the indirect object and direct object could be marked in various ways, but they were always marked differently from the subject. Why not mark the indirect object the same way as the subject? In fact, let's do these three possibilities with a duative-unitive system, just for kicks:

54.

a. *Kelinar sapu.* "The woman's sleeping."
b. *Kelina lamu palino.* "The woman's petting the panda."
c. *Kelina kanu kitapo palinos.* "The woman's giving a book to the panda."

In this pattern, the direct object of both transitive and ditransitive verbs are treated alike. And, as you can see, they're both marked with the duative case. The subjects of the transitive verbs are as well. The subject of the intransitive is marked with the unitive, and the indirect object in (54c) is marked with the dative. Now for the next one:

55.

a. *Kelinar sapu.* "The woman's sleeping."
b. *Kelina lamu palino.* "The woman's petting the panda."
c. *Kelina kanu palino kitapos.* "The woman's giving a book to the panda."

Same thing here as with the "give to (someone)" verbs we've seen before, where the R is assigned the objective case, which is in this case the duative. And here, the *-s* probably stands for an instrumental case. Last one:

56.

a. *Kelinar sapu.* "The woman's sleeping."
b. *Kelina lamu palino.* "The woman's petting the panda."
c. *Kelina kanu palino kitapo.* "The woman's giving a book to the panda."

And this is about as duative as you get. Here the subject of the intransitive verb in (56a) is marked with the unitive, and *everything else* is marked with the duative, the status of each object being determined by word order in (56c).

Oh, one thing I forgot about: What about a valency reduction system in a duative-unitive system? This would be odd, because in this case (and in this case only), the case that would be reduced would be the unmarked/default case, rather than the marked/special case. (Well, that is *if* the duative is the unmarked case.) Anyway, the result is that the transitive verb becomes intransitive, and the duative argument becomes a unitive argument. But *which* duative argument?! You don't know. Therefore, the resulting verb would mean something like, "Y is a participant (either agent or patient) in an X action". Thomas Wier suggested this might be like the Ancient Greek middle voice construction (see his post to CONLANG by clicking here). In any case, here's what it'd look like in Ergato:

57.

a. *Kelina lamu palino.* "The woman is petting the panda."
b. *Kelinar lamuto (palinok).* "The woman's petting (the panda)/being petted (by the panda)."
c. *Palinor lamuto (kelinak).* "The panda's petting (the woman)/being petted (by the woman)."
d. *Kelina hopokos kanu kitapo.* "The woman's giving the book to the man."

e. *Kelinar hopokos kanuto (kitapok).* "The woman is giving to the man (and what she's giving is a book)/being given to the man (by the book)."
f. *Kitapor hopokos kanuto (kelinak).* "The book is giving to the man (and what it's giving is a woman)/being given to the man (by the woman)."

Given a system like the above, one can easily imagine that discourse context and animacy would help you decide which reading is the correct one (for example, if giving is the act, and you're talking about a woman and a book, it's pretty likely that the book's the one being given). Anyway, that's what a duative-unitive system would look like, in toto (I believe). As for the valency-reduction system, if you already have passive and antipassive, then I propose that the name of this system should be an *ambipassive*, since it can apply to either of the arguments in a transitive clause.

Here's a thought I don't think I've run across before: What if the subjects of intransitive verbs, tranisitive verbs, and ditransitive verbs *all* had different subject marking? This would be treating the subjects of ditransitive verbs as something inherently different from transitive verbs. This is probably unattested, but nevertheless, a possible pattern.

Those are some ideas to mull over. There's a lot more that's possible than is attested in the world's languages (though they *do* do a lot more than most universalists would have you believe).

8.0 CONCLUSION:

The intention of this section has been to document the basics of ergativity. It's my hope that this is a starting point. If you have more information, or if you think I've made a mistake (or if you spot any typos--I know there are tons!), my hope is that you'll e-mail me, so that I can further improve this section. Though I did write all this, I prefer to think of this as a collaborative effort, since I got my information from many different sources. I hope you've got something from this section on ergativity, and that if you have something to share, you'll let me know, so I can make improvements in the future.

9.0 REFERENCES AND THANKS:

These are a list of references I used and some shout outs:

- Bell, David. *ámman îar Reference Grammar*. http://www.graywizard.net/Conlinguistics/amman_iar/amman_iar.htm, 2004.
- Dryer, Stephen S. "Clause Types". SUNY Buffalo, 2001. Download .pdf
- Henning, Jeffrey. *Langmaker.com*. http://www.langmaker.com, 2004.
- Hiller, P.J. *The Georgian Language: An outline grammatical summary*. http://www.armazi.com/georgian/, 2004.
- Loos, Eugene E. (gen.ed.). *Glossary of linguistics terms.* http://www.sil.org/linguistics/GlossaryOfLinguisticTerms/Index.htm, 2004.
- Payne, Thomas E. *Describing Morphosyntax: A Guide for Field Linguists.* Cambridge University Press, 1997. Amazon.com Info

- Rye, Justin B. *Learn Not to Speak Esperanto: Case.* http://www.xibalba.demon.co.uk/jbr/ranto/r.html, 2004.
- Wier, Thomas. *Re: ergative + another introduction*. http://listserv.brown.edu/archives/cgi-bin/wa?A2=ind0411c&L=conlang&F=&S=&P=31479, 2004.
- Wier, Thomas. *Re: Ergativity Reference Done*. http://listserv.brown.edu/archives/cgi-bin/wa?A2=ind0411D&L=conlang&P=R12196, 2004.

I'd like to thank all those who contributed to the recent discussion of ergativity on the CONLANG list (well, recent as of November 28, 2004), as well as all those who've discussed ergativity many, many times on CONLANG over the years. In particular, I'd like to thank Thomas Wier for reminding me of the *escapee* example, which, despite its fame, always seems to elude me in times of need. I'd also like to thank Roger Mills for reminding me of David Bell's section on ergativity in ámman îar. I'd also like to thank Taliesin for his design advice. (As you can probably tell, I'm not too good a judge of what is and is not easy to read on the screen.) And, of course, I'd like to thank Christophe Grandsire for providing me with webspace. *Vive la France!*

The Language Creation Kit - http://www.zompist.com/kit.html

© Mark Rosenfelder - markrose@rcn.com

Models

Natural and unnatural languages

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

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

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

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

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

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

Sounds

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

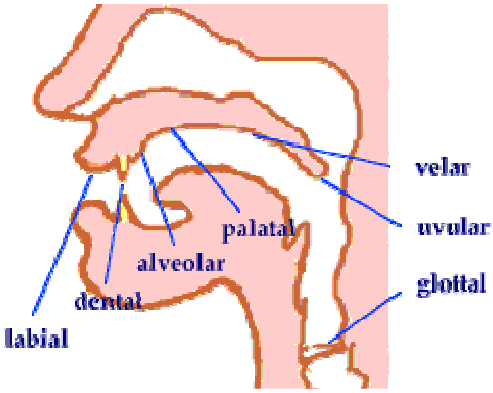
You'll get better results the more you know about **phonetics** (the study of the possible sounds of language) and **phonology** (how sounds are actually used in language). Useful

references are J.C. Catford, *A Practical Introduction to Phonetics* (excellent for home study), and Roger Lass, *Phonology*. Below is a quick overview.

TYPES OF CONSONANTS

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

- **Place of articulation**-- where the obstruction occurs:
  - ***labial***: lips (w), lips + teeth (f)
  - ***dental***: teeth (th, French or Spanish t)
  - ***alveolar***: behind the teeth (s, English t, Spanish r)
  - ***palato-alveolar***: further back from the teeth (sh, American r)
  - ***palatal***: top of palate (Russian ch)
  - ***velar***: back of the mouth (k, ng)
  - ***uvular***: way back in the mouth (Arabic q, French r)
  - ***glottal***: back in the throat (h, glottal stop as in John Lennon saying bottle).



- **Degree of closure**. This proceeds in steps
  - from ***stops*** (stopping the airflow entirely: p t k)
  - to ***fricatives*** (impeding it enough to cause audible friction: f s sh kh)
  - to ***approximants*** (barely impeding it: r l w y).
  - An ***affricate*** is a stop plus a fricative, which must occur at the same place of articulation: t + sh = ch, d + zh = j.
- **Voicing**: whether the vocal cords are vibrating or not. That's the difference between f and v, t and d, k and g, sh and zh.
- **Nasalization**: whether air travels through the nose as well as the mouth. For instance, m, n, and ng are stops like b, d, g, but only the oral airflow is stopped.
- **Aspiration**: whether stops are released lightly, or with a noticeable puff of air. In Chinese, Hindi, or Quechua, there are series of aspirated and non-aspirated stops.
- **Palatalization**: whether the tongue is raised toward the top of the mouth while pronouncing the consonant. In Russian and Gaelic, there are distinct series of palatalized and non-palatalized consonants.

English consonants can be arranged in a grid like this:

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

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

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

INVENTING CONSONANTS

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

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

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

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

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

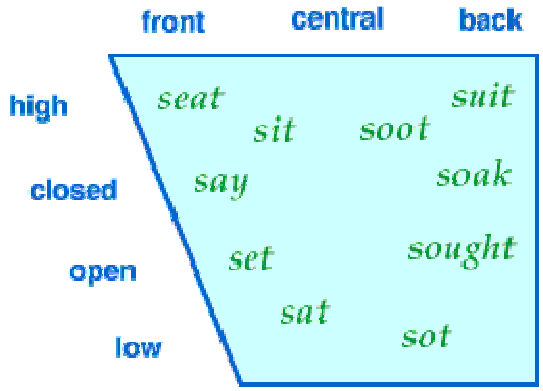
Some consonants are more common than others. For instance, virtually all languages have the simple stops p t k. Lass's book gives examples; see also David Crystal's *The Cambridge Encyclopedia of Language,* p. 165.

VOWELS

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

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

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



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

Vowels can vary along other dimensions as well:

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

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

English has a rather complicated vowel system:

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

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

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

STRESS

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

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

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

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

TONE

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

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

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

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

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

PHONOLOGICAL CONSTRAINTS

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

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

Start with a distinctive syllable pattern. For instance,

- Japanese basically allows only (C)V(V)(n): *Ranma, Akane, Tatewaki Kunoo, Rumiko Takahashi, Gojira, Tookyoo, konkuuru, sushi,* etc.
- Mandarin Chinese allows (C)(i, u)V(w, y, n, ng): *wô, shì, Mêiguó, rén, wényán, chìàn, mànhuà, Wáng, Zhang,* etc.

- Quechua allows (C)V(C): *Wallpakuna sarata mikuchkanku, achka allin hatun mosoq puka wasikuna,* etc.
- English goes as far as (s) + (C) + (r, l, w, y) + (V) + V + (C) + (C) + (C): *sprite, thinks.*

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

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

ALIEN MOUTHS

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

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

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

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

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

Alphabets

ORTHOGRAPHY

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

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

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

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

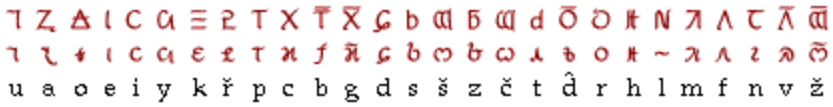
Or you may go for neatness, as I did in inventing Verdurian. I don't like digraphs, so I adapted Czech orthography-- **č**for ch, **š**for sh, etc. This ultimately involved creating a special Macintosh font, so I was probably crazy. (Note however that fonts for non-Western-European languages are plentiful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

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

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

AN EXAMPLE

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



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

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

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

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

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

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

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

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

DIACRITICS

Some advice: never use a diacritical mark without giving it a specific meaning, preferably one which it retains in all uses. I made this mistake in Verdurian: I used ö and ü as in German, but ë somewhat as in Russian (indicating palatalization of the previous consonant), and ä as a mere doubling of a. I was smarter by the time I got to Cuêzi: the circumflex consistently indicates a low-pitch accent.

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

FANCIER WRITING SYSTEMS

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

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

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

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

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

Word building

HOW MANY WORDS DO YOU NEED?

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

Long before, Ogden and Richards had shown that eight hundred and fifty words were sufficient vocabulary to express anything that could be expressed by "normal" human vocabularies, with the aid of a handful of special words-- a hundred odd-- for each special field, such as horse racing or ballistics. About the same time phoneticians had analyzed all human tongues into about a hundred-odd sounds, represented by the letters of a general phonetic alphabet.
... One phonetic symbol was equivalent to an entire word in a "normal" language, one Speedtalk word was equal to an entire sentence.
--"Gulf", in *Assignment in Eternity*, 1953

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

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

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

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

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

However, go ahead and play with the Speedtalk idea. It's good for some hours of fun, working out as minimal a set of primitives as you can; and the habit of paraphrase it gives you is very useful in creating languages. Just don't take it too seriously; if you do, your

punishment is to learn 850 words of any actual foreign language and be set down in a city of monolingual speakers of that language.

ALIEN OR *A PRIORI* LANGUAGES

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

This can fairly quickly get tiresome. I don't advise you to sit down and come up with a hundred words at once; you're likely to run out of inspiration, or find that all the words are starting to sound the same. You may also be creating new roots where you could more easily derive the word from existing roots.

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

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

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

A FEW HALF-RECOGNIZABLE BORROWINGS

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

Sul Aḑ e otál mudray dy tü, dalu esë, er ya ceč̣el rho sen e sënul.
Only God is as wise as you, my king, and even there I'm not certain.
So cuon er so ailuro eu druki. Cuon ride še slušir misotém ailurei. So ailuro e araš́ó rizuec.
The dog and the cat are friends. The dog laughs at the cat's jokes. The cat is quite amusing.

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

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

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

LANGUAGES BASED ON EXISTING LANGUAGES

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

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

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

SOUND SYMBOLISM

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

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

Inventing alien languages, authors also simply make use of what we might call phonetic stereotypes. Tolkien's Orkish, for instance, makes heavy use of guttural sounds and is full

of consonants, while his Elvish tongues are more vocalic, and seem to have plenty of pleasant-sounding l's and r's.

SOME GUIDELINES FOR NOT REINVENTING THE ENGLISH VOCABULARY

- If the literal meaning of an expression doesn't make sense (e.g. "make good", "go all out", "have it in for someone", "look lived-in"), you're probably dealing with an idiom. Translate using expressions that make sense literally ("succeed", "work at full capacity", "have a grudge against someone", "seem inhabited"), or create your own idioms ("laugh at hell", "play bee", "circle your eye at someone", "be breathed and worn").
- Look through the foreign-to-English section of a bilingual dictionary. Look at the range of English meanings particular foreign words have: think about what kind of root concept could cover all of them. Look at the foreign words used to translate a single English word: try to see what distinctions the foreign language is making where English uses that one word.
- Derive your lexicon from basic roots using regular derivation processes.
- Look up the etymology of the English word. See if you can come up with an alternative process.
- Consider a whole class of related English words-- verbs of motion, for instance. Design the related class of words in your language, dividing up the conceptual space in your own way.
- Read Lakoff and Johnson, *Metaphors We Live By*. Create your own metaphors and the vocabulary that goes with them.
- Read a text on semantics (Palmer's *Semantics* is short; Takao Suzuki's *Japanese and the Japanese: Words in Culture*, a.k.a. *Words in Context*, is wonderful), for a greater awareness of the structure of the lexicon.
- For a fantasy language, think about the culture that your language serves. What concepts are most important to it? They will likely have many synonyms, or even be reflected directly in the grammar. What's its history or mythology? They will probably generate a number of derived words.

Grammar

Once you've bundled together some words and perhaps an alphabet, you may think you're done. If you do, it's likely that you've just created an elaborate cipher for English. You still have the grammar to do, bucko.

This section doesn't attempt to cover all the issues in morphology, syntax, and pragmatics. Instead, it suggests what your grammar should minimally do, mentions some of the issues, and lists some interesting approaches taken by various languages.

IS YOUR LANGUAGE INFLECTING, AGGLUTINATING, OR ISOLATING?

Inflections are of course affixes used to conjugate verbs and decline nouns. Examples from English are the -s we add to verbs for the 3rd person present form, the -s added to pluralize nouns, and the -ed of the past tense. Languages such as Russian or Latin have complex, not to say baroque, inflectional systems.

A single inflection may encode multiple meanings. For instance, in the Russian form *domóv,* the *-óv* ending indicates both plurality and the genitive case; it doesn't bear any evident relationship with other plural endings (e.g. nominative *-á*) or the singular genitive ending (*-a*). In Spanish *comí* 'I ate', the *-í* ending indicates the 1st person singular, past tense, indicative mood-- quite a job for one vowel, even accented.

In **agglutinating** languages, one affix has one meaning. Compare Quechua *wasikunapi* 'in the houses'; the plural suffix *-kuna* is separate from the case suffix *-pi.* Or *mikurani* 'I ate', in which the past tense suffix *-ra-* is kept separate from the personal ending *-ni.*

In **isolating** languages, there are no suffixes at all; meanings are modified by inserting additional words. In Chinese, for instance, *wô chi fàn* could mean 'I eat' or 'I was eating', depending on the context; the verb is not inflected at all. For precision, adverbs can be brought in: *wô chi fàn zuótiàn* 'I was eating yesterday'.

(In practice natural languages are all a bit mixed; some inflections have a single meaning; Quechua does have a few inflections, for instance, and Chinese does have required grammatical particles, such as the aspect particle *le*, used to show completed action: *wô chi fàn le* 'I ate.')

Conlang creators seem to gravitate toward agglutinating or isolating languages; but there's something to be said for inflections. They tend to be compact, for instance. You can't beat *-í* for succintness.

DO YOU HAVE NOUNS, VERBS, AND ADJECTIVES?

Why not get rid of one or two of them?

It's not hard to get rid of **adjectives**. One easy way is to treat them as verbs: instead of saying "The wall is red", you say "The wall reds"; likewise, instead of "the red wall" you say "the redding wall".

With such tricks you can even get rid of the verb **be**, which according to some theorists is responsible for most of the sloppy thinking in the world today. (Heinlein was careful to

ban 'to be' from Speedtalk.) About the only response this notion deserves is: would that clear thinking was that easy.

You can extend the idea to get rid of **nouns**. For instance, in Lakhota, ethnic names are verbs, not nouns. There's a verb 'to be a Lakhota': the present forms mean 'I am a Lakhota, you are a Lakhota, etc.'

