A thesis submitted in partial fulfillment of the requirements
For the degree of Master of Science In
Computer Science

By
Mark Alan Bondurant

December 2011
The thesis of Mark Alan Bondurant is approved:

_________________________________________  ___________
Rick Covington  Date

_________________________________________  ___________
Diane Schwartz  Date

_________________________________________  ___________
Gloria Malara, Chair  Date
Table of Contents

Signature Page ..................................................................................................................... ii
Abstract .................................................................................................................................. iv
1.0 Creativity ........................................................................................................................ 1
2.0 Being Foreign ................................................................................................................... 3
3.0 Simple Plumbing ............................................................................................................. 9
4.0 Typing Chinese ............................................................................................................... 26
5.0 Ordering Things ............................................................................................................. 33
6.0 Universal Meanings ....................................................................................................... 44
7.0 Quantifying Culture ..................................................................................................... 51
8.0 Final Summary .............................................................................................................. 63
References ......................................................................................................................... 66
ABSTRACT

Internationalization and Localization

By

Mark Alan Bondurant

Master of Science in Computer Science

The purpose of this thesis is to identify the causes of recent international business failures, why our software in many cases is not only failing to make inroads into foreign markets but is actually fostering the development of foreign competition, essentially turning receptive audiences into dangerous competitors. I propose that these problems are not just political or business related, but are caused by fundamental problems in way we approach software development. I intend to expose some of these weaknesses and to suggest ways that we can remediate them.

Internationalization for the purposes of this thesis is the adaption of software for use anywhere in the world. It’s the expansion of standards to cover the planet as a whole. Localization is the adaptation of computers to a particular locale or culture. It’s the adaption to and inclusion of a specific locale, or in the general sense to all locales. Computers must relate to the people who use them, otherwise the user cannot operate them to maximum advantage. The people who use them must be able to input commands and data and they must understand the response provided by the computer.

We are shoehorning ourselves into standards, and stretching and bending standards to fit around ourselves. Like the currency union in Europe, we are looking for a point of expedient commonality around which we can balance ourselves. Pitted against this is the need to retain local cultures in the face of that expediency in the Internet, where the pressure of the market demands that local culture compromise with dominant world cultures, often to the destruction of the individual.

A world standard is evolving, one which we no longer control. We must participate or be left behind. To participate, we must change the way we think about development, raise our awareness of standards developing outside the US, and develop our software for this new world market rather than expect that market to come to us.

This is a subsection of a larger work and deals mostly with problems of language and the measurement and quantification of culture as a means to unify our approach to international software development.
1.0 Creativity

The last century in computing was an American one. Any way you measure it, any piece of software or hardware was either built by Americans or was cloned from something built by Americans. The world danced to our tune. Companies like Microsoft and Apple built their business models around the knowledge that 90% of their business would be American or European. The core of their construction was English. This was just practicality. We built for the existing market.

In this rapidly expanding technical world, building for existing markets in the same ways Western computing had for the last six decades worked well. But the winds of the technical world shift very quickly. The world market in computing is looking less and less American each day. We drag the short term decisions made in the last century with us in the legacies such as words like “Print” or “Write” in computer languages, assumptions about word and sentence structure, and the forms of input devices like keyboards and phones, are the kinds of standards that we Americans still naively take for granted.

If US computing is going to continue to exist to face the competition that is rising throughout the world, then it’s going to have to embrace this world, and change. This doesn’t necessarily mean leading the change. This is not always practical since cultures as they emerge will find their own zone of comfort developing their own solutions. Being second in innovation too is often the wisest path. It just means understanding their solutions as they evolve and incorporating them into our products in order to maintain and perhaps even expand our market dominance.

But we are by default, as is every other human being on this planet, culturally blind. In this thesis, I will attempt to illustrate ways we miss things without knowing it, through a lack of training and experience in things foreign. Not the colorful travel show foreign, but in basic habits of thought that were evolved for practical reasons.

Figure 1, no country is mono cultural. For instance, H-Mart is US based and has 17 locations in the US. Notice how crowed the page is.
We are no longer working in a monochromatic marketplace. It is time for computing to incorporate a little anthropology into its thinking. We, as Americans, must understand our customers all over the world if we are to garner their business. We must produce a truly international product. Even if you are a lowly programmer or engineer, you produce a product that can be seen all over the world. You have international customers.