You can have some fun with this. "The rock is under the tree" could be expressed as something like "There is stonying below the growing, greening, flourishing",or perhaps "It stones whileunder it grows greeningly." If we really encountered a language like this, however, I'd have to wonder whether we weren't just fooling ourselves. If there's a word that refers to stones, why translate it as 'to stone' rather than simply 'stone'?

Jorge Luis Borges, in "Tlön, Uqbar, Tertius Orbis", posits a language without nouns; but this was because its speakers were Berkeleyan idealists, who didn't believe in object permanence. However, linguists really do not like using semantic classes-- or metaphysics-- to define syntactic categories. (It's not the right level of analysis; and it tends to obscure how languages really work by making them all look like Latin.)

Jack Vance (in *The Languages of Pao*) posited a language without **verbs**. For instance, "There are two matters I wish to discuss with you" comes out something like "Statement-of-importance -- in-a-state-of-readiness-- two; ear-- of [place name]-- in-a-state-of-readiness; mouth-- of this person here-- in-a-state-of-volition." Vance may be in a state of pulling our legs.

HOW DO YOU INDICATE PLURAL, CASE, AND GENDER FORMS OF ADJECTIVES AND NOUNS?

What's case? It's a way of marking nouns by function: e.g. Latin

```
mundus   subject or nominative: the world (is, does, ...)

mundum   object or accusative: (something affects) the world

munde    vocative: O world!

mundi    possessive or genitive: the world's

mundo    indirect object or dative: (given, sold, etc.) to the world

mundo    ablative: (something is done) by the world
```

English actually has cases: possessives like 'world's' are actually genitive case forms; while the subject/object distinction is made with pronouns (I vs. me, we vs. us).

Conlang enthusiasts generally either love case (because it makes a language compact and frees up word order) or hate it (because English doesn't do much with it).

Some languages, such as Basque, have a different arrangement of cases. Instead of the subject of the sentence always being in the same case (the nominative), the *subject* of *intransitive* sentences (e.g. "The *window* broke") and the *object* of *transitive* sentences (e.g. "I broke the *window*) are in the same case, the **absolutive**, while the subjects of transitive sentences (e.g. "*I* broke the window") are in the **ergative** case.

If you think that's weird, a few languages, such as Dyirbal, use the nominative/accusative system for 1st and 2nd person pronouns (I, we, you), and the ergative/absolutive system for nouns and for 3rd person pronouns.

If a language doesn't have case it may rely on word order to indicate the relationship between a verb's arguments; but there is another alternative: **head-marking** on the verb. For instance, in the Swahili *Kitabu umekileta?* 'Did you bring the book?', the verb *leta* has prefixes indicating the subject (*u-* 'you') and the object (*-ki-*, a third person prefix agreeing in gender with *kitabu*). (*-me* marks the perfect tense.) The gender-specific object marker on the verb allows free word order even without case marking on the nouns.

DO NOUNS HAVE GENDER?

Note that gender need not be simply masculine/feminine. Swahili, for instance, has eight gender classes, none of them masculine/feminine: one is for animals, one for human beings, one for abstract nouns, one forms diminutives, etc.

I daresay not many conlangs have grammatical gender. (Verdurian has it, because it's intended to be naturalistic.) People ask, what is gender **for**? Gender is remarkably persistent: it's persisted in the Indo-European, Semitic, and Bantu language families for at least five thousand years. It must be doing *something* useful.

A few possibilities:

- It helps tie adjectives and nouns together, reducing the functional load on word order and adding useful clues for parsing.
- It gives language (in John Lawler's terms) another dimension to seep into. In French, for instance, there are many words that vary only in gender: *port/porte, fil/file, grain/graine, point/pointe, sort/sorte,* etc. Changing gender must have once been an easy way to create a subtle variation on a word.
- It allows indefinite references to give someone's sex.
- It offers some of the advantages of obviative pronouns (see below): one may have two or more third person pronouns at work at the same time, referring to different things.
- It can support free word order without case marking, as in the Swahili example above.

DOES THE VERB INFLECT BY PERSON, GENDER, AND/OR NUMBER?

Like case, **personal endings** make for nice compact sentences, since if you have them you can generally omit subject pronouns.

Some languages, such as Swahili and Quechua, include the **object pronoun** in the verb as well, usually as an infix.

The Romance languages have **clitic** forms of the pronouns, which stop just short of being verb inflections: e.g. French *Je le vois,* 'I see him'; Spanish *Digame,* 'Tell me'.

Basque verbs can inflect to encode information about the **listener**. For instance, *ekarri digute* is a neutral way of saying 'They brought it to us'; *ekarri zigunate* means the same, but also indicates that the listener is a woman addressed with the informal personal pronoun.

WHAT DISTINCTIONS ARE MADE IN THE VERB?

Some distinctions languages make:

- time, of course (**tense** strictly speaking)
- whether the action is completed (grammarians say **perfect**) or not
- whether the focus is on the ongoing process (**progressive**), or a single action, or a habitual action, or a repeated action (all these are **aspects**)
- whether the action can be counted on (**indicative mood**), or is doubtful or merely to be desired (**subjunctive**), or isn't happening at all (**negative**)
- whether I'm telling you (indicative again) or ordering you (**imperative**)
- whether the speaker knows about the action from personal experience, or merely from hearsay, or merely considers it probable (**evidentiality**)
- whether the verb is **intransitive** (it just happens) or **transitive** (it happens **to** something) or **reflexive** (it happens to the subject)
- whether the verb simply describes a state (**static**) or reports a change in state (**dynamic**). In Arabic, for instance, *rukubun* means 'ride' in its static forms, 'mount' in its dynamic forms; *'iqamatun* is static 'reside' and dynamic 'settle'.
- degree of **deference** between speaker and listener

Any language can *express* these distinctions, but they differ in which features are **grammaticalized**: reflected in the morphology and syntax of the language. English, for instance, grammaticalizes person and number in its verbal system, while Japanese does not. On the other hand Japanese verbs have positive and negative forms, as well as a morphological indication of levels of deference.

Languages also differ in how many distinctions are made in these categories.

- There is an Austronesian language which has four past **tenses** (last night, yesterday, near past, remote past) and three futures (immediate, near, remote).
- The languages of the Vaupés river basin distinguish five levels of **evidentiality**: visual perception; non-visual perception; deduction from obvious clues; hearsay; and mere assumption.

WHAT ARE THE PERSONAL PRONOUNS?

The basic, universal persons are first (referring to the speaker), second (the hearer), and third (everybody else). However, there's lots of room to play around. Distinctions may be made:

- by **gender** (not necessarily just in the third person)
- **not** by gender (many languages don't distinguish 'he' and 'she')
- by **number** (I vs. we... sometimes there's special **dual** forms for pairs of things)
- **not** by number (an optional distinction in Chinese)
- by **animacy** (cf. he/she vs. it)
- whether 'we' includes 'you' (**inclusive** we) or not (**exclusive** we)
- by level of **formality** or politeness
- by whether third persons are **present** or not
- between two sets of third persons (**proximate** and **obviative**)-- imagine having two forms of 'he' to distinguish two different persons
- between real and hypothetical reference: e.g. English 'one', French *on*

I invented an alien race once that used different pronouns on land and underwater (they were amphibians), and had the inclusive/exclusive and proximate/obviative distinctions. They also had a pronoun for group minds, and pronouns for each of their three sexes. The complete list was impressive.

WHAT ARE THE OTHER PRONOUNS?

To me, the best idea Zamenhof had was his table of **correlatives**, a nice way to organize all these pronouns. For English, it looks like this:

	QUERY	THIS	THAT	SOME	NO	EVERY
ADJECTIVE	which	this	that	some	no	every
PERSON	who	this	that	someone	no one	everyone
THING	what	this	that	something	nothing	everything
PLACE	where	here	there	somewhere	nowhere	everywhere

TIME	when	now	then	sometime	never	always
WAY	how	thus		somehow		
REASON	why					

It's easy and diverting to regularize the table, although natural languages generally leave holes, which must be filled in with phrases ('in that way', 'for no reason').

You might ask yourself whether the interrogative pronouns ("*Who* did it?") and the relative pronouns ("Is this the man *who* did it?") are the same; in some languages they aren't.

Generally, if nouns decline, these pronouns decline the same way. Sometimes they're worse-- English, for instance, retained separate 'from' and 'to' forms for pronouns of place (here / hence = from here / hither = to here) long after such distinctions were lost for ordinary nouns.

WHAT ARE THE NUMBERS?

Are the numbers based on tens, or something else? Many human number systems are based on fives instead. My pronoun-happy aliens had a duodecimal system. Intelligent machines would surely prefer hexadecimal...

How do you form higher numbers? 'Forty-three', for instance, may be formed in several ways:
forty three
four three
forty with three
three and forty
four tens and three
eight fives and three
fifty less seven
twice twenty and three

Where nouns decline, numbers may also. Or they may not. In Latin, you stop declining the numbers at four.

In Indo-European languages we are used to unanalyzable roots for the numbers; but in other families number names are derivations, often related to the process of counting on fingers and toes-- e.g. Choctaw 5 = *tahlapi* 'the first (hand) finished'; Klamath 8 *ndan-ksahpta* 'three I have bent over'; Unalit 11 *atkahakhtok* 'it goes down (to the feet)'; Shasta 20 *tsec* 'man' (considered as having 20 countable appendages).

For more on numbers, see the Sources page of my Numbers from 1 to 10 in Over 2000 Languages page.

WHAT ABOUT ADJECTIVES?

Adjectives can be something like nouns, something like verbs, or like neither. If they're like nouns, they generally agree with their head noun in gender, case, and number. If they're like verbs, they conjugate like verbs.

How are comparative expressions ("holier than thou", "most holy", "as holy as thou") formed?

It's useful to have some regular derivations for or from adjectives:
opposite (un-)
lack (-less) or surfeit (-ful)
possibility (-able)
liking (-phile) or disliking (-phobe)
inhabitant (-er, -ian, -an, -ese)
weakening of meaning (-ish)
strengthening of meaning (to the max)
adverb (-ly)

ARE THERE ARTICLES (A, THE)?

Many languages, such as Latin and Russian, get by quite happily without them.

It may help to understand what the distinction really means. Ordinarily it's pragmatic: *the* can be paraphrased 'You know which one I'm talking about'. Consider:

I saw a man at the rodeo. The man had on a horrid plaid suit.
A man in the first sentence signals that this character is being introduced in this conversation; *the* in the second sentence signals that he's old news, he is in fact the same guy we just started talking about. *The* before *rodeo* also indicates that the speaker expects that the hearer can figure out which rodeo-- if not, he'd have said *a rodeo*.

Word order serves the same function in Russian. There you'd say, in effect,

I saw man in rodeo. Man wore horrid plaid suit.
When he's introduced, the man lives near the end of the sentence; when he's old news, he appears at the front.

(Actually, they don't have many rodeos in Russia.)

WHAT ORDER DO THE VARIOUS COMPONENTS OF A NOUN PHRASE APPEAR IN?

Consider articles, numbers, quantifiers, adverbs, adjectives, possessives, subordinate clauses-- e.g.

The ten very happy robots who passed the bar exam

You can generally divide phrases into **heads** and **modifiers**. Some languages are very consistent about placing all modifiers before, or all after the head. English is head-final, with the exception of subordinate clauses. Japanese is head-final too, but it's more consistent: it would say "the bar exam passed robots".

WHAT ORDER DO THE VARIOUS COMPONENTS OF A SENTENCE APPEAR IN?

Linguists like to talk about the order of subject, object, and verb, which of course can occur in just six combinations: SVO (as in English or Swahili), SOV (Latin, Quechua, Turkish), VSO (Welsh), OVS (Hixkaryana), OSV (Apurinã), VOS (Malagasy). The last three are for some reason rare, although they do exist.

Combinations and complications are common; for instance, German is basically SOV, but a finite verb (anything but a participle or an infinitive) appears after the subject in a main clause:

Mein Vater *ist* vor einigen Tagen nach London *gefahren*.
My father has several days ago to London travelled.

(German isn't usually described this way; but my way is equally correct, and requires only one exception. The usual approach requires two exceptions, one for nonfinite verbs in the main clause, one for subclauses.)

HOW DO YOU FORM A RELATIVE CLAUSE (THE MAN WHO...)?

It can be useful to think about relative clauses using transformational grammar. For instance, a sentence like

The man that John hit yesterday prefers beer to wine.
can be seen as deriving by transformation from one sentence that's embedded in another:
The man [John hit him yesterday] prefers beer to wine.

In English, you can think of relativization as proceeding in two steps: a) replacing the pronoun in the subclause with an interrogative pronoun (or *that*)

The man [John hit whom yesterday] prefers beer to wine.
and b) moving that pronoun to the head of the clause.
The man [whom John hit yesterday] prefers beer to wine.

Your language may also put limits on what exactly can be relativized. The following examples are legal in English, for instance, but not in certain other languages.

the girl [you think [I love her]
>> the girl you think I love
the neighbor [I traumatized his pastor]
>> the neighbor whose pastor I traumatized
the cat [I said [Alesia brought it home]]
>> the cat that I said Alesia brought home

Not everything is possible in English:

This is the man [my girlfriend's father is a friend of John and him]
>> This is the man that my girlfriend's father is a friend of John and.
or (thanks to Leo Connolly for this example)
There's the barn [more people have gotten drunk down in back of it than any other barn in the county]
>> There's the barn that more people have gotten drunk down in back of than any other barn in the county.

Some languages can handle such sentences simply by leaving the pronoun in the subclause. S.J. Perelman liked to do this in English:

"That's the man which my wife is sleeping with him!"

If your language has cases, you must be careful to put the pronouns in the right case--English doesn't give you the right instincts here, now that *whom* is used only by pedants like me. Generally the proper case to use is the one that would be appropriate in the subclause. In *The cat that I said Alesia brought home*, for instance, the *that* representing the cat should be in the case appropriate for *the cat* in *Alesia brought the cat home*.

Quechua has an interesting way of forming clauses, using participles. For instance:

Chakra-y yapu-q runa-ta qaya-mu-saq
field-my plow-participle man-accusative call-[movement-toward]-[I-future]
I'll call the man that plowed my field.
The subclause has, rather than the form of an ordinary sentence ("the man plowed my field") the form of a participle ("the my-field-plowing man").

HOW DO YOU FORM YES-NO QUESTIONS?

English has a rather baroque procedure (inverting subject and verb). Other languages simply make use of a rise in intonation, or add a particle at the beginning of the sentence (e.g. Polish *czy*) or to the verb.

Many languages offer ways of suggesting the answer to the question. For instance, the Latin particle *num* expects the answer 'no' (*Num ursi cerevisiam imperant?* Bears don't order beer, do they?), while *nonne* expects 'yes' (*Nonne ursus animal implume bipes?* Bears are featherless bipeds, aren't they?).

Where questions are formed by appending a particle (e.g. *-ne* in Latin, or *-chu* in Quechua), the particle can be added directly to the word being questioned. We can only achieve the same effect in English by emphasis (Is the *bear* drinking beer? Is the bear drinking *beer?*) or by rearrangement (Is it beer that the bear is drinking?).

One way of asking a quesion in Chinese is to offer the listener a choice: *Nî shì bu shì Bêijing rén?* "You're from Beijing?", literally "You be, not be from Beijing?"

Some folks, believe it or not, get by without having words for 'yes' or 'no'. The usual workaround is repeat the verb from the question: "Do you know the way to San José?" can be answered "I know" or "I don't know", as in Portuguese:

--Você conhece o caminho que vai a São José?
--Conheço. ['I know']

HOW ABOUT OTHER QUESTIONS?

English usually moves the question word to the beginning of the sentence, but other languages don't, asking in effect "You said *what*?" or "She's going out with *whose* boyfriend?"

Also note that some languages have different pronouns for relative clauses ("The man who fishes") and questions ("Who is this man?").

HOW DO YOU NEGATE A SENTENCE?

Again, there are many options:

- add a particle before the verb (as in Russian or Spanish)
- ...or after the verb (as we used to do: thou rememberest not?),
- ...or both (French *je ne sais pas*)
- use a special mood of the verb (Japanese *nageru* 'throw', *nagenai* 'not throw')

- add a particle at the beginning or end of the sentence (e.g. Quechua *mana*, which however also requires a supporting suffix on the verb)
- insert a special verb and negating *that*, as English does
- use a special inflected auxiliary (e.g. Finnish *e-*)-- it's as if 'not' was an inflected verb: I not, you not, he nots...

HOW DO CONJUNCTIONS WORK?

Latin has a neat trick: to express *X and Y,* you can say *X Y-que,* using a clitic. The expression SPQR, *Senatus Populusque Romae,* is an example of this construction: the Senate and the People of Rome.

Latin also distinguishes inclusive and exclusive or: *vel X vel Y* means that you can have X or Y or both, but *aut X aut Y* means you get one or the other but not both.

Quechua (before the Spanish conquest) got by without conjunctions at all. For adding things together, you can usually get by with juxtaposition. Or you can use a case ending meaning *with*: in effect you say 'X and Y' by saying 'X with Y'. I'm not sure how disjunctions ('or') were handled-- today Quechua uses forms borrowed from Spanish.

Style

A natural language has a wide variety of **registers**, or styles of speech: from the ceremonial or ritual, to the official or scientific, to the journalistic or novelistic, to ordinary conversation, to colloquial, to slang. Children talk in their own way; so do poets. The upper crust speaks differently from the lower classes.

Some of these registers work in predictable ways. For instance, rites are often conducted in an archaic form of the language (or sometimes another language entirely). Educated speech usually includes older, longer, foreign, or technical words. In Verdurian, for instance, educated speech borrows many words from the parent language, Cad̂inor.

Slang often provides humorous substitutions for common words. Some such substitutions from Vulgar Latin have become the normal word in the Romance languages: *testa* 'pot' replaced *caput* 'head', giving French *tête; bucca* 'cheek' replaced *os* 'mouth', giving *bouche; caballus* 'nag' replaced *equus* 'horse', giving *cheval.*

Slang also borrows from minority groups: e.g. French *toubib, chnouf, bled* from Arabic; English *shiv* and *pal* from the Gypsies, *schlock* from Yiddish, *jazz* and *jive* from blacks; Spanish *calato* and *cachaco* from Quechua.

POLITENESS

All cultures have ways of expressing politeness, but they differ in the methods used, and in what ways politeness is grammaticalized.

According to Anna Wierzbicka, polite speech in English lays great stress on respecting others and avoiding imposition. English has a vast array of **indirect forms** for asking people to do things, or even for offering them things: *Will you have a drink? Would you like a drink? Sure you wouldn't like a beer? Why don't you pour yourself something? How about a beer? Aren't you thirsty?* We're so used to such pseudo-questions that we use them rather than a direct imperative even when actual politeness is far from our minds: *Will someone put this fucking idiot out of his misery? For Christ's sake, will you get lost?*

In Polish, by contrast, a courteous host pushes his hospitality on the guest, dismissing the guest's expressed remonstrances and desires as irrelevant: *Prosze bardzo! Jeszcze troszke! --Ale juz nie moge! --Ale koniecznie!* "Please, a little more!" "But I can't!" "But you must!" And Polish is very free with imperatives-- indeed, to be really forceful you must use the infinitive instead.

Japanese is often even more indirect than English: e.g. it avoids the imperative "Drink Coca-Cola!" in favor of *Koka kora o nomimashou!* (lit. "We will drink Coca-Cola!").

Japanese is also notable for having **verbal inflections** which add a level of politeness (e.g. *tetsudau* 'helps'; polite form *tetsudaimasu*), as well as entirely different lexical items with the same purpose (e.g. *iku* 'go', humble form *mairu*, honorific *irassharu*).

Terms of address are a fertile field for exquisite complications; so are **pronouns**. In quite a few languages it's perceived as rather a familiarity to address someone using the second person pronoun: to be polite you use the plural (French *vous*), or a third-person form (Italian *Lei*, Spanish *Usted* from *vuestra merced* 'your mercy', Portuguese *o senhor* 'the gentleman'), or a title (Japanese *sensei* 'teacher', *otousan* 'father', etc.). If this seems odd, it's worth noting that English took the first approach, so thoroughly that the second person singular pronoun 'thou' disappeared.

Attempts have been made to formulate **universals** of politeness, but this can be tricky. E.g. it's been suggested that politeness involves *avoiding disagreement*; but in Jewish culture disagreement expresses sociability and is taken as bringing people closer together. Or, it's been said that *direct praise* of oneself is avoided, and praise of others is approved; but self-praise among Black American speakers is good form, and direct praise of others is avoided in Japanese.

POETRY

For poetry you must consult your own Muse. However, it's worth pointing out that rhyme is not the only thing poetry can be based on:

- Old English verse was based on **alliteration**.
- Latin and Greek poetry was based on **quantity**, that is, patterns of long and short vowels.
- Blank verse, of course, is based on patterns of **stress**, without having to rhyme.
- French verse is generally based on lines of a certain **syllable length**, e.g. the alexandrine, of twelve syllables. Similarly, the haiku is composed of three lines, of 5, 7, and 5 syllables each.
- Ancient Hebrew poetry was based on **parallelism**, the near repetition of an idea ("But let justice roll down like waters, and righteousness like an ever-flowing stream."), or on successive sentences or verses each beginning with a different letter (notably Psalm 119).

Language families

You can add enormous depth to a fantasy language by giving it a history, and relatives. Verdurian and its sister language Barakhinei, for instance, derive from Ca**d̂**inor, as French and Spanish derive from Latin. Ca**d̂**inor, Cuêzi, and Xurná**š**, in turn, all derive from Proto-Eastern, and thus are related in systematic ways, much as Latin, Greek, and Sanskrit all derive from proto-Indo-European.

What can you do with such relationships?

- Create **doublets** of words to enrich the language: one that derives from the ancient language and is worn down by milennia of sound change, one that has been borrowed more recently in its ancient form. Verdurian has doublets such as these:
  fež**ir** 'hurl' **/ pegeio** 'force'
  sönil 'saddle' **/asuena** 'seat'
  žž**anec** 'coming' **/ ctanec** 'future tense'
  elut 'fair play' **/ aelutre** 'virtuous'
- Create **learned borrowings**. Legal, scientific, medical, literary, and theological terms in Verdurian are often reborrowed from Ca**d̂**inor: e.g. **vocet** 'summons'; **gutia** 'epilepsy' (from a Ca**d̂**inor word meaning 'shaking'), **menca** 'style, school'. Verdurian has also borrowed educated terms from Cuêzi: **avisar** 'school', **deyon** 'matter', **risunen** 'draw'. Moreover, some terms were borrowed direct from Cuêzi; others were borrowed from Cuêzi into Ca**d̂**inor in ancient times, and then inherited in Verdurian: e.g. **risunen** << *risunden* << Cuêzi *risonda* 'drawing', ultimately from *risi* 'reed pen'.
- Set up **borrowings from related languages**, e.g. Verdurian **kenek** 'camel', borrowed from Barakhinei *kêntek,* derived from Ca**d̂**inor *kentos* 'plain', which has also come down into Verdurian as **kent**. Or compare **či**š**te** 'guitar', borrowed from

another sister language, Ismaîn, and cognate with native **sista** 'box', both going back to Ca**d̂**inor *cista* 'box'.

Words often **change meaning** as they're borrowed. Some cute examples from Verdurian:

- **čayma** 'tent' << Western *chaimba* 'shelter'-- because the shelters of the Western barbarians were in fact tents
- **dalu** 'king' << C. *dalu* 'prince'-- because when the Ca**d̂**inorian empire fell, its princes each became independent rulers
- **garlo** 'sorcerer' << C. *garorion* 'wise or clever man'; note the **dissimilation** of the two r's; compare Latin *arbor* >> Spanish *arbol*
- **kestora** 'natural philosophy' << C. *kestora* 'the categories (of study)'
- **minyón** 'cute' << C. *mingondul* 'beggar' << *mingonda* 'large mat', i.e. all that a beggar possessed
- **nočula** 'together' << C. *nodatula* 'tied up'
- **ponyore** 'baritone' << Cuêzi *pomioro* 'manly'

HOW DO YOU DO IT?

To do this well you have to know something about historical linguistics. The sci.lang faq will give a brief overview. Better yet, read Theodora Bynon's excellent *Historical Linguistics*, or Hans Henrich Hock's more thorough *Principles of Historical Linguistics*.

The basic principle is that sound change is almost completely **regular**. This is good news: it means all you have to do is devise a set of sound changes between the parent language and its derivative(s), and apply them to each word.

Here, for instance, are just some of the **sound changes** from Ca**d̂**inor to Verdurian:

- loss of final **-os**: **corsos >> cos**
- **p** fricativizes to **f** before **s** or **t**: **psis >> fsiy**
- **c** becomes **s** before a front vowel, or before **n**: **cisir >> sisir; aracnis >> arasni**
- **g** becomes **ž**before a front vowel: **gina >> žina**
- **l** becomes **y** between vowels: **bileta >> biyeta**
- **nd, dr, lg, kr** simplify to **n, d, ly, rh** respectively: **sudrir >> sudir, unge >> unye**
- diphthongs normally simplify: **aid̂os >> ad̂, caer >> cer, Endauron >> Enäron**

A different set of sound changes can be used to create a sister language. For instance, Barakhinei changes unvoiced consonants to voiced between vowels (this is an extremely common change in languages), loses the final sound of each word, etc. The net result is a language related to but subtly different from Verdurian:

Cadhinor Verdurian Ismaîn Barakhinei gloss

prosan	prosan	prozn	proza	'walk'
molenia	mólnia	moleni	molenhi	'lightning'
ueronos	örn	rone	feron	'eagle'
aestas	esta	este	âshta	'summer'
laudan	lädan	luzn	laoda	'go'
geleia	želea	jeleze	gelech	'calm'

If you're interested in applying sound changes to one language in order to generate a descendent language, you may find my Sound Change Applier program useful.

DIALECTS

You can use the same technique to create **dialects** for a your language. Linguistically, dialects are simply a set of language varieties which haven't diverged far enough apart that their speakers can't understand each other. Dialects can be created simply by specifying a smaller number of less dramatic sound changes.

For instance, the Verdurian dialect of Avéle is characterized by the following changes:

- Unstressed vowels are reduced to **i** (front vowels), schwa (back vowels), or vocalic **r** (before r)
- Consonants between vowels become voiced: standard **epese** 'thick' becomes **ebeze**
- Where Cadînor *c* changes to **s** in standard Verdurian, in Avéle it changes to **š**
- Where Cadînor *ct* changes to **ž**in standard Verdurian, in Avéle it also changes to **š**

Dialects can also have their own lexical terms, of course, perhaps borrowed from neighbors or previous inhabitants of the local territory.

People often suppose that the dialect of the capital city (or whatever other place has supplied the standard language) is more 'pure' or more conservative than provincial speech. In fact the opposite is likely to be true: the active center of a culture will see its speech change fastest; rural or isolated areas are more likely to preserve older forms.

If you're inventing an interlanguage you may of course want to do everything possible to *prevent* the rise of dialects. This is probably an expression of the fascistic streak common to language tinkerers. Why not *design* your interlanguage with dialects, reflecting the phonology of various linguistic regions? The resulting language, with varieties close to the major natural languages, might achieve more acceptance than uniform interlanguages have.

What is Writing? - http://www.omniglot.com/writing/index.htm

What is writing?

There are a number of different ways to describe writing and writing systems.

In *the world's writing systems*, Peter T. Daniels defines writing as:

a system of more or less permanent marks used to represent an utterance in such a way that it can be recovered more or less exactly without the intervention of the utterer.

In *The Blackwell Encyclopedia of Writings Systems*, Florian Coulmas defines a writing system as:

a set of visible or tactile signs used to represent units of language in a systematic way, with the purpose of recording messages which can be retrieved by everyone who knows the language in question and the rules by virtue of which its units are encoded in the writing system.

All writing systems use visible signs with the exception of the raised notation systems used by blind and visually impaired people, such as Braille and Moon. Hence the need to include tactile signs in the above definition.

In *A History of Writing*, Steven Roger Fischer argues that no one definition of writing can cover all the writing systems that exist and have ever existed. Instead he states that a 'complete writing' system should fullfill all the following criteria:

- Complete writing must have as its purpose communication;
- Complete writing must consist of artificial graphic marks on a durable or electronic surface;
- Complete writing must use marks that relate conventionally to articulate speech (the systematic arrangement of significant vocal sounds) or electronic programing in such a way that communication is achieved.

Types of writing system

- Abjads / Consonant Alphabets

  Abjads, or consonant alphabets, represent consonants only, or consonants plus some vowels. Full vowel indication (vocalisation) can be added, usually by means of diacritics, but this is not common. Most of abjads,

with the exception of Divehi hakura and Ugaritic, are written from right to left.

Some scripts, such as Arabic, are used both as an abjad and as an alphabet.

- Alphabets

  Alphabets, or phonemic alphabets, represent consonants and vowels.

- Syllabic Alphabets / Abugidas

  Syllabic alphabets, alphasyllabaries or abugidas consist of symbols for consonants and vowels. The consonants each have an inherent vowel which can be changed to another vowel or muted by means of diacritics. Vowels can also be written with separate letters when they occur at the beginning of a word or on their own.

  When two or more consonants occur together, special conjunct symbols are often used which add the essential parts of first letter or letters in the sequence to the final letter.

- Syllbaries

  A syllabary is a phonetic writing system consisting of symbols representing syllables. A syllable is often made up of a consonant plus a vowel or a single vowel. In Japanese, for example, you use different symbols to write ka, ki, ku, ke or ko (か、き、く、け、こ).

- Logographic writing systems (Chinese, Hieroglyphs, etc.)

  The symbols used in these complex scripts may represent both sound and meaning. As a result, these scripts generally include a large number of symbols: anything from several hundred to tens of thousands. In fact there is no theoretical upper limit to the number of symbols in some scripts, such as Chinese.

  Complex scripts may include the following types of symbol:

  - Logograms - symbols which represent parts of words or whole words. Some logograms resemble the things they represent and are sometimes known as pictograms or pictographs.
  - Ideograms - symbols which graphically represent abstract ideas.
  - Semantic-phonetic compounds - symbols which include a semantic element, which represents or hints at the meaning of the symbol, and a phonetic element, which denotes or hints at the pronunciation.

        - Sometimes symbols are used for their phonetic value alone, without regard for their meaning.

- Alternative writing systems (fictional and constructed alphabets, and other communication systems)
- Undeciphered writing systems

Numerals in many different writing systems

	0	1	2	3	4	5	6	7	8	9	10	100	1000	10000
Arabic	٠	١	٢	٣	٤	٥	٦	٧	٨	٩				
Bengali	০	১	২	৩	৪	৫	৬	৭	৮	৯				
Chinese (simple numerals)	〇	一	二	三	四	五	六	七	八	九	十	百	千	万
Chinese (complex numerals)	零	壹	貳	參	肆	伍	陸	柒	捌	玖	拾	佰	仟	萬
Chinese 花碼 (huā mǎ)	〇	𠄡	𠄢	𠄣	𠄤	𠄥	𠄦	𠄧	𠄨	𠄩				
Devanagari	०	१	२	३	४	५	६	७	८	९				
Ethiopic		፩	፪	፫	፬	፭	፮	፯	፰	፱	፲	፻		፼
Gujarati	૦	૧	૨	૩	૪	૫	૬	૭	૮	૯				
Gurmukhi	੦	੧	੨	੩	੪	੫	੬	੭	੮	੯				
Kannada	೦	೧	೨	೩	೪	೫	೬	೭	೮	೯				
Khmer	០	១	២	៣	៤	៥	៦	៧	៨	៩				
Lao	๐	๑	๒	๓	๔	๕	๖	๗	๘	๙				
Limbu	᥆	᥇	᥈	᥉	᥊	᥋	᥌	᥍	᥎	᥏				
Malayalam	൦	൧	൨	൩	൪	൫	൬	൭	൮	൯				
Mongolian	᠐	᠑	᠒	᠓	᠔	᠕	᠖	᠗	᠘	᠙				
Myanmar	๐	๑	๒	๓	๔	๕	๖	๗	๘	๙				
Oriya	୦	୧	୨	୩	୪	୫	୬	୭	୮	୯				
Tamil		க	உ	ரூ	சு	ரூ	கூ	எ	அ	க	ப	ஊ	சூ	
Telugu	౦	౧	౨	౩	౪	౫	౬	౭	౮	౯				
Thai	๐	๑	๒	๓	๔	๕	๖	๗	๘	๙				
Tibetan	๐	๑	๒	๓	๔	๕	๖	๗	๘	๙				
Urdu	٠	١	٢	٣	۴	۵	۶	٧	٨	٩				

Arabic script

Origin

The Arabic script evolved from the Nabataean Aramaic script. It has been used since the 4th century AD, but the earliest document, an inscription in Arabic, Syriac and Greek, dates from 512 AD. The Aramaic language has fewer consonants than Arabic, so during the 7th century new Arabic letters were created by adding dots to existing letters in order to avoid ambiguities. Further diacritics indicating short vowels were introduced, but are only generally used to ensure the Qur'an was read aloud without mistakes.

There are two main types of written Arabic:

1. **Classical Arabic** - the language of the Qur'an and classical literature. It differs from Modern Standard Arabic mainly in style and vocabulary, some of which is archaic. All Muslims are expected to recite the Qur'an in the original language, however many rely on translations in order to understand the text.
2. **Modern Standard Arabic** - the universal language of the Arabic-speaking world which is understood by all Arabic speakers. It is the language of the vast majority of written material and of formal TV shows, lectures, etc.

Each Arabic speaking country or region also has its own variety of colloquial spoken Arabic. These colloquial varieties of Arabic appear in written form in some poetry, cartoons and comics, plays and personal letters. There are also translations of the bible into most varieties of colloquial Arabic.

Arabic has also been written with the Hebrew, Syriac and Latin scripts.

Notable Features

- The Arabic alphabet contains 28 letters. Some additional letters are used in Arabic when writing placenames or foreign words containing sounds which do not occur in Standard Arabic, such as /p/ or /g/.
- Words are written in horizontal lines from right to left, numerals are written from left to right
- Most letters change form depending on whether they appear at the beginning, middle or end of a word, or on their own. (see below)
- Letters that can be joined are always joined in both hand-written and printed Arabic. The only exceptions to this rule are crossword puzzles and signs in which the script is written vertically.
- The long vowels /a:/, /i:/ and /u:/ are represented by the letters *'alif*, *yā'* and *wāw* respectively.
- Vowel diacritics, which are used to mark short vowels, and other special symbols apppear only in the Qur'ān (Koran). They are also used, though with less consistancy, in other religious texts, in classical poetry, in textbooks children and foreign learners, and occasionally in complex texts to avoid ambiguity.

Sometimes the diacritics are used for decorative purposes in book titles, letterheads, nameplates, etc.

Arabic consonants

IPA	Latin	Name		Final	Medial	Initial	Isolated	IPA	Latin	Name		Final	Medial	Initial	Isolated
[ʈ]	ṭ	ṭā’	طاء	ـط	ـطـ	طـ	ط	[ʔ]	’(a)	alif	ألف	ـا	__	__	ا
[ʐ]	ẓ	ẓā’	ظاء	ـظ	ـظـ	ظـ	ظ	[b]	b	bā’	باء	ـب	ـبـ	بـ	ب
[ʕ]	‘	‘ayn	عين	ـع	ـعـ	عـ	ع	[t]	t	tā’	تاء	ـت	ـتـ	تـ	ت
[ɣ]	gh	ghayn	غين	ـغ	ـغـ	غـ	غ	[θ]	th	thā’	ثاء	ـث	ـثـ	ثـ	ث
[f]	f	fā’	فاء	ـف	ـفـ	فـ	ف	[ʤ]	j	jīm	جيم	ـج	ـجـ	جـ	ج
[q]	q	qāf	قاف	ـق	ـقـ	قـ	ق	[ħ]	ḥ	ḥā’	حاء	ـح	ـحـ	حـ	ح
[k]	k	kāf	كاف	ـك	ـكـ	كـ	ك	[x]	kh	khā’	خاء	ـخ	ـخـ	خـ	خ
[l]	l	lām	لام	ـل	ـلـ	لـ	ل	[d]	d	dāl	دال	ـد	__	__	د
[m]	m	mīm	ميم	ـم	ـمـ	مـ	م	[ð]	dh	dhāl	ذال	ـذ	__	__	ذ
[n]	n	nūn	نون	ـن	ـنـ	نـ	ن	[r]	r	rā’	راء	ـر	__	__	ر
[h]	h	hā’	هاء	ـه	ـهـ	هـ	ه	[z]	z	zāy	زاي	ـز	__	__	ز
[w]	w	wāw	واو	ـو	__	__	و	[s]	s	sīn	سين	ـس	ـسـ	سـ	س
[j]	y	yā’	ياء	ـي	ـيـ	يـ	ي	[ʃ]	š	shīn	شين	ـش	ـشـ	شـ	ش
		hamza	همزة	ء	__	__	__	[sˤ]	ṣ	ṣād	صاد	ـص	ـصـ	صـ	ص
								[dˤ]	ḍ	ḍād	ضاد	ـض	ـضـ	ضـ	ض

Arabic vowel diacritics and other symbols

لا بُّ بِّ بَّ بّ بْ بو بي بِ بَا بُ بِ بَ

lām ’alif				šadda	sūkun				ḍamma	kasra	fatḥa
lā	bbu	bbi	bba	bb	b	bū	bī	bā	bu	bi	ba

Arabic numerals and numbers

٠	١	٢	٣	٤	٥	٦	٧	٨	٩	١٠
0	1	2	3	4	5	6	7	8	9	10
صفر	واحد	إثنان	ثلاثة	أربعة	خمسة	ستة	سبعة	ثمانية	تسعة	عشرة
ṣifr	waḥid	ithnān	thalātha	ʿarbaʿa	khamsa	sītta	sabʿa	thamānya	tisʿa	ʿashara
صفر	واحد	جوج	تلاتة	ربعة	خمسة	ستة	سبعة	تمنية	تسعود	عشرة
ṣifr	wahed	zhuzh	tlata	rebʿa	khamsa	setta	sebʿa	tmenya	tesʿūd	ʿashra

The first lot of number names are Modern Standard Arabic. The second lot are Moroccan Arabic.