Great companies, innovators, have attempted to walk across the surface of our planet and have stumbled at great cost to our country and the future of our children. They didn’t just run into barriers they found as they moved across our borders. They built their own barriers through ignorance. They took receptive and eager audiences and turned them into competition. This is a process that we as human beings have repeated not just in Internet development, but time and time again throughout history. It is through history that we can learn what to do.

We have entered a post technical era in terms of information science, where the new issues are in use rather than plumbing. Our operating systems, our compilers, our desktops are largely built out. The Internet is, except for the problems posed by expansion and embellishment, what it will be. We may see more movies, better resolution on our PDAs, thinner laptops, bloating and reoptimization of development frameworks, more surveillance, more censorship, and perhaps even cleverer ways to organize things. But this is just more, not different or even better.

When we create a web page, it’s visible all over the world. There is art and craft in our work and even the choice of the color of our fonts can have significant implications to different peoples. That choice of color is communication of ideas and emotion. Art at its best is a language that communicates ideas and feelings at multiple levels. Aren’t our creations as programmers the same, especially within the multimedia context of the Internet? There is a movement called “convergence” in science in which scientists are required to carry a mixed focus. For example: chemistry/biology, chemistry/physics, and astronomy/biology. It’s time for computer scientists to carry a mix as well.

We as engineers must grasp hold of a wider view because our work is seen the world over. Creating a building, a fixed object, can be accomplished with fixed ideas and methods. Creating something as slippery as a successful web page is a little more difficult, even if that page is only geared to a single national audience. None of us can say we will only ever work within one culture. After all, even nations themselves are multicultural.

How can be become better at working with foreign cultures? There may be ways that we can generalize cultures as a whole that have not yet been visited. There are problems
coming from localization that we have not yet come to terms with. Issues involving the unexpected aspects of language that upset the data handling structures we have developed for instance. There are problems of sensibility and propriety, are we offending our audiences? The problem could be as simple as understanding the differences in the way others tell time and measure things, which is certainly part of it. But most of all, there are the problems in ourselves that we cannot see, this paragraph being a prime example.

All of the ways listed above are about finding ways for us to deal with others, not us in general. This is a subtle difference. When you are attempting to work with programmers from Nairobi, designers from Vietnam, and copy editors from France, you cannot think in terms of us and them. You must work in terms of cultural generalisms and cultural specificity, internationalization and localization. There is no longer any us and them. This is the topic of this paper.
2.0 Being Foreign

How do problems of culture manifest themselves? If you think about China you may think foreign. The writing, the history, the languages, and even the plants and animals are different. But who is really foreign? Foreign is a state of ignorance, an unfilled gap that can only exist in the space between two people. Neither one is foreign. To us they are foreign. To them we are foreign. It is relative. If you only think that China is foreign then you are only half correct. If you think that there are cultural differences that exist for practical reasons, necessities that are not easily compromised, then you are well on your way to dumping the idea of foreign. Unfortunately we are a long way from that.

In March of 2010, Google announced the shut down its China operations. Google’s decision stemmed from pressure from the Chinese government to censor search results, a laudable and moral decision in keeping with Google’s moral motto to “Don’t be Evil”. Sometimes however, things aren’t what they seem.

With a quarter of the world Chinese they cannot be ignored as a market. America’s software companies have all pushed hard to enter the Asia market, especially China, and for the most part were initially unopposed. But arrayed against them there have risen indigenous giants who have done battle for the hearts and minds of the Asian peoples and have almost uniformly across the board\(^1\) trounced us. How can companies rise like this to overwhelm western companies, some with decades of feature development in their products? Many of these Asian companies are censored or are otherwise controlled by their governments. Yet competitors with more freedom produced by companies like EBay, MSN, and Google, products that are tried and true, like polished stones that should fit like keys to the needs of their users have been beaten back. The answer is that what matters is which river these stones were polished in.

In the past, companies such as Coca-Cola and McDonalds successfully entered into foreign markets because they had the American “caché”. We no longer carry that heft. Features are important, but those features must fit with the users who use them. They must be familiar and comfortable. We are discussing culture. To ignore it, as Americans are in the habit of, is an arrogance that will be punished in the markets. In this case the market we will look at is China.