The Arabic language

Arabic is a Semitic language with about 221 million speakers in Afghanistan, Algeria, Bahrain, Chad, Cyprus, Djibouti, Egypt, Eritrea, Iran, Iraq, Israel, Jordan, Kenya, Kuwait, Lebannon, Libya, Mali, Mauritania, Morocco, Niger, Oman, Palestinian West Bank & Gaza, Qatar, Saudi Arabia, Somalia, Sudan, Syria, Tajikistan, Tanzania, Tunisia, Turkey, UAE, Uzbekistan and Yemen.

There are over 30 different varieties of colloquial Arabic which include:

- **Egyptian** - spoken by about 46 million people in Egypt and perhaps the most widely understood variety, thanks to the popularity of Egyptian-made films and TV shows
- **Algerian** - spoken by about 22 million people in Algeria
- **Moroccan/Maghrebi** - spoken in Morocco by about 19.5 million people
- **Sudanese** - spoken in Sudan by about 19 million people
- **Saidi** - spoken by about 19 million people in Egpyt
- **North Levantine** - spoken in Lebannon and Syria by about 15 million people
- **Mesopotamian** - spoken by about 14 million people in Iraq, Iran and Syria
- **Najdi** - spoken in Saudi Arabia, Iraq, Jordan and Syria by about 10 million people

For a full list of all varieties of colloquial Arabic click here (format: Excel, 20K).

Source: www.ethnologue.com

Sample Arabic text

يولد جميع الناس أحراراً متساوين في الكرامة والحقوق. وقد وهبوا
عقلاً وضميراً وعليهم ان يعامل بعضهم بعضا بروح الإخاء.

Translation

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood. *(Article 1 of the Universal Declaration of Human Rights)*

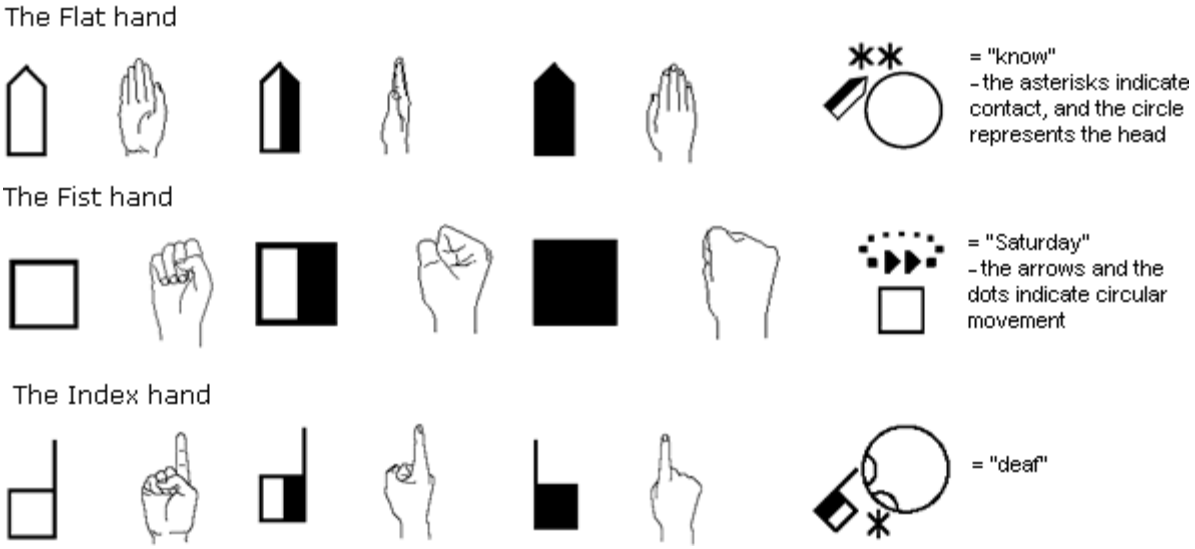
Sutton SignWriting

Sutton SignWriting, or SignWriting, was created in 1974 by Valerie Sutton. It uses visual symbols to represent the handshapes, movements, and facial expressions of signed languages. SignWriting is based on Sutton DanceWriting, a notation system for representing dance movements which Valerie Sutton developed in 1972.

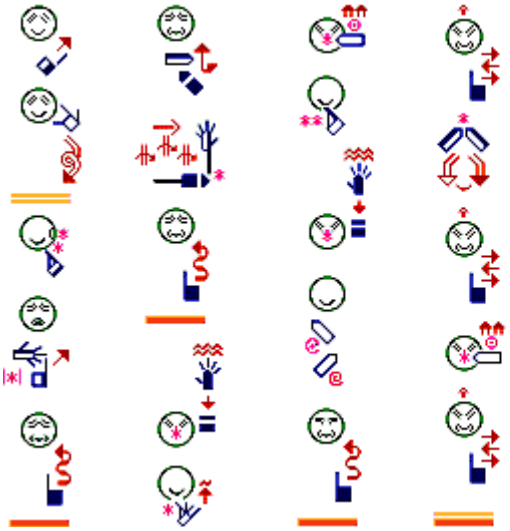
SignWriting is a "movement-writing-alphabet", which can be used to write any signed language. It is the written form of 27 Sign Languages. The SignWriting alphabet writes the way the body looks, when people sign, just as the Roman alphabet writes the way words sound, when people speak.

SignWriting can be used to write American Sign Language (ASL), British Sign Language (BSL) or any other variety of sign language. There are newspapers, magazines, dictionaries, and literature written in SignWriting. It is also used to teach signs and signed language grammar to novice signers, and can be used to teach skilled signers other subjects, such as maths, history or English.

A selection of basic ASL SignWriting signs



Sample text in ASL SignWriting (from Goldilocks and the Three Bears)



Gloss

THERE GOLDILOCKS , HOME ESCAPE
WANDER . ENTER FOREST WANDER .
SMELL FAVORITE SCENT
FOOD SMELL ENJOY WANDER .
WHERE HOUSE WHERE SCENT WHERE ?

English version

Goldilocks wandered away from her home and into the forest. She smelled the scent of her favorite food and wandered towards the pleasing scent. Where was the house where the scent was coming from?

Gloss and English version provided by Marq Thompson

Korean 한국어

Origin of writing in Korea

Chinese writing has been known in Korea for over 2,000 years. It was used widely during the Chinese occupation of northern Korea from 108 BC to 313 AD. By the 5th century AD, the Koreans were starting to write in Classical Chinese - the earliest known example of this dates from 414 AD. They later devised three different systems for writing Korean with Chinese characters: *Hyangchal* (향찰/鄕札), *Gukyeol* (구결/口訣) and *Idu* (이두/吏讀). These systems were similar to those developed in Japan and were probably used as models by the Japanese.

The *Idu* system used a combination of Chinese characters together with special symbols to indicate Korean verb endings and other grammatical markers, and was used to in official and private documents for many centuries. The *Hyangchal* system used Chinese characters to represent all the sounds of Korean and was used mainly to write poetry.

The Koreans borrowed a huge number of Chinese words, gave Korean readings and/or meanings to some of the Chinese characters and also invented about 150 new characters, most of which are rare or used mainly for personal or place names.

The Korean alphabet was invented in 1444 and promulgated it in 1446 during the reign of King Sejong (r.1418-1450), the fourth king of the Joseon Dynasty. The alphabet was originally called *Hunmin jeongeum*, or "The correct sounds for the instruction of the people", but has also been known as *Eonmeun* (vulgar script) and *Gukmeun* (national writing). The modern name for the alphabet, *Hangeul*, was coined by a Korean linguist called Ju Si-gyeong (1876-1914).

King Sejong and his scholars probably based some of the letter shapes of the Korean alphabet on other scripts such as Mongolian and 'Phags Pa, and the traditional direction of writing (vertically from right to left) most likely came from Chinese, as did the practice of writing syllables in blocks.

Even after the invention of the Korean alphabet, most Koreans who could write continued to write either in Classical Chinese or in Korean using the *Gukyeol* or *Idu* systems. The Korean alphabet was associated with people of low status, i.e. women, children and the uneducated. During the 19th and 20th centuries a mixed writing system combining Chinese characters (*Hanja*) and *Hangeul* became increasingly popular. Since 1945 however, the importance of Chinese characters in Korean writing has diminished significantly.

Since 1949 *hanja* have not been used at all in any North Korean publications, with the exception of a few textbooks and specialized books. In the late 1960s the teaching of *hanja* was reintroduced in North Korean schools however and school children are expected to learn 2,000 characters by the end of high school.

In South Korea school children are expected to learn 1,800 *hanja* by the end of high school. The proportion of *hanja* used in Korean texts varies greatly from writer to writer and there is considerable public debate about the role of *hanja* in Korean writing.

Most modern Korean literature and informal writing is written entirely in *hangeul*, however academic papers and official documents tend to be written in a mixture of *hangeul* and *hanja*.

Notable features of Hangeul

- There are 24 letters (*jamo*) in the Korean alphabet: 14 consonants and 10 vowels. The letters are combined together into syllable blocks.

For example, *Hangeul* is written: 한 (han) ㅎ(h) + ㅏ(a) + ㄴ(n) 글 (geul) ㄱ(g) + ㅡ(eu) + ㄹ(l)

- The shapes of the the consontants g/k, n, s, m and ng are graphical representations of the speech organs used to pronounce them. Other consonsants were created by adding extra lines to the basic shapes.
- The shapes of the the vowels are based on three elements: man (a vertical line), earth (a horizontal line) and heaven (a dot). In modern Hangeul the heavenly dot has mutated into a short line.
- Spaces are placed between words, which can be made up of one or more syllables.
- The sounds of some consonants change depending on whether they appear at the beginning, in the middle, or at the end of a syllable.
- A number of Korean scholars have proposed an alternative method of writing *Hangeul* involving writing each letter in a line like in English, rather than grouping them into syllable blocks, but their efforts have been met with little interest or enthusiasm.
- In South Korea *hanja* are used to some extent in Korean texts.
- Korean can be written in vertical columns running from top to bottom and right to left, or in horizontal lines running from left to right.

Used to write

Korean, a language spoken by about 63 million people in South Korea, North Korea, China, Japan, Uzbekistan, Kazakhstan and Russia. The relationship between Korean and other languages is not known, though some linguists believe it to be a member of the Altaic family of languages. Grammatically Korean is very similar to Japanese and about half its vocabulary comes from Chinese.

The Hangeul alphabet (한글)

Consonants (자음/子音)

ㄱ	ㄲ	ㄴ	ㄷ	ㄸ	ㄹ	ㅁ	ㅂ	ㅃ
기역	쌍 기역	니은	디귿	쌍 디귿	리을	미음	비읍	쌍 비읍
giyeok	ssang giyeok	nieun	digeut	ssang digeut	rieul	mieum	bieup	ssang bieup
g, k	kk	n	d, t	tt	l	m	b, p	pp
k, g	kk	n	t, d	tt	l, r	m	p, b	pp
[k/g]	[k']	[n]	[t/d]	[t']	[l/r]	[m]	[p/b]	[p']

ㅅ	ㅆ	ㅇ	ㅈ	ㅉ	ㅊ	ㅋ	ㅌ	ㅍ	ㅎ
시옷	쌍 시옷	이응	지읒	쌍 지읒	치읓	키읔	티읕	피읖	히읗
shiot	ssang shiot	ieung	jieut	ssang jieut	chieut	kiuek	tieut	pieup	hieut
s	ss	ng	j	jj	ch	k	t	p	h
s	ss	-ng	ch, j	tch	ch'	k'	t'	p'	h
[s]	[s']	[Ø/-ŋ]	[ʧ/ʤ]	[ʧ']	[ʧʰ]	[kʰ]	[tʰ]	[pʰ]	[h]

Vowels (모음/母音)

ㅏ	ㅐ	ㅑ	ㅒ	ㅓ	ㅔ	ㅕ	ㅖ	ㅗ	ㅘ	ㅙ
a	ae	ya	yae	eo	e	yeo	ye	o	wa	wae
a	ae	ya	yae	ŏ	e	yŏ	ye	o	wa	wae
[a]	[æ]	[ja]	[jæ]	[ʌ]	[e]	[jʌ]	[je]	[o]	[wa]	[wæ]

ㅚ	ㅛ	ㅜ	ㅝ	ㅞ	ㅟ	ㅠ	ㅡ	ㅢ	ㅣ
oe	yo	u	wo	we	wi	yu	eu	ui	i
oe	yo	u	wŏ	we	wi	yu	ŭ	ŭi	i
[we]	[jo]	[u]	[wʌ]	[we]	[wi]	[ju]	[ɨ]	[ɨj]	[i]

Note on the transliteration of Korean

There are a number different ways to write Korean in the Latin alphabet. The methods shown above are:

1. (first row) the official South Korean transliteration system, which was introduced in July 2000. You can find further details at www.mct.go.kr.
2. (second row) the McCune-Reischauer system, which was devised in 1937 by two American graduate students, George McCune and Edwin Reischauer, and is widely used in Western publications. For more details of this system see: http://mccune-reischauer.org

Sample of in Korean

모든 인간은 태어날 때부 터 자유로우며 그 존엄과 권리에 있어 동등하다. 인간은 천부적 으로 이성과 양심을 부여받았으며 서로 형제애의 정신으로 행동하여야한다.

Translation

All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood. *(Article 1 of the Universal Declaration of Human Rights)*

Mongolian alphabets (Монгол)

Origin

The Mongolian alphabet was adapted from the Uighur alphabet in the 12th Century. The Uighur alphabet was a derivative of the Sogdian alphabet, which ultimately came from Aramaic.

Between the 13th and 15th Centuries, Mongolian was also written with Chinese characters, the Arabic alphabet and a script derived from Tibetan called Phags-pa.

As a result of pressure from the Soviet Union, Mongolia adopted the Latin alphabet in 1931 and the Cyrillic alphabet in 1937. In 1941 the Mongolian government passed a law to abolish the Mongolian alphabet.

Since 1994, the Mongolian government has been trying to bring back the Mongolian alphabet and it is starting to be used more widely and is now taught in schools.

In Inner Mongolian Autonomous Region of China the traditonal Mongolian alphabet is still used.

Notable features

- This is a phonemic alphabet with separate letters for consonants and vowels.
- Written vertically from top to bottom and from left to right. This is very unusual as all other scripts that are written vertically (Chinese, Japanese and Korean) are written from right to left
- The letters have a number of different shapes, the choice of which depends on the position of a letter in a word and which letter follows it.

Used to write

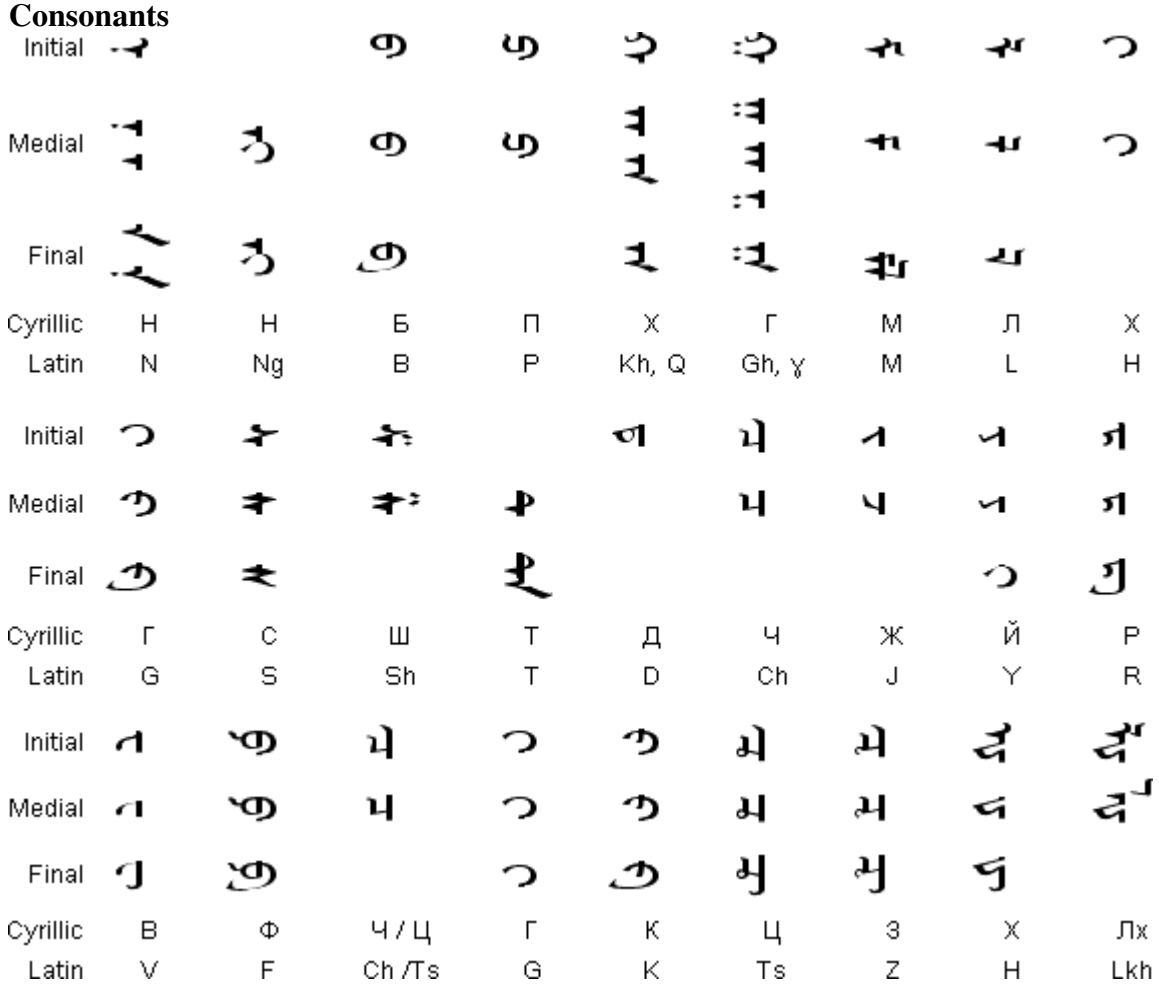
Mongolian, an Altaic language spoken by approximately 5 million people in Mongolia, China, Afghanistan and Russia. There are a number of closely related varieties of Mongolian: **Khalkha** or **Halha**, the national language of Mongolia, and **Oirat**, **Chahar** and **Ordos**, which are spoken mainly in the Inner Mongolian Autonomous Region of China.

Other languages considered part of the Mongolian language family, but separate from Mongolian, include **Buryat** and **Kalmyk**, spoken in Russia and **Moghul** or **Mogul**, spoken in Afghanistan.

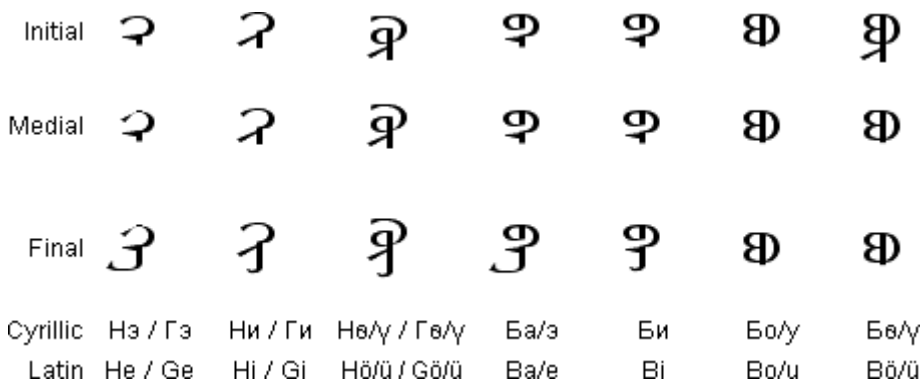
Traditional Mongolian alphabet

Vowels

Consonants

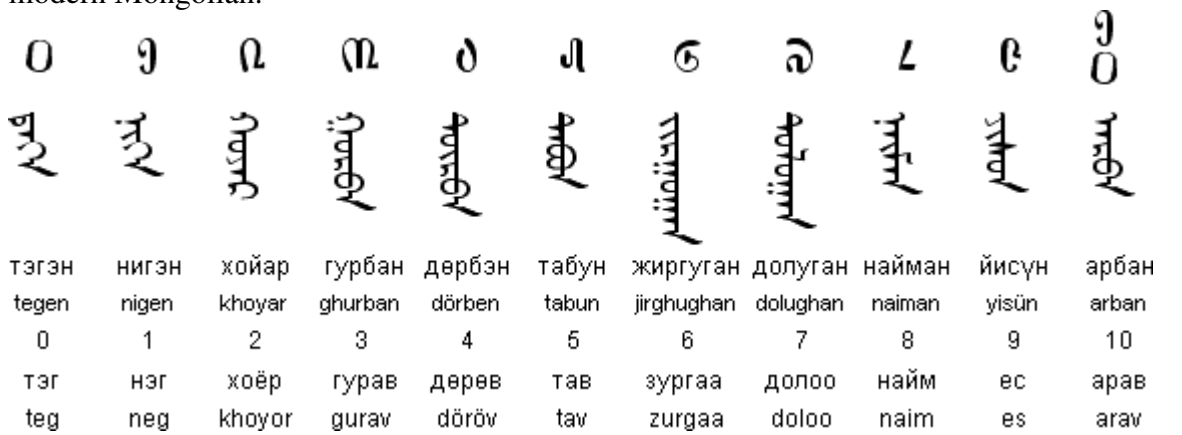


Consonant/vowel combinations

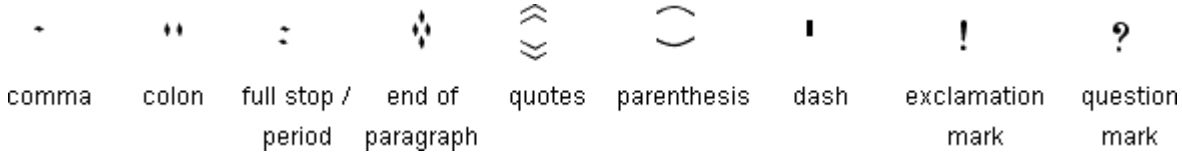


Numerals

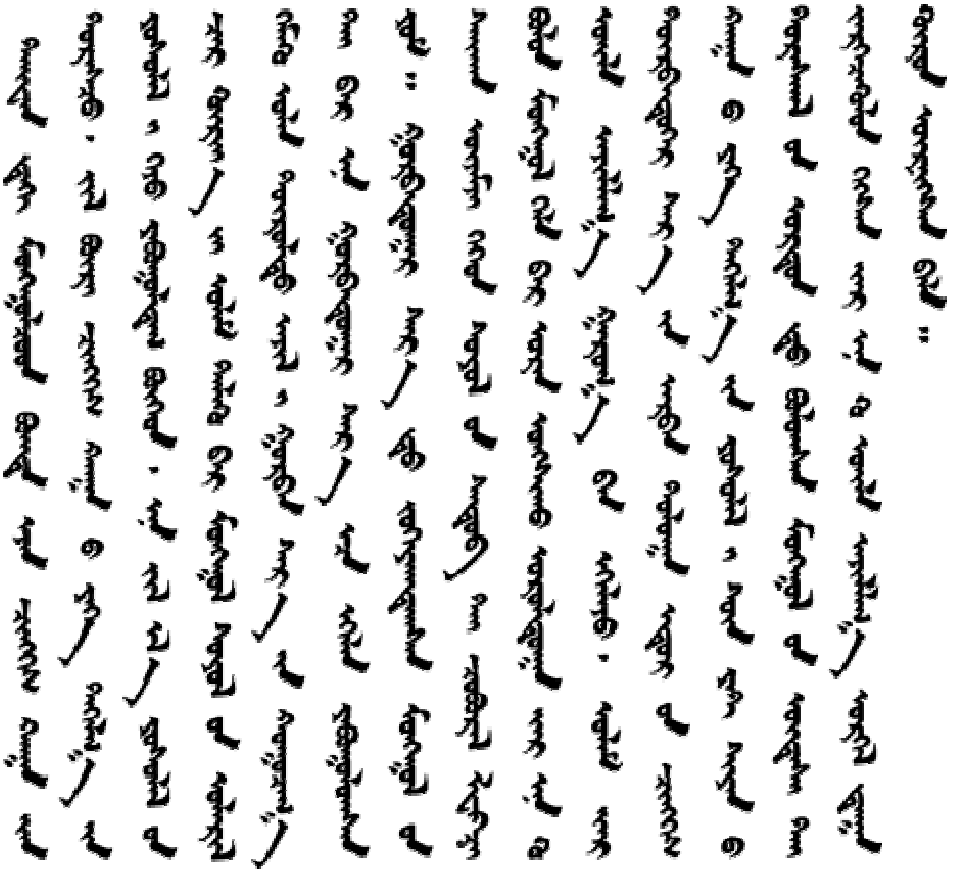
The first set of numbers (tegen, nigen, etc.) are Classical Mongolian, the others are modern Mongolian.



Punctuation



Sample of Mongolian written in the traditional alphabet



Cyrillic alphabet for Mongolian (Khalkha)

А а	Б б	В в	Г г	Д д	Е е	Ё ё	Ж ж	З з	И и	Й й	К к
a	b	w	g	d	ye/yö	yo	j	j (dz)	i	i	k
[ɒ]	[b]	[v, w]	[g, k]	[d]	[jɛ, jɜ]	[jɔ]	[ʤ]	[z, dz]	[i]	[i]	[k]
Л л	М м	Н н	О о	Ө ө	П п	Р р	С с	Т т	У у	Ү ү	Ф ф
l	m	n-, -ng	o	ö	p	r	s	t	u	ü	f
[l]	[m]	[n, ŋ]	[ɔ]	[ɜ:]	[p]	[r]	[s]	[tʰ]	[ʊ]	[u:]	[f]
Х х	Ц ц	Ч ч	Ш ш	Щ щ	Ъ ъ	Ы ы	Ь ь	Э э	Ю ю	Я я	
kh	c	č	š	šč	hard	ii	soft	e	yu/yü	ya	
[x]	[c]	[ʧ]	[ʃ]	[ʃʧ]	sign	[i:]	sign	[ɛ]	[ju]	[ja]	

Sample Mongolian text in the Cyrillic alphabet

Хүн бүр төрж мэндлэхдд зрх ёөлөөтэй, адилхан нэр төртэй, ижил эрхтэй байдаг. Оюун ухаан, нандин ёанар заяасан хүн гэгё өөр хоорондоо ахан дүүгийн үзэл санаагаар харьцах уёиртай.

Transliteration

Khün bür törzh mendlekhdd erkh ëolööteĭ, adilkhan ner törteĭ, izhil erkhteĭ baĭdag. Oyuun ukhaan, nandin ëanar zayaasan khün gegë öör khoorondoo akhan düügiĭn üzel sanaagaar khar'tsakh uëirtaĭ.

Translation
All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.
(Article 1 of the Universal Declaration of Human Rights)

Devanāgarī alphabet देवनागरी

Origin

The Nāgarī (lit. 'of the city') or Devanāgarī ('divine Nagari') alphabet descended from the Brahmi script sometime around the 11th century AD. It was originally developed to write Sanksrit but was later adapted to write many other languages.

Notable Features

- Some scholars use the term alphasyllabary to describe Devanāgarī, while others call it an abugida.
- Consonant letters carry an inherent vowel which can be altered or muted by means of diacritics or *matra*.
- Vowels can be written as independent letters, or by using a variety of diacritical marks which are written above, below, before or after the consonant they belong to. This feature is common to most of the alphabets of South and South East Asia.
- When consonants occur together in clusters, special conjunct letters are used.
- The order of the letters is based on articulatory phonetics.

Used to write:

Awadhi, Bagheli, Balti, Bateri, Bhili, Bhojpuri, Bihari, Braj bhasha, Chhattisgarhi, Garhwali, Gondi, Harauti, Hindi, Ho, Kachchi, Kanauji, Kankan, Kashmiri, Konkan, Limbu, Marwari, Marathi, Nepali, Newari, Sanskrit, Santali, Sherpa, Sindhi

Devanāgarī alphabet

Primary vowels

	Short				Long				Diphthongs			
	Initial		Diacritic		Initial		Diacritic		Initial		Diacritic	
Unrounded low central	अ	a	प	pa	आ	ā	पा	pā				
Unrounded high front	इ	i	पि	pi	ई	ī	पी	pī				
Rounded high back	उ	u	पु	pu	ऊ	ū	पू	pū				
Syllabic variant	ऋ	ṛ	पृ	pṛ	ॠ	ṝ	पॄ	pṝ				
Secondary vowels												
Unrounded front					ए	e	पे	pe	ऐ	ai	पै	pai
Rounded back					ओ	o	पो	po	औ	au	पौ	pau

Other symbols

अं aṅ *anusvāra* - nasalises vowel अँ aṃ *anunāsika/candrabindu* - nasalises vowel

अः aḥ *visarga* - adds voiceless breath after vowel प् p *virāma* - mutes vowel

Consonants

Occlusives

	Voiceless plosives				Voiced plosives				Nasals	
	unaspirated		aspirated		unaspirated		aspirated			
Velar	क	ka	ख	kha	ग	ga	घ	gha	ङ	ṅa
Palatal	च	ca	छ	cha	ज	ja	झ	jha	ञ	ña
Retroflex	ट	ṭa	ठ	ṭha	ड	ḍa	ढ	ḍha	ण	ṇa
Dental	त	ta	थ	tha	द	da	ध	dha	न	na
Labial	प	pa	फ	pha	ब	ba	भ	bha	म	ma

Sonorants and fricatives

	Palatal		Retroflex		Dental		Labial	
Sonorants	य	ya	र	ra	ल	la	व	va
Sibilants	श	śa	ष	ṣa	स	sa		

Other letters

ह ha ळ ḷa

Variant letters used in Mumbai

झ jha ण ṇa

A selection of conjunct consonants

क्ष kṣa ज्ञ jña त्त tta त्र tra प्य pya त्क tka ट्क ṭka ह्य hya त्व ttva

Numerals

०	१	२	३	४	५	६	७	८	९	१०
0	1	2	3	4	5	6	7	8	9	10

ひらがな

Japanese Hiragana

Origin

Hiragana syllables developed from Chinese characters, as shown below. Hiragana were originally called *onnade* or 'women's hand' as were used mainly by women - men wrote in kanji and katakana. By the 10th century, hiragana were used by everybody. The word hiragana means "oridinary syllabic script".

In early versions of hiragana there were often many different characters to represent the same syllable, however the system was eventually simplified so that there was a one-to-one relationship between spoken and written syllables. The present orthography of hiragana was codified by the Japanese government in 1946.

The hiragana syllabary

In each column the rōmaji appears on the left, the hiragana symbols in the middle and the kanji from which they developed on the right. There is some dispute about which kanji the hiragana developed from.

平仮名（ひらがな） hiragana

a	あ	安	i	い	以	u	う	宇	e	え	衣	o	お	於
ka	か	加	ki	き	幾	ku	く	久	ke	け	計	ko	こ	己
sa	さ	左	shi	し	之	su	す	寸	se	せ	世	so	そ	曽
ta	た	太	chi	ち	知	tsu	つ	川	te	て	天	to	と	止
na	な	奈	ni	に	仁	nu	ぬ	奴	ne	ね	祢	no	の	乃
ha	は	波	hi	ひ	比	fu	ふ	不	he	へ	部	ho	ほ	保
ma	ま	末	mi	み	美	mu	む	武	me	め	女	mo	も	毛
ya	や	也				yu	ゆ	由				yo	よ	与
ra	ら	良	ri	り	利	ru	る	留	re	れ	礼	ro	ろ	呂
wa	わ	和	wi	ゐ	為				we	ゑ	恵	wo	を	遠
												n	ん	无

The symbols for 'wi' and 'we' were made obsolete by the Japanese Minsitry of Education in 1946 as part of its language reforms. The symbols 'ha', 'he' and 'wo' are pronounced 'wa', 'e' and 'o' respecitvely when used as grammatical particles.

Additional sounds are represented using diacritics or combinations of syllables:

ga	が	gi	ぎ	gu	ぐ	ge	げ	go	ご	kya	きゃ	kyu	きゅ	kyo	きょ
za	ざ	ji	じ	zu	ず	ze	ぜ	zo	ぞ	gya	ぎゃ	gyu	ぎゅ	gyo	ぎょ
da	だ	ji	ぢ	zu	づ	de	で	do	ど	sha	しゃ	shu	しゅ	sho	しょ
ba	ば	bi	び	bu	ぶ	be	べ	bo	ぼ	ja	じゃ	ju	じゅ	jo	じょ
pa	ぱ	pi	ぴ	pu	ぷ	pe	ぺ	po	ぽ	cha	ちゃ	chu	ちゅ	cho	ちょ
										nya	にゃ	nyu	にゅ	nyo	にょ
										hya	ひゃ	hyu	ひゅ	hyo	ひょ
										bya	びゃ	byu	びゅ	byo	びょ
										pya	ぴゃ	pyu	ぴゅ	pyo	ぴょ
										mya	みゃ	myu	みゅ	myo	みょ
										rya	りゃ	ryu	りゅ	ryo	りょ

Characteristics and usage of hiragana

The hiragana syllabary consists of 48 syllables and is mainly used to write word endings, known as *okurigana* in Japanese. Hiragana are also widely used in materials for children, textbooks, animation and comic books, to write Japanese words which are not normally written with kanji, such as adverbs and some nouns and adjectives, or for words whose kanji are obscure or obselete.

Hiragana are also sometimes written above or along side kanji to indicate pronunciation, especially if the pronunication is obscure or non-standard. Hiragana used in this way are known as *furigana* or ruby. In horizontal texts, the furigana appear above the kanji and in vertical texts, the furigana appear on the right of the kanji. In newspapers it is a legal requirement for furigana to be attached to kanji which are not included in the official list of the 1,945 most frequently-used kanji. Newspapers in fact rarely use kanji not included in this list.

Furigana in action

The furigana in the following text are the small red symbols.

Horizontal text

すべての人間は、生まれながらにして自由であり、かつ、尊厳と権利とについて平等である。人間は、理性と良心、とを授けられてあり、互いに同胞の精神をもって行動しなければならない。

Vertical text

すべての人間は、生まれながらにして自由であり、かつ、尊厳と権利とについて平等である。人間は、理性と良心、とを授けられてあり、互いに同胞の精神をもって行動しなければならない。

Hiragana are sometimes used to write words which would normally written with katakana to make them appear more "feminine", particularly in comic books and cartoons for young girls. In children's video games texts are often written entirely in hiragana or katakana.

Japanese Katakana

Origin

The katakana syllabary was derived from abbreviated Chinese characters used by Buddhist monks to indicate the correct pronunciations of Chinese texts in the 9th century. At first there were many different symbols to represent one syllable of spoken Japanese, but over the years the system was streamlined. By the 14th century, there was a more or less one-to-one correspondence between spoken and written syllables.

The word katakana "part (of kanji) syllabic script". The "part" refers to the fact that katakana characters represent parts of kanji.