The Internet in China is filtered by the “Golden Shield”, which blocks content deemed unsuitable by the Chinese government. In this, Internet traffic within China is a pool unto itself, separate from the world. But within that pool is a free market. Chinese users

\(^1\) Facebook seemed a notable exception, but has recently announced that it is moving to form a joint venture with Baidu to form a social networking site to operate within China in place of their regular site.
are not directed to choose one product over another. In the case of search engines, all are working on an equal footing, suffering under the same content restrictions. So market performance within China is a matter of free market competition, quality of service.

Baidu is an Internet services provider, and in this case most notably a search engine, founded by two partners in 2000. Robin Li, who worked in the US during the 90's helping to develop the online version of the Wall Street Journal and also patented a system of search engine ranking which he later used in Baidu, and Dr. Eric (Yong) Xu, businessman and financier, who currently runs a venture capital fund in China. They met at a company picnic in Silicon Valley in 1996.

In June 2004, Google made its first venture into China investing $10 million in Baidu, giving it a minority interest in the company. Google’s chief competitor Yahoo had already invested in Baidu’s competitor 3721.com. At the time no US internet companies had yet received permission to open branches within China itself, although of course all were visible outside the Golden Shield. Even without a Chinese site, Google held close to 30% of the Chinese market and hoped to improve this by creating a Chinese oriented site, sited within China itself. A presence inside China would give them a place to meet local customers, give them a better sense of the market, and give them easier access to the government. This initial investment gave Google an “in” into China and a chance to learn from Baidu.

In May 2005, Google was given permission to enter China. From the beginning it was with the caveat that Google abide with the rules that governed censorship in China, which it already had been doing quietly in its regular search results for China from its American site. In January 2006, it launched Google.cn. By the time it went online Google’s market share had dropped to 20.4% of the China market. When it finally left in 2009, it had only 12.4%.

2.1 Community Verses Precision

When Google left and moved to Hong Kong, Baidu had 64.5% of the market. Today it’s closer to 75%. What made an upstart like this so popular? Was it just support for the domestic brand, or is there more?

Is there something in the name? It has been pointed out that most of the .com companies that succeeded through the Internet’s explosion of growth in the US had two syllable names, and both Baidu and Google certainly fit that. However “Google”, an English word, is not an easy for the average Chinese to say. Google was named after the mathematical term googol, which is equal to $10^{100}$. Both the number and the name were invented by mathematics Edward Kasner’s 8 year old nephew, a charming mathematical in-joke from the math geek period of the Internet.
The name Baidu has an interesting meaning too:

"Many people have asked about the meaning of our name. 'Baidu' was inspired by a poem written more than 800 years ago during the Song Dynasty. The poem compares the search for a retreating beauty amid chaotic glamour with the search for one's dream while confronted by life's many obstacles. '...hundreds and thousands of times, for her I searched in chaos, suddenly, I turned by chance, to where the lights were waning, and there she stood.' Baidu, whose literal meaning is hundreds of times, represents persistent search for the ideal."²

"The name "Baidu" is a quote from the last line of Xin Qiji's classical poem "Green Jade Table in The Lantern Festival" saying: "Having searched for her hundreds and thousands of times in the crowd, suddenly turning back by chance, I find her there in the dimmest candlelight."³

There is a great distance here that I will discuss later in this paper. Let us just say that poetry carries more weight in China than it does in the US, for practical reasons that involve the complexities of the language, especially in its written form.

Both search engines have their strengths within the market. Let’s start with their basic product: search results. In the beginning before its entry into China, Google was at a great disadvantage because it was English based. In Chinese for instance there are 38 different forms of the word “I”, none of which were accounted for in Google’s US database indices. Google though, knew it had a problem and put together a strong Chinese R&D team to try to quickly address these issues before they formally entered the Chinese market. But even to this day, although Google’s database is certainly more comprehensive than Baidu’s, many feel that Baidu is still better at obscure words and antiquated terms. So Google returns more comprehensive results, but Baidu’s results are more localized to Asia.

Both Google and Baidu divide their databases into functional sections. Google has Web, Images, Videos, and News. Baidu has Map, News, Knows, MP3 Search, Images, Video, Encyclopedia, Web Directory, Patents, Games, Finance, Statistics, Entertainment, Dictionary, Desktop Search, and TV. The Chinese language has difficulties inherent in the categorization and subdivision of written Chinese, which is evident when you look at their dictionaries. If you think about how difficult it must be to create a sensible collation of Chinese text since it has no alphabetical order, then importance of the larger number of categories should become evident. Google's interface is simple and uncluttered, thinking in terms of the rules of interface complexity where the human mind has difficulty dealing with

Figure 4, Baidu's feature menu, translated ironically with Google translate.

---

4 See Figure 4.
more than 7 choices at any one time. The complexities of organizing the Chinese language on paper and now screens have led the Chinese to become accustomed to complex interfaces with many choices. This is a complex subject we will discuss in detail later in the collation section.

The number of categories here are different, but there is another factor at work here. These categories are friendlier. Think about children doing schoolwork, people trying to find their way around cities that grow at a bewildering pace, and businesses forming out of thin air. Google’s site is generalist, while Baidu directly addresses current Chinese concerns. Which site is more helpful?

The most important asset a company can have is its image. Google is popular among the highly educated elite. It’s also popular amongst the technically savvy because the US version can be accessed when users are using common methods to bypass the golden shield. Google returns a mix of Chinese and English results which favors more highly educated users who know English. Google also provides practical research tools such as Google Earth. So Google projects a bright intellectual image.

Something that Baidu does well that Google does not is that it incorporates social networking into its product. Many people in China are leaving the village life that they grew up in and are entering the cities in search of new opportunities. Community has been very important to the Chinese people. Business for most part has always been a face to face thing, but as China has grown and its cities swelled, the opportunities for that direct human contact have dried up. They are discovering the loneliness of the western lifestyle. Baidu offers a large array of social opportunities. Baidu has a product called “Baidu Postbar”, where users can establish theme bars for anything they’re interested in, basically chat rooms. It also has “Baidu Knows” where users can post questions and answers, a feature discontinued on Google. Google at the time had a social networking feature called Orkut, but it was by invitation only. Not open to everyone. Orkut these days is ranked 65th in worldwide in user participation. User participation in Baidu not only builds community, it builds a sense of ownership as users help build the site.

Baidu was first into the market. Google even helped it. This is not to be underestimated. Users develop habits that can be infectious. Google built on this here in the US. The name Google in America is becoming synonymous with searching just as Kleenex has with tissue. When school children in China are collaborating on a school project and ask their friends to Baidu something, they are spreading the brand to their peers and as we have said, “Google” is hard for Chinese to say.

Lastly, supporting western copyright and patents puts Google at a disadvantage. Baidu allows users to share music and videos, which makes its MP3 and video services strong

---

draws for the site as a whole. However, this advantage is mitigated to a certain extent by strong domestic competition.

So Google lost China because it failed to account for local culture and sensibility. China for them represented a shallow technical adjustment to their product, which they made without truly making the effort needed to understand their audience. They created a good product that fit the world as they knew it, but didn’t stop to think that there might be more to the world than the neighborhood they grew up in.

In losing China, Google lost a quarter of the world’s population, the single largest and fastest growing Internet market on Earth. In raw numbers it also gained a strong rival. IV A rival that will continue to grow even without moving into outside markets6, while Google’s home ground in the US and Europe is largely mined out. As if to rub salt in the wound, Baidu is leaking rumors that it will extend its extensive mobile support to include their own operating system in direct competition with Android. To maintain its position in the world against this pressure from Baidu, Google will have to win decisively in the emerging markets of Asia and South and Central America. It can’t afford to make this mistake again.

---

6 Baidu currently only supports Chinese and Japanese.
3.0 Simple Plumbing

How did Google fail in China? Did the problems encountered stem from inadequacies in our programming languages? Clearly not, since Baidu and Google use the same tools. Was it the hands that wielded the tools? Clearly not again, since many if not most of the technicians at Baidu, including its founders were trained here in the US. The problem was clearly us. We need to incorporate the ability to account for culture into our development methodologies. This is more than just using UTF-8 in our databases.

In order to understand what the steps are that need to be taken, we need to understand our customers. A global perspective needs to be incorporated into our technical training. Designers and architects need exposure to world culture and training on the tools required to work with it. This means more than just using the correct date formats and language fonts. To understand, to illustrate this change, we need some examples. Let’s start with language.