Characteristics and usage of katakana

The katakana syllabary consists of 48 syllables and was originally considered "men's writing". Since the 20th century, katakana have been used mainly to write non-Chinese loan words, onomatopoeic words, foreign names, in telegrams and for emphasis (the equivalent of bold, italic or upper case text in English). Before the 20th century all foreign loanwords were written with kanji.

The Japanese katakana syllabary

In each column the rōmaji appears on the left, the katakana symbols in the middle and the kanji from which the symbols were derived on the right.

片仮名（カタカナ） katakana

a	ア	阿	i	イ	伊	u	ウ	宇	e	エ	江	o	オ	於
ka	カ	加	ki	キ	幾	ku	ク	久	ke	ケ	介	ko	コ	己
sa	サ	散	shi	シ	之	su	ス	須	se	セ	世	so	ソ	曽
ta	タ	多	chi	チ	千	tsu	ツ	川	te	テ	天	to	ト	止
na	ナ	奈	ni	ニ	二	nu	ヌ	奴	ne	ネ	祢	no	ノ	乃
ha	ハ	八	hi	ヒ	比	fu	フ	不	he	ヘ	部	ho	ホ	保
ma	マ	万	mi	ミ	ミ	mu	ム	牟	me	メ	女	mo	モ	毛
ya	ヤ	也				yu	ユ	由				yo	ヨ	與
ra	ラ	良	ri	リ	利	ru	ル	流	re	レ	礼	ro	ロ	呂
wa	ワ	和	wi	ヰ	井				we	ヱ	恵	wo	ヲ	乎
												n	ン	无

The symbols for 'wi' and 'we' were made obsolete by the Japanese Minsitry of Education in 1946 as part of its language reforms.

Additional sounds are represented by diacritics or combinations of syllables:

ga	ガ	gi	ギ	gu	グ	ge	ゲ	go	ゴ	kya	キャ	kyu	キュ	kyo	キョ
za	ザ	ji	ジ	zu	ズ	ze	ゼ	zo	ゾ	gya	ギャ	gyu	ギュ	gyo	ギョ
da	ダ	ji	ヂ	zu	ヅ	de	デ	do	ド	nya	ニャ	nyu	ニュ	nyo	ニョ
ba	バ	bi	ビ	bu	ブ	be	ベ	bo	ボ	hya	ヒャ	hyu	ヒュ	hyo	ヒョ
pa	パ	pi	ピ	pu	プ	pe	ペ	po	ポ	bya	ビャ	byu	ビュ	byo	ビョ
sha	シャ			shu	シュ	she	シェ	sho	ショ	pya	ピャ	pyu	ピュ	pyo	ピョ
ja	ジャ			ju	ジュ	je	ジェ	jo	ジョ	mya	ミャ	myu	ミュ	myo	ミョ
cha	チャ			chu	チュ	che	チェ	cho	チョ	rya	リャ	ryu	リュ	ryo	リョ
fa	ファ	fi	フィ	fu	フ	fe	フェ	fo	フォ						
va	ヴァ	vi	ヴィ	vu	ヴ	ve	ヴェ	vo	ヴォ						

The katakana for with the initial "v" are recent creations. This sound used to written with the ones with the initial "b" and some people still prefer to use those katakana.

Chinese

- Origins of writing in China
- The Chinese writing system
- Evolution of characters
- Types of characters
- Chinese numerals
- Simplified characters
- Chinese links
- Recommended books
- Phonetic transcription of Chinese
- Braille for Chinese
- Spoken Chinese

Origins of writing in China

Most linguists believe that writing was invented in China during the latter half of the 2nd millenium BC and that there is no evidence to suggest the transmission of writing from elsewhere. The earliest recognisable examples of written Chinese date from 1500-950 BC (Shang dynasty) and were inscribed on ox scapulae and turtle shells - "oracle bones".

In 1899 a scholar from Beijing named Wang Yirong noticed symbols that looked like writing on some "dragon bones" which he had been prescribed by a pharmacy. At that time "dragon bones" were often used in Chinese medicine and were usually animal fossils. Many more "oracle bones" were found in the ruins of the Shang capital near Anyang in the north of Henan province.

The script on these "oracle bones" is known as 甲骨文 (jiǎgǔwén) - literally "shell bone writing". They were used for divination, a process which involved heating them then inspecting the resulting cracks to determine to answers to one's questions. The bones were then inscribed with details of the questions and the answers. Most of the questions involved hunting, warfare, the weather and the selection of auspicious days for ceremonies.

Further information about the oracle bones:
http://www.chinapage.com/oracle/oracle00.html
http://www.lib.cuhk.edu.hk/uclib/bones/bones.htm

A collection of oracle bones in the National Palace Museum near Taipei.

Recently archaeologists in China have unearthed many fragments of neolithic pottery, the oldest of which date from about 4800 BC, inscribed with symbols which could be a form of writing. None of these symbols resemble any of the Shang characters and the likelyhood of deciphering them is remote given the paucity of material.

The Chinese writing system 中文

Chinese is written with characters known as 漢字 [汉字] (hànzi). Each character represents a syllable of spoken Chinese and also has a meaning. The characters were originally pictures of people, animals or other things but over the centuries they have become increasingly stylised and no longer resemble the things the represent. Many of the characters are actually compounds of two or more characters

How many characters?

The Chinese writing system an open-ended one, meaning that there is no upper limit to the number of characters. The largest Chinese dictionaries include about 56,000 characters, but most of them are archaic, obscure or rare variant forms. Knowledge of about 3,000 characters is sufficient to read Modern Standard Chinese. To read Classical Chinese though, you need to be familiar with about 6,000 characters.

Usage

Characters can be used on their own, in combination with other characters or as part of other characters. Click here to see how this works for the character for horse: 馬

Strokes

Chinese characters are written with the following twelve basic strokes:

A character may consist of between 1 and 84 stokes. The strokes are always written in the same direction and there is a set order to write the strokes of each character. In dictionaries, characters are ordered partly by the number of stokes they contain.

一	二	三	心	玉	竹	見	金	面	骨
1 stroke	2 strokes	3 strokes	4 strokes	5 strokes	6 strokes	7 strokes	8 strokes	9 strokes	10 strokes
魚	黃	鼎	鼻	齒	龍	龜	簡	識	覺
11 strokes	12 strokes	13 strokes	14 strokes	15 strokes	16 strokes	17 strokes	18 strokes	19 strokes	20 strokes
鐵	韃	體	鱷	黌	鼉	鸞	鸚	鸛	鱺
21 strokes	22 strokes	23 strokes	24 strokes	25 strokes	26 strokes	27 strokes	28 strokes	29 strokes	30 strokes
灪	籲	麤	齾	龘	龍龍	龍龍龍	龘龘		
31 strokes	32 strokes	33 strokes	35 strokes	36 strokes	48 strokes	64 strokes	84 strokes		

When writing Chinese, every character is given exactly the same amount of space, no matter how many strokes it contains. There are no spaces between characters and the characters which make up multi-syllable words are not grouped together, so when reading Chinese, you not only have to work out what the characters mean and how to pronounce them, but also which characters belong together.

Homophones

There are approximately 1,700 possible syllables in Mandarin, which compares with over 8,000 in English. As a result, there are many homophones - syllables which sound the same but mean different things. These are distinguished in written Chinese by using different characters for each one.

Not all the following characters are pronounced with the same tone, so to Chinese ears they sound different. To Westerner ears however they all sound the same. These syllables can be distinguished in speech from the context and because most of them usually appear in combination with other syllables.

傍	幫	梆	邦	榜	牓	綁	膀	傍	棒	磅	鎊
bāng	bāng	bāng	bāng	bǎng	bǎng	bǎng	bǎng	bàng	bàng	bàng	bàng
near	to help	watchman's wooden clapper	nation, state, country	list of names	tablet, plaque	to tie up	shoulder	to depend on, to draw near	stick, club, cudgel	pound (lb), scales	pound (£)

If you look closely, you will notice that some of the characters above have parts in common. These parts give you a clue to how to pronounce the characters.

More examples

Compound words

Chinese verbs and adjectives generally consist of one character (syllable) but nouns often consist of two, three or more characters (syllables):

電腦	飛機	大學	收音機	貓頭鷹	精神分裂症
computer	aeroplane	university	radio	owl	schizophrenia
(electric brain)	(flying machine)	(great learning)	(receive sound machine)	(cat-headed eagle)	(split mind disease)

More examples

Simplified characters
In an effort to increase literacy, about 2,000 of the characters used in China have been simplified. These simplified characters are also used in Singapore, but in Taiwan, Hong Kong, Macau and Malaysia the traditional characters are still used. Here are some examples (simplified characters in red):

語语 見见 間间 銀银 飯饭 魚鱼 紅红

More examples

Chinese characters, with some modifications, are also used in written Japanese and Korean, and were once used to write Vietnamese.

[contents]

Sample text in Chinese

繁體中文字 (Traditional Chinese characters)
人人生而自由，在尊嚴和權利上一律平等。他們賦有理性和良心，並應以兄弟關係的精神互相對待。

简体中文字 (Simplified Chinese characters)
人人生而自由，在尊严和权利上一律平等。他们赋有理性和良心，并应以兄弟关系的精神互相对待。

Hànyǔ pīnyīn transliteration
Rénrén shēng ér zìyóu, zài zūnyán hé quánlì shàng yīlǜ píngděng. Tāmen fùyǒu lǐxìng hé liángxīn, bìng yīng yǐ xīongdì guānxì de jīngshén hùxiāng dùidài.

Translation
All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.
(Article 1 of the Universal Declaration of Human Rights)

How the Chinese writing system works

The illustrations below of the various ways the character for horse is used in Chinese give you an idea of how the Chinese script works.

Evolution of the character
You can see below how the character for horse has evolved since it first appeared in the Oracle Bone Script during the Shang Dynasty (c. 1400-1200 BC).

Further information about the evolution of chinese characters

Basic meaning
The character 馬is pronounced *mǎ* in Mandarin and *máh* in Cantonese. It means horse and is also used as a family name.

Usage in compound words
The character is also used in horse-related compound words such as:

馬力 mǎlì - horse power 馬房 mǎfáng - stable (lit. "horse house")
馬上 mǎshàng - on horseback, immediately (lit. "horse on") 馬鬃 mǎzōng - mane
馬戲團 mǎxìtuán - circus (lit. "horse play group") 馬夫 mǎfu - groom (lit."horse man")
馬路 mǎlù - street, road (lit. "horse path") 馬磴 mǎdèng - stirrup (lit."horse step")

Radicals and phonetics
About 90% of Chinese characters contain a radical or *bùshǒu*, which gives you a clue to the meaning of a character, and a phonetic component, which hints at how to pronounce the character. The character for horse is used both as a phonetic compontent and as a radical.

The charcter for horse is used as a **phonetic component** in the following characters:

媽媽 māma - mother 瑪瑙 mǎnǎo - agate, cornelian 碼頭 mǎtou - dock, quay, wharf
螞蟻 mǎyǐ - ant 鎷 mǎ - masurium 禡祭 màjì - ritual of offering sacrifices to the god of war by troops on the eve of a battle
罵 mà - to swear, curse 嗎 ma - a particle which turns a statement into a question

The charcter for horse is used as a **radical** in the following characters:

馮 píng - to gallop 駕 jià - to ride, drive, pilot 騙 piàn - to cheat, swindle, deceive
騷 sāo - to disturb, agitate, worry, stinking 騾 lúo - mule 駝 túo - camel
驚 jīng - to startle, surprise, frightened, to marvel 驫 biāo - horses, to run, high speed

Further information about types of Chinese characters.

Usage in the transliteration of foreign words

The character for horse is also used for its phonetic value alone when writing foreign loanwords or the names of foreign people or places.

馬達 mǎdá - motor 馬賽克 mǎsàikè - mosaic 馬虎 mǎhū - perfunctory, careless, so-so
馬達加斯加 mǎdájiāsījiā - Madagascar 馬來西亞 mǎláixīyà - Malaysia
馬克斯 mǎkèsī - Karl Marx 馬哥孛羅 mǎgēbólúo - Marco Polo

The few foreign loanwords that exist in Chinese come mainly from English but the word *mǎhū* comes from the Sanskrit *moha* - ignorance. The syllables of *mǎhū* are usually doubled to make it *mǎmahūhu*. This is a common way to intensify the meaning of adjectives.

Simplified Chinese characters

The Simplified script (a.k.a. Simplified Chinese) was officially adopted in the People's Republic of China in 1949 in an effort to eradicate illiteracy. The simplified script is also used in Singapore but the older traditional characters are still used in Taiwan, Hong Kong, Macau and Malaysia.

About 2,000 characters have been simplified in a number of different ways (the simplified characters are shown in red):

Many simplified characters are based on commonly used abbreviations:

Others retain only one part from the traditional character.

Some replace the phonetic element of the traditional character with a simpler one that is pronounced in the same or in a similar way:

In some cases, several traditional characters are represented by one simplified character:

Recently the traditional characters have started to make a come back, particularly in southern China.

Blissymbolics

Origin

Blissymbolics were developed by Charles K. Bliss (1897-1985). Bliss originally called his invention "Semantography" and intended for it to be used as a universal written language which would enable speakers of different languages to commuicate with one another.

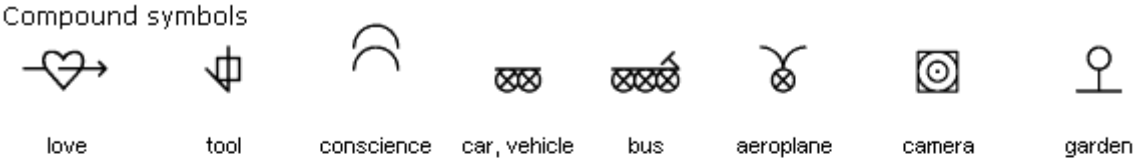
Since 1971 Blissymbolics have been used mainly as a communication aid for people with communication, language and learning difficulties. Such people have limited or no ability to use ordinary spoken and/or written language but manage to learn Blissymbolics.

Notable features

- Blissymbolics consists of over 2,000 basic symbols which can be combined together to create a huge variety of new symbols.
- The symbols can be formed into sentences and their order is based on English word order
- The symbols are made up of simple shapes designed to be easy to write.
- Blissymbolics are used in over 33 countries.

A selection of Blissymbolics symbols

Compound symbols



love | tool | conscience | car, vehicle | bus | aeroplane | camera | garden

Words

friend | pet | happiness | like | dislike | education | teacher | school

theatre | library | hospital | post office | city | village | telephone | office

Tengwar 

Origin

J.R.R. Tolkien created many languages throughout his life. He wrote in one of his letters that the tales of Middle-earth (The Hobbit, The Lord of the Rings, The Silmarillion, etc) grew from these languages, rather than the languages being created for use in the stories.

Tolkien also created a number of different alphabets to write his languages - Tengwar, or Feanorian letters, is the one which appears most frequently in his work. The way the vowels are indicated in Tengwar resembles Tibetan and other Brahmi-derived scripts.

Notable features

- Written from left to right in horizontal lines.
- Tengwar is written is a number of different ways known as "modes". For example there is a Quenya mode, a Sindarin mode and even an English mode. The phonetic values of the consonants (tengwa) and the ways vowels are indicated varies from mode to mode.
- Vowels are indicated by diacritics (tehtar) which appear above the consonant which precedes them (in Quenya mode) or above the consonant which follows them (in Sindarin mode). When vowels stand on their own or come at the beginning of a word, the diacritics appear over a special vowel holder. Long vowels are always attached to a vowel holder.
- Consonants are doubled by adding a wavy line below them.
- When followed by a vowel, the letters /s/ /ss/ and /r/ are written with the tengwa silme nuquerna, esse nuquerna and rómen respectively. Otherwise these letters are written with the the tengwa silme, esse and óre.
- When the letter /s/ follows another consonant it is written with a little downward hook.

Used to write

A number of different languages of Middle-Earth, such as:

Quenya, Qenya or High-Elven, the most prominent language of the Amanya branch of the Elvish language family. Tolkien complied the "Qenya Lexicon", his first list of Elvish words, in 1915 at the age of 23 and continued to refine the language throughout his life. It is based mainly on Finnish, but also partly on Greek and partly on Latin.

Sindarin, the language of the Grey-elves or *Sindar*. Tolkien based Sindarin on Welsh and originally called it gnomish.

Sylvan, Westron, etc

Tengwar can also be used to write English, Welsh, Scottish Gaelic, Swedish, Polish, Esperanto and a variety of other languages.