20% of the world speaks Chinese in one form or another, 17% speak Arabic, two languages that take careful handling and yet the problems inherent in their handling are not taught in US technical education. Unicode does not completely cover Chinese and many other Asian languages and yet we rarely cover the limitations of Unicode or the reasons for them. And it’s not just the portrayal of languages in text, but their organization as well.

Language is central to culture and is a major dividing factor in humanity. Its expression in the Internet is vital. It can define who is included and who is excluded. Poor language support can lead to the exclusion of cultures and even to their deaths. This may seem an extreme statement, but I hope by the end of this paper that you will see that it is true and already happening.

What is the relationship between Unicode code points and fonts? How do we handle diacritic characters such as the hamza in Arabic and why does Arabic need two parallel fonts? We use UTF-8, but we also use UTF-16 and UTF-32, and still see UTF-7, UTF-4, UTF-2, and UTF-1. Unicode is not yet world encompassing. There are regional encoding schemes still in common use such as JIF, Big5, and Guobaio (also known these days as UTF-GB18030), and to a lesser extent Unix EUC which is still used in some parts of Asia. Why don’t people just use UTF-8? Doesn’t it fit everything? And what about compromises made for mobile devices? These are issues of localization that need context for complete understanding.

To give these issues and others context, you need to start with a bit of history. For example in the case of Unicode we could begin with the reasons why different languages are as they are and why the machinery we use today is as it is. This naturally leads into the vocabulary necessary to understand the later, more complex issues. There are
practical reasons for everything humans do. The places not yet covered by Unicode are
that way for good reasons.

3.1 Unicode

Over the last 70 years software has been in a continual state of improvement and new
standards have come into being which have allowed the inclusion of greater and greater
segments of the world’s population in the computer and communications revolution. One
of the most central and basic parts of this cultural compatibility is the written word. The
Internet has increasingly become the world’s access point for knowledge. The inability
to access the Internet in their native tongue can stunt an entire people’s access to the
outside world, but more important still denies us access to their knowledge and works.
Even ancient languages such as Assyrian and Old Norse provide text that must be at least
catalogued if not eventually encoded in their entirety. The Internet has become the great
repository for the world’s knowledge. Without access, cultures risk marginalization or
even destruction.

It has come to the point where nations are actually rewriting their languages in order to
gain admittance. Without native language support there is risk that members of that
culture will resort to the use of other languages and that members of foreign cultures
capable of consuming that culture’s knowledge and products will instead turn to other
sources, in effect culturally and economically marginalizing them. As yet no system has
been created that is capable of encoding and digitally rendering all the world’s languages.
To understand the difficulty of the problem we must begin by reviewing the efforts to
date to solve it.

3.2 Early History

Before the Internet, back when computers were large and solitary, communication
between people and computers for the most part was accomplished using proprietary
numeric codes, with a number being mapped to each character. The communication
devices that connected to the computer would display the correct character depending on
the number it received. There were and dozens of different character encoding schemes
produced by different companies for different computers. The lack of standards was a
real problem because devices such as terminals and tape drives from one machine
couldn’t be used on another, sometimes not even on different versions of the same
machine. The code didn’t match the desired letters. IBM alone used over nine different
character sets\textsuperscript{8}. In 1963, the American National Standards Institute (ANSI) X3.4
Committee released the ASCII or American Standard Code for Information Interchange
standard. It was adopted uniformly by the American computer industry except for IBM,
which adopted its own proprietary standard EBCDIC. The ASCII standard is still used today, embedded in more comprehensive standards such as UNICODE.

Most computers, with a few exceptions, then and to this day organize their memories around an 8-bit word, or "byte", the smallest section of memory that can be accessed. It can hold the numbers 0 to 255, or 256 digits altogether. ASCII is a 7-bit coding scheme, which is to say that it uses 7 binary digits to encode a single character such as "R" or "%". The number inside the computer is called the code value. The character it represents is called the code point. Seven bits left room for 128 code values that represented 128 code points. Plenty of room for the 26 character alphabet and punctuation needed to express English. ASCII code values have a one to one relationship to their code points. Not all encoding schemes have this relationship.