Quenya mode

Consonants

t	tinco (metal)	p	parma (book)	c/k	calma (lamp)	kw	quesse (feather)
nd	ando (gate)	mb	umbar (fate)	ŋg	anga iron	ŋgw	ungwe (spider's web)
th/s	thúle/súle (spirit/wind)	f	formen (north)	kh	harma/aha (treasure/rage)	khw	hwesta (breeze)
nt	anto (mouth)	mp	ampa (hook)	ŋk	anca (jaws)	ŋkw	unque (hollow)
nt	númen (west)	mp	malta (gold)	ŋ	ŋoldo/ŋoldo (one of the Noldor)	ŋw	ŋwalme/ŋwalme (torment)
r	óre (heart / inner mind)	v	vala (angelic power)	y	anna (gift)	w/v	wilya/vilya (air/sky)

Additional letters

r	rómen (east)	rd	arda (region)	l	lambe (tongue)	ld	alda (tree)
s	silme (starlight)	s	silme nuquerna (silme reversed)	z/r/ ss	áze/áre/esse (sunlight/name)	z/r/ ss	áze nuquerna (áze reversed)
h	hyarmen (south)		hwesta sindarinwa	y	yanta (bridge)	w	úre (heat)
h	halla (tall)		short carrier		long carrier		

Sindarin mode

Consonants

t tinco	p parma	kh calma	kw quesse
d ando	b umbar	g anga	gw ungwe
th thúle	f formen	kh harma	khw/hw hwesta
dh anto	v ampa	gh anca	ghw/w unque
n númen	m malta	ŋ ŋoldo/holdo	ŋw ŋwalme/nwalme
r óre	w vala	y anna	w wilya

Additional letters

r rómen	rh arda	l lambe	lh alda
s silme	s silme nuquerna	z/ss áze/áre/esse	z/ss áze nuquerna
h hyarmen	hw hwesta sindarinwa	y yanta	w úre
h halla	short carrier	long carrier	

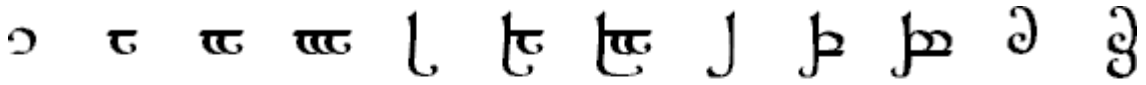
Vowels (same for Quenya and Sindarin modes)

a	e	i	o	u	y	á/aa	é/ee	í/ii	ó/oo	ú/uu	ý/yy

Punctuation marks

comma period exclamation mark question mark

Numerals



0 1 2 3 4 5 6 7 8 9 10 11

Tengwar numbers are written "backwards" and read from right to left, for example 19352 is written: 

Pronunciation of Quenya

Short vowels					Long vowels				
a	e	i	o	u	á	é	í	ó	ú
[a]	[e]	[i]	[ɔ]	[u]	[a:]	[e:]	[i:]	[o:]	[u:]

Diphthongs

ai	ei	oi	ui	au	eu	iu
[aɪ]	[eɪ/ej]	[oɪ/oj]	[wɪ]	[au]	[eu]	[iu]

Consonants

c	d	f	g	h	l	m	n	p	r	s	t	v	y	w
[k]	[d]	[f]	[g]	[h]	[l]	[m]	[n]	[p]	[r]	[s]	[t]	[v]	[j]	[w]

Palatized and labialized consonants

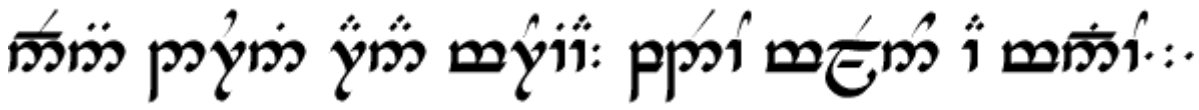
gw	hy	hw	ly	nw	ny	qu	ry	ty
[gʷ]	[x/hʲ]	[hʷ]	[lʲ]	[nʷ]	[nʲ]	[kʷ]	[rʲ]	[tʲ]

Quenya pronounciation provided by Joshua Boniface

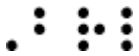
Sample text (Quenya)

Transliteration / Translation
Elen síla lumenn' omentielvo / A star shines on the hour of our meeting

Sample text (Sindarin)



Transliteration / Translation
Ennyn Durin Aran Moria: pedo mellon a minno.
The Doors of Durin, Lord of Moria. Speak, friend, and enter.
(inscription on the Gate of Moria)

Braille 

Braille is writing system which enables blind and partially sighted people to read and write through touch. It was invented by Louis Braille (1809-1852), a French teacher of the blind. It consists of patterns of raised dots arranged in cells of up to six dots in a 3 x 2 configuration. Each cell represents a letter, numeral or punctuation mark. Some frequently used words and letter combinations also have their own single cell patterns.

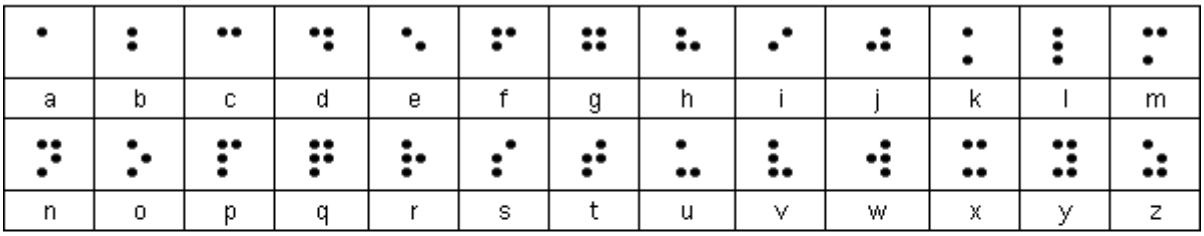
There are a number of different versions of Braille:

- **Grade 1**, which consists of the 26 standard letters of the alphabet and punctuation. It is only used by people who are first starting to read Braille.
- **Grade 2**, which consists of the 26 standard letters of the alphabet, punctuation and contractions. The contractions are employed to save space because a Braille page cannot fit as much text as a standard printed page. Books, signs in public places, menus, and most other Braille materials are written in Grade 2 Braille.
- **Grade 3**, which is used only in personal letters, diaries, and notes. It is a kind of shorthand, with entire words shortened to a few letters. Examples: brl=braille. bl=blind. gd=good.

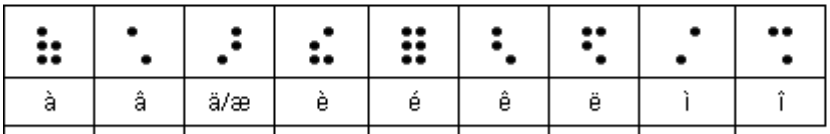
Braille has been adapted to write many different languages, including Chinese, and is also used for musical and mathematical notation.

Braille

Basic letters

												
a	b	c	d	e	f	g	h	i	j	k	l	m
n	o	p	q	r	s	t	u	v	w	x	y	z

Accented letters

								
à	â	ä/æ	è	é	ê	ë	ì	î
ï	ò	ô	ö/œ	ù	û	ü	ç	

Words and abreviations

a	but	can	do	every	from	go	have	just	knowledge	like	more	not
people	quite	rather	so	that	us	very	will	it	you	as	and	for
of	the	with	child/ch	gh	shall/sh	this/th	which/wh	ed	er	out/ou	ow	bb
cc	dd	en	gg; were	in	st	ing	ar					

Punctuation

,	;	:	.	!	( )	? “	*	”	’	-

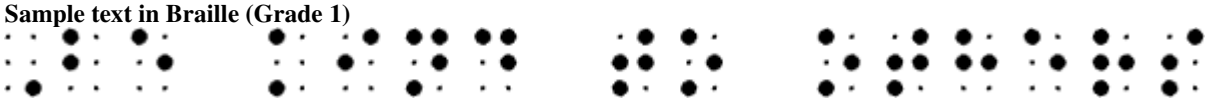
Numerals

1	2	3	4	5	6	7	8	9	0

Special signs

letter sign	capital sign	numeral sign	numerical index sign	literal index	italic sign

Sample text in Braille (Grade 1)

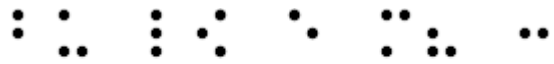


Transliteration: "Be kind to others"

Sample text and other information provided by Samuel Barnes

Braille for Chinese

When Braille is used to write Chinese, it represents the sounds of the language rather than the characters. It is written from left to right in horizontal lines running from top to bottom. Each syllable is made up of three Braille letters: one for the initial, one for the final and one for the tone, though the tones marks are rarely used. Words are separated by spaces. Where there is no possibility of confusion, some initials are written in the same way. For example g and j, and h and x in Mandarin Braille.



Braille for Mandarin 布来叶盲文 (bùláiyè mángwén)

Initials

b	c	ch	d	f	g, j	h, x	k, q	l	m	n	p	r	s	sh	t	z	zh

Finals

a	ai	an	ang	ao	o, e	ei	en	eng	er	i / yi	ia / ya	ie / ye	iao / yao	iu / you	ian / yan	in / yin	iang / yang
ing / ying	iong / yong	ong / weng	ou	u / wu	ua / wa	uai / wai	uan / wan	uang / wang	ui / wei	un / wen	uo / wo	ü / yu	un / yuan	ue / yue	en		

Tones

1	2	3	4

Punctuation

,	.	;	:	?	!	(	)	–	…

Source: www.braille.ch/pschin-e.htm

Braille for Cantonese 盲人點字 (màahngyàhn dímjih)

Initials

f	h	g	h	l	m	b	p	s	d	t	w	j	dz	tz	gw	kw	n

Finals

a	ɛ	i	ou	ɔ	ei	ai	ɐu	au	ei	iu	u	ɔi	y
œy	uí	ɐm	am	im	ɐn	an	in	ɔn	œn	un	yn	ɐŋ	
aŋ	eŋ	iŋ	ɔŋ	uŋ	œŋ	œ	ɐk	ak	ik	ɛk	ɔk	uk	
œk	ɐt	at	it	ɔt	œt	ut	yt	ɐp	ap	ip	m	ŋ	

Tones

high level	high falling	mid rising	mid level	low level	low rising	low falling

Punctuation

,	.	;	:	?	!	“	”	( )	–	…

Source: www.hadley-school.org/Web_Site/8_d_chinese_braille_alphabet.asp

12480 Alphanumeric System 

12480 was designed in 2002 by Bradley Tetzlaff from Waukesha, Wisconsin, USA. It was invented for both use in a computer game named Ecclemony (1E78) and as a basis for constructed languages. It was also designed to show how a true alphanumeric* writing system looks and works.

12480 is not based upon phonemes, but rather upon binary. It achieves complete universality with an optimal amount of applications from its binary basis. A writing system based on phonemes will only last as long as the human voice is used. 12480's binary foundation will last as long as numbers exist.

* "Alphanumeric" is used here to describe the combination of an alphabet and a numeral system.

Notable features

- 12480 is composed of various scripts, each of which could be considered a separate writing system on their own. Each script has its own specialities and advantages.
- Each script is used to represent either a word or a number by default. Visit http://www.12480.8m.com/scripts.html to see a list of what each script's default is.
- Each alphanumeric has both a consonant and a vowel assigned to it. They can be used interchangeably except for the initial phoneme--An initial consonant represents a word and an initial vowel represents a number.
- The punctuation is limited to break symbols, grouping symbols, and radix indicators, but it may be extended in future versions.
- Words are typically separated with a circle instead of a space. A space is used to group symbols in radixes lower than 16 into hexadecimal segments.
- 12480 is usually written from top to bottom and from left to right. A baseline underline is used to show how the text is oriented.

Used to write

Binary (radix 2), quadnary (radix 4), hexadecimal (radix 16), radix 256, and all other numeral systems based on a power of two. Anything that can be expressed with a numeric value can be written using 12480.

Hexadecimal Number / Binary Number	Bubble Script	Four Line Script	Dot Script	Slash Script	Diamond Script	Letter Script	IPA Pronunciation [Consonant, Vowel]	Associated Color
0 0000							[h or □ʰ, a]	
1 0001							[j or □ʲ, i]	
2 0010							[ɹ or ɾ or □, æ]	
3 0011							[s, ɪ]	
4 0100							[ɸ or f, ɛ]	
5 0101							[p, e]	
6 0110							[t, ʌ]	
7 0111							[k, ɤ]	
8 1000							[ɣ or □ˠ, ɒ]	
9 1001							[w or □ʷ, u]	
A 1010							[l or □ˡ, ɐ or ɜ]	
B 1011							[ʒ, ʊ]	
C 1100							[ð, œ]	
D 1101							[m, ø]	
E 1110							[n, ɔ]	
F 1111							[ŋ, o]	

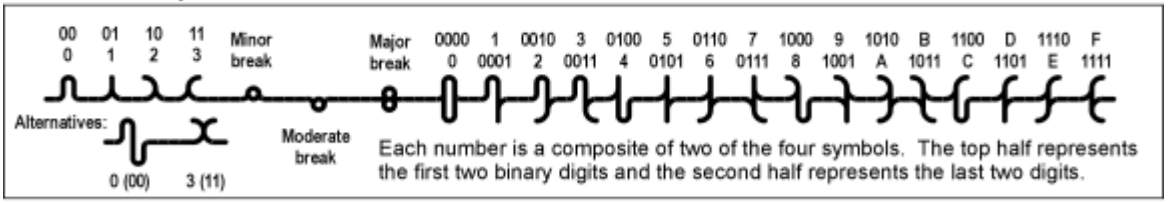
Leading zeros may be ignored. Both unvoiced and voiced consonants may be used, but the consonant displayed is recommended. The consonant and vowel both represent the same number.

Punctuation

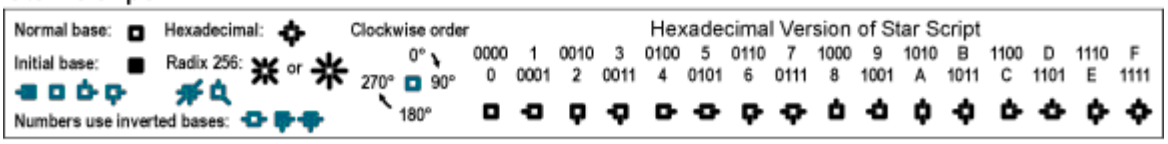
Minor break (Comma)	·	Grouping Symbols		Integer-fraction divider (number mark)		Radix 256 mark	
Moderate break (Space)	o			Letter mark		Hexadecimal (radix 16) mark	
Major break (Period)	8	Capital underline		Negative mark/bracket		Quadnary (radix 4) mark	
Colon	:	Baseline underline		Hyphen	~	Slashes		Binary (radix 2) mark	

All other punctuation should be written using binary. The integer-fraction divider should be used in the same way as a decimal point is used.

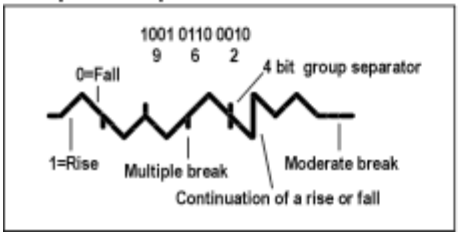
Cursive Script



Star Script



Graph Script



Quadnary (Radix 4) Scripts

Number	Slash Script	Circle Code	Tone	Click Signal	Amplitude-Frequency Wave (AFM)	DNA Nitrogen Base
0 00				⊙ or ʷ/ʷ	A↓ F↓	Adenine
1 01				| or ‼	A↓ F↑	Guanine
2 10				!	A↑ F↓	Cytosine
3 11				‖	A↑ F↑	Thymine (RNA Uracil)

Sample texts

Betamaze alphabet

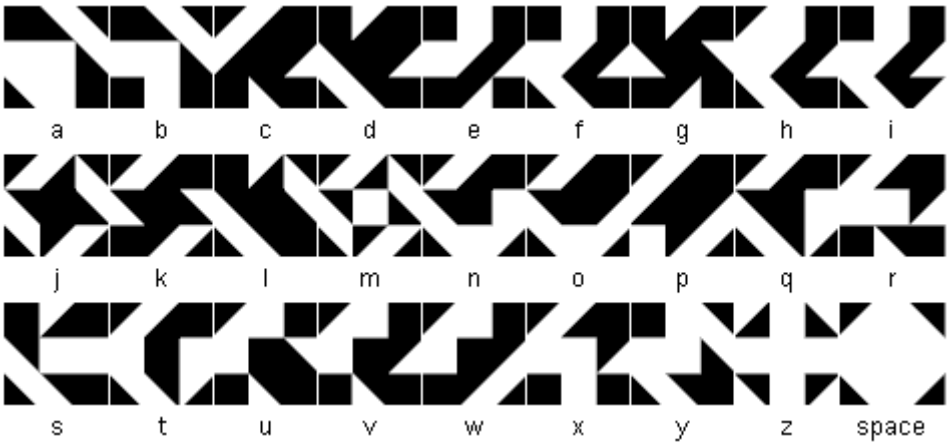
The Betamaze alphabet is the creation by Terrana Cliff (rillani@yahoo.com), an American art student in California. It is designed to draw mazes, which Terrana has been interested in for a long time.

Terrana would like to encourage other people to find new (perhaps more artful) ways to meet the simple demands of the concept.

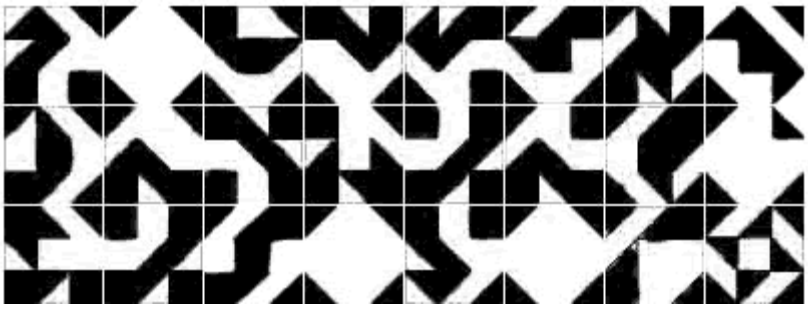
Notable features

- *All the letters connect together so they can form paths.*
  To make sure this happens, they all fit within a 3x3 grid. Letters are made from black squares and triangles in the grid. To allow the paths to connect, every letter has white space on the sides of the 3x3 grid.
- *Paths can branch, terminate, and come together.*
  The locations on the 3x3 grid that are not used for connecting are used for giving each letter its shape. Within each letter, the black space is used to close or alter the path between the white connection spaces. Some letters have more black space in the grid than others. Some letters only allow a 3-way path, some are 2-way, some turn the path 90 degrees, some close in all directions, and some open to all directions.
- *Path structure can be altered without having to alter spelling, word order, etc.*
  Every letter has a unique shape, unlike in the english alphabet, where some letters have the same shape (m and w are the same shape, just vertically flipped). Each letter can be turned upsidedown or flipped without a change in its meaning, so the direction of the path can be changed.

The Betamaze alphabet



Sample text/maze



Transliteration
I think; therefore I am.

Ihathvé Sabethired 

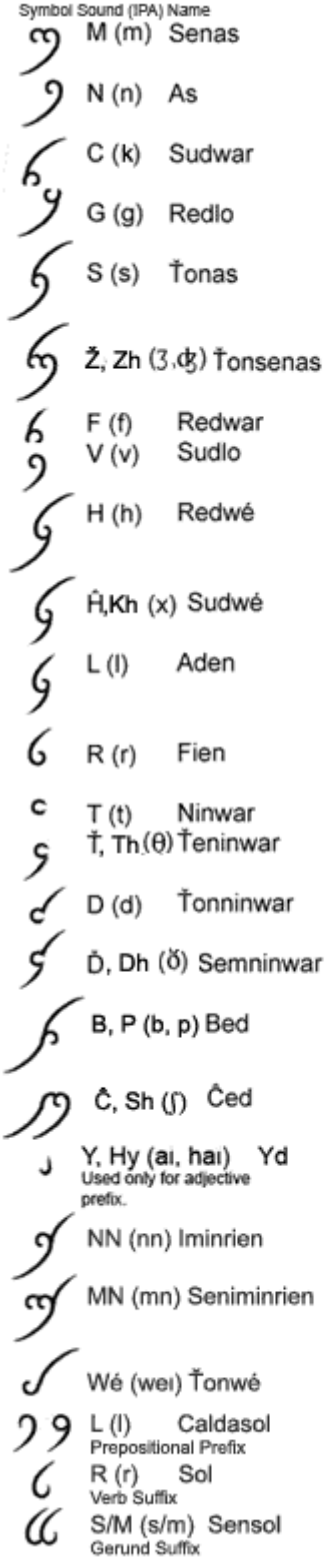
Ihathvé Sabethired is the creation of Jason Liekhus. It developed from an older alphabet called Ihadva, which Jason based on of Arabic and Tengwar. The script is used to write a language called Sabethir, meaning "Eastern Language", which Jason invented for use in a fictional world.

Noteable features

- Ihathvé Sabethired is an abjad which is written fully vocalised.
- It includes a number of ideographs for verb conjugations, some conjunctions and pronouns.
- It is cursive and is written from right to left

Ihathvé Sabethired script

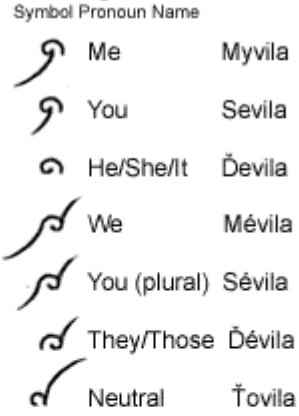
Common Characters
Symbol Sound (IPA) Name



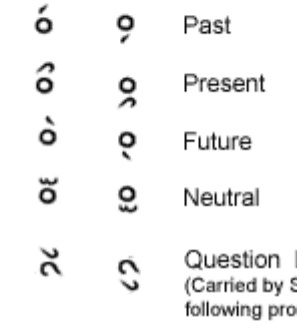
Diacritics
Symbol Sound (IPA) Name

Punctuation
Symbol Use Name

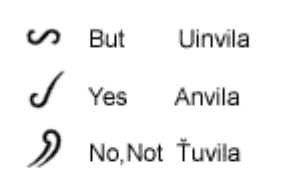
Verb Ideograms
Symbol Pronoun Name



Verb Tenses (Add to Pronoun Symbol)
Positive Negative Tense

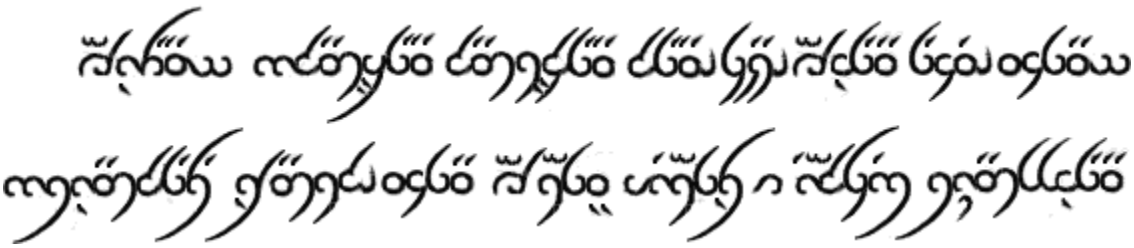


Other Ideograms
Symbol Meaning Name



Numerals

Sample text



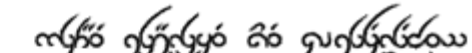
Transliteration
Erthéhyathra eratidhiahythuelyared arethoved aregoled. Aceidhia eratisevuin maĥdya i sirvya, orvydhia erthéhydavenin saradén.

Translation
All human beings are born free and equal in dignity and rights. They are endowed with reason and conscience and should act towards one another in a spirit of brotherhood.
(Article 1 of the Universal Declaration of Human Rights)

Longer sample text (Tower of Babel)

Some Invaluable Phrases in Sabethir – By Jason Liekhus

My hovercraft is full of eels.
Idrauharn'ymon adh agvillolehin afel.



Merry Christmas and a Happy New Year.
Seluin ieĉwaĥeraté'hysél i brason'yavras.

I don't understand.
Ceninan

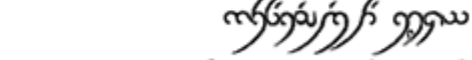
Please speak more slowly.
Athwar hiaré'hyomana.

Thank you.
Ameth.

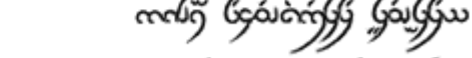
Welcome.
Illanesil.

I love you.
Yam se.

Happy Easter.
Seluin baĉa'hyavras.



Tongue Twister: Halvu'hyalvo halmadhvo'hyathra vy'hythvo.
A woman's mind will always keep count of one's actions.



'Useful' Phrase: Get that Buddhist monk out of my coffee.
Athucearé bodhiĥavanty'hyar cavélin'ymon.

Idiom: To cut the blue from one's face – To dishonor and shame one's self.
Silidhar idsalin fien.

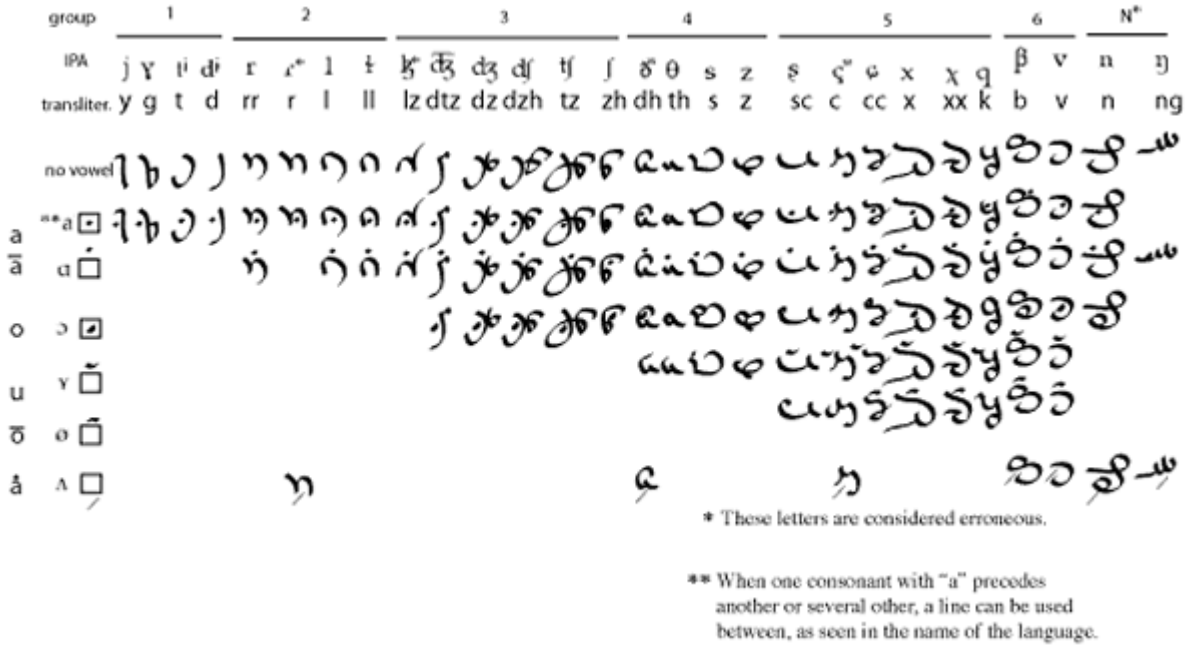
Sunscript

Sunscript is the creation of Colin Williams. He created it when he had nothing better to do in school and based its appearance partly on Arabic and partly on some of the Indian syllabic alphabets.

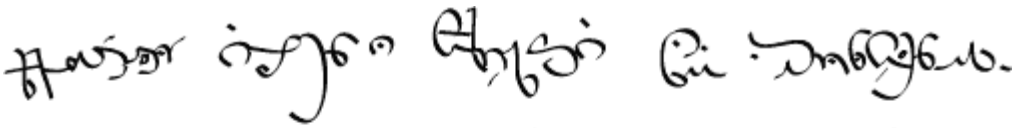
Colin uses Sunscript to write "navthāladasa□", a language he invented after the creating the alphabet. The language is based primarily on German and Latin but has been distorted almost totally out of recognition so as to sound more like an Indian language.

Notable features

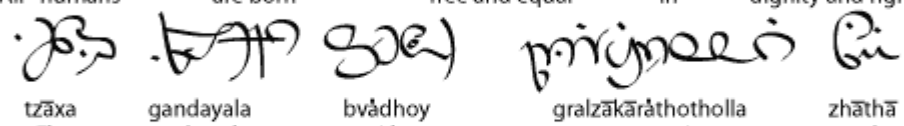
- Sunscript is a fully vocalized abjad
- It is cursive and written left to right in horizontal lines
- Vowels are represented with diacritics, however; the vowel "a" can be simplified if it occurs in more than one leter in a row by drawing a line between consonants (e.g. the example in the name of the language).
- The language uses a system of consonant-vowel groups. The first group takes the first vowel, the second the first and second vowels, the third the first three, etc. The letters "r", "lz", "dh" and "c" are erroneous letters and take slightly different vowels than their greater group.



Sample text in Sunscript



gayangāvōcolzā	lānodzla	dhāyarångablā	zhāthā	xarazhadhådzong .
Humans-all	born	free-equal	and	dignity-rights .
All humans	are born	free and equal	in	dignity and rights .



tzāxa	gandayala	bvådhoy	gralzākārå thotholla	zhāthā
They	endowed	with	reason-conscience	and
They	are endowed	with	reason and conscience	and

vrādagacādhåddhāla	sāvandzho	xxuyathāla .
brotherhood-acting	selves-to	should .
should act towards one another	in a spirit	of brotherhood .

How to Create a Language - http://www.angelfire.com/ego/pdf/ng/lng/how/

If you enjoy this, Pablo would love to get a postcard from you. Mail it to:

Pablo Flores
J. J. Paso 6038
2007AKT Rosario
Argentina

How to create a language

by Pablo David Flores
(partly based on Mark Rosenfelder's Language Construction Kit)

[All the pages of **How to create a language** can be downloaded for offline browsing in a .zip file. That doesn't include multimedia content. A big consolidated page with all the topics is also available for reading, and is a bit more suitable for printing.]

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 auxilliary 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, diachronical change of both grammar and phonology, and methods of word generation. The full table of contents is available elsewhere. 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 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. 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 which 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.

These pages do not require any plug-in or fancy gadget in order to be viewed correctly (not Flash, not Shockwave, not even Java). However, it is recommended that you use a browser with the ability to interpret Cascaded Style Sheets (CSS specification). Though not required, these pages are compatible with **Opera**, which provides support for certain innovations in the standard allowing for easier navigation.

Also, a couple of topics are accompanied by sound samples in MP3 format, which was chosen since it produces compact files that can be listened to, recorded and/or modified with software tools anyone can access for free. These samples are not indispensable for the comprehension of the rest of the content.

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. 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.

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 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 [ɸ] (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 /ts/).

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 Indoeuropean 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, [pʰ] (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 /pʰ/ 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 English or German, but it's quite a common phoneme in other languages, like Hawai'an (the apostrophe ʻ represents /ʔ/, the glottal stop). 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 **implossives**. 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 implossive 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):

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

(where /w/ is actually labiovelar, not just labial; /j/ is palatal, not alveolo-palatal; and /ɹ/ 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 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 says, the key to a

naturalistic language is to add (or substract) 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 /ptk/. From what I've been able to gather, the average language has twice as much 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 (/aiu/, 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 /u/).

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, /ɯ/ and /ɨ/ 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 alive, rodent) is in the middle, right over the *a*; it's 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, /ɰ̃/, 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 disbalanced. 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 /ɑ/ 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 accute 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* ['sekulus]); else stress falls on the first-to-last syllable (as in *secundus* [se'kundus]).

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. (Jump here for more details.)

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-reeducation 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", positing and answering an interesting question: how do you sing a song in a tonal language?

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 intrincate 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 inconditional. An inconditional 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 especifically 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.e'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* /sEmwa/ vs. *C'est Anne* /sEt'an/.
- **Prothesis**: 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 inconditional 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 inconditional. For example, the change of Latin /k/ (written *c*) in Romance languages is regarded as inconditional, 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 treated in 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 irrestricted**. 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 **reborrowed** from the ancient language (or another one), like English *foreign*, which comes from Proto-Indoeuropean **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, viril-werewolf, blanch-blank, etc.

Harmony

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 cathegorized 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 Northeastern 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 diaeresis over vowels are diacritic marks: *á è î ÿ*. English doesn't use any diacritic marks. Spanish shows some stressed vowels with an accute 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 diaeresis over them (*ö ü*). Danish writes a kind of rounded *a* with *å*, and a fronted *o* with *ø*. Many languages have nonstandard 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 /ʒ/, but most people will surely read something completely different from /ʒ/ 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 /waˈzo/, 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 viceversa.

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. [See a description of Japanese and Chinese writing here. Includes a *hiragana-katakana* chart!]

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, 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 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 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 cathegorization is that languages can be **inflecting**, **agglutinating**, or **isolating**. This cathegorization 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 Indoeuropean 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 polisynthetic 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 grammaticalized 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 Indoeuropean 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, 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.

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, started by a question by Fredrik Ekman: are there languages without adjectives?

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: 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.

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, 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).

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:

> Kanojo 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

Sentence order	Phrase order	Adpositions
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 SOV, 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 auxilliary) 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, noteably 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 Indoeuropean 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 (absolutivo, 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 Indoeuropean 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).

See also this article about split ergativity.

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:

- a. 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. Lenguages with a fluid S-role (**Fluid-S**), in which the decission 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 langauge, 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 auxilliaries ('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 willfully 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):

- a) event vs. state
- b) control
- c) 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 Lakhota 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-Indoeuropean (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 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 inintelligible 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 Rigveda 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 (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.
- 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. It lets you specify beginning, medial and final consonants, clusters, vowels and diphthongs, and the number of syllables you want.

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 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, 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, 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.

If you're a J. R. R. Tolkien fan, you can find descriptions of the languages he invented in Ardalambion, the Tongues of Arda.

For a look at some real world scripts, you can visit Ancient Scripts, 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. Follow the arrows!

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 a 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. 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 inintelligible (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. Don't go away without checking out Language Creation.

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.

Conlang Errors - http://www.angelfire.com/ego/pdf/ng/lng/lang_errors.html

The purpose of this page is to display and correct several errors I've found (newbie) language creators make all the time. I'm certainly not up to the challenge of a complete, well-articulated essay on the matter; I'm not a linguist or a philologist or a phonologist, and almost everything I know I owe to people who corrected me. That's why I'm risking to be named Obnoxious Pedantic Lecturer of the Millenium by some people who are the source of these errors, and the target for the corrections. I have a compulsion for correcting mistakes.

I will say it in Spanish: *La verdad no ofende* ('Truth does not offend'). The truth is many people are creating languages (so to speak) without real knowledge. I was one of those a few years ago. La verdad no ofende, so I didn't resent it when my lack of knowledge was pointed out. But then, I like to learn. Most people I've met in the conlanging environment like to learn too, though many would not bother to learn too much. Some people don't like to learn; they just want to do as they please. All of them have the right to do so -- just don't write to me telling me "I do as I please, my language is nice and you're a stupid because you dismiss it". On the other hand, "You're a geek" is accepted, though not welcome given the implicit tone.

Enough. Let's enter the slaughterhouse now...

Here's my language (points to a dictionary)

If you can enclose it in a dictionary (in the normal meaning of the word), then it's not a language, but a code. Now, an encyclopedia would be useful. A language doesn't consist of words and meanings only; it has a phonology, and a grammar, and many many subcategories under those. If you replace English words for [your language] words and maybe add some strange letters and diaeresis over vowels, you're creating a nice code, but nothing else.

As I said, you can do as you please with your creation, but if you call it a language, it should be a language. I can't boast to have mastered chess if I use the board to play checkers.

I don't have that sound -- there's no letter for it in my con-script

This one is very frequent. It seems many people blend sound with sound representation -- and even worse, they do it in the opposite order. I'll just go biblical here: *in the beginning there was the (spoken) Word!* Are you telling me you can't produce a sound that you don't have a letter for? Did you learn to read *before* you learned to speak?

English has no letters for many very common sounds. English has no single letters for several sounds found in English -- it has to use digraphs which usually don't have a single reading. This is not important at all. On Earth, first you learn to speak, and then, if you're lucky, you go to school and learn to read and write.

Recipe: don't mix sounds and letters. Letters are not sounds. The same letter or combination of letters can be used to represent many sounds. The letter *j* is used for four different sounds in English, French, German and Spanish. Letters do not *exist* in a language -- they are conventional marks that belong in other fields of study. Once you have your sounds, assign them to letters, but don't delete sounds only because they're unrepresentable -- no sound is, since you can always invent.

The sound [X] and the sound [Y] are the same in my language

Nope. The sound [X] and the sound [Y] are different in all languages. Lemme guess: you mentioned them because they both exist in English, right? What you're saying here is that people do not distinguish between them. Actually, [X] and [Y] are called allophones; they are *not* the same sound, but they're treated similarly by speakers. They are the same phoneme -- you can't distinguish two words only by them. In general, if [X] and [Y] are allophones, they're in complementary distribution: you can't have one in the same environment as the other (for example, between vowels you pronounce [X], but elsewhere you pronounce [Y]). If you exchange them, it sounds wrong, but you can't produce a different word.

You have to say when you will pronounce one or the other. Free allophonic variation, if I got it right in the first place, is not common.

On the other hand, maybe you just wanted to say you only have [X], not [Y] (or the other way round). As in "I have [p], but no [b]". That's all right -- you don't have to clarify that. There are many sounds you don't have, even common sounds. You can't mention them all.

How do you say that in English?

This one is close to the one that immensely bothers abstract artists: "What does it mean?" Sometimes you can translate more or less properly and convey the original meaning. Sometimes you cannot. As for myself, I love it when you cannot. Two languages need not be terribly different or alien to each other in order to have untranslatable utterances. Off the top of my head, the English expressions 'go ballistic', 'how come' and 'set sail' are untranslatable in Spanish (you can certainly find rough equivalents, but no literal translations, and they lack the original force). And in Spanish you can say 'se mató' and not knowing if it means 'he killed himself' or 'he got killed' or just 'he died by accident'. Such ambiguities and quirks are what gives a language a definite character.