7-bit ASCII, because it only used 7 of those 8-bits, fit in a byte sized slot in computer memories with an extra bit left over. Why did it only use 7-bits when there were 8 available? Back then we were still using the old American Teletype and paper tape devices left over from the days of telegrams, which only had room to store 7 data bits and one parity bit, a code used to help detect mistakes. With a 7-bit code we could keep using our older equipment and new equipment could still be developed with the assurance that it would connect and operate properly.

Computers in the late 60's all spoke English encoded in either ASCII or EBCDIC. There were numerous efforts overseas to create domestic computer industries, but their efforts at the time were focused mainly on cloning American technology. There was no support for the differences seen in other languages. They spoke English because most of the computers were designed by English speaking people for use within the English speaking world or because American dominance of the computer market forced compatibility with American standards. So a 128 character set was it. And this worked well, until our allies
in Europe started wanting to use computers in their own languages. Businesses wanted computers of their own and wanted to produce output that could be read not just by engineers and scientists, but also by their managers and customers. So how could we handle the umlauts and accents of Europe?

### 3.3 Code Pages

In 1964 IBM began selling its ground breaking IBM System/360 line of computers, which used a new character set called Extended Binary Coded Decimal Interchange Code, or EBCDIC. EBCDIC introduced a new concept called code pages. The characters represented by different numbers in the computer's memory could be mapped to different symbols depending on a number, the code page number. So you could have a German, Greek, or even Russian mapping of code points to code values. In German code page 500 for instance, the number 95 maps to "p." In Hebrew code page 424, the number 95 maps to "ס." Unfortunately this system was not just proprietary, but aggressively proprietary. IBM excluded character mappings necessary for competing operating systems and languages so that they could not be used with IBM equipment.

EBCDIC was brilliant, but excluded the rest of computer industry. What to do? The answer was easy. Simply extend ASCII using that 8\textsuperscript{th} unused bit doubling the space to 256 characters. Leave the first 128 characters the same as 7-bit ASCII and fill the upper 128 spots in with the new characters needed to meet market demands. The old teletypes could ignore the 8\textsuperscript{th} bit and still print out their regular English output and new devices, which could handle the 8\textsuperscript{th} bit, could print out their umlauts and accents. There was so much room that there was space for common monetary symbols, math symbols, and even Greek letters. The new solution was called Extended ASCII and it was free and open to everyone. So Europe was covered and it seemed like we had a perfect solution that everybody could use.

Scientists and engineers can sometimes compromise when it comes to language, until the end of the 19\textsuperscript{th} century for instance, the universal language of science was German, but business is universal to all of humanity and they must to talk to their customers who cannot. You can't produce bills in a foreign language and expect them to be paid. And businesses all over the world, not just Europe, wanted computers.

Proprietary variations of ASCII began popping using the upper 128 characters to represent different specialized symbol sets. The arrival of the microcomputer in the early 70's accelerated this since they often contained their own input and output devices which didn't have to be compatible with other computers, so they could use any character encoding they wanted to. Commodore PETSCII, Atari ATASCII, and Serbian Galaksija 2716 were popular examples of ASCII variants. Clearly there needed to be a single
standard that could be extended to cover the world’s language needs, something that covered more than just Europe.

Unfortunately there were many standards, each major player in the computer industry had its own, often with conflicting numbering. All of them used a similar scheme, borrowed from EBCIDIC, to solve the problem – code pages.

The upper 128 characters, noted by that 8th bit, became a numbered page mapping code values of the upper 128 characters to different code points depending on the user’s need. It could be full of Cyrillic or Greek characters, or even mathematical symbols. As long as you had a code page number and font that could display the correct symbols you were fine. You mapped those upper 128 character codes to whatever symbols you wanted and as long as you owned both ends of the equation, the computer and the terminal, you were fine.

Code pages became possible because the terminal had become a computer itself, the old dedicated CRT terminals had been replaced by the personal computer. In older PC’s and terminals characters and symbols were built into the terminal hardware. If you wanted Cyrillic, you had to build a terminal to express it. But with more graphic oriented computers fonts became dynamic images that could be changed at will. The means of expression had moved from hardware to software. New fonts could be loaded dynamically to support new code pages. So if you wanted to show a document in Arabic, you would select an Arabic code page and your 256 character codes would be mapped to 256 Arabic characters that could be displayed on the page.

It should be noted that later on, code page came to mean any encoding in or out of the 8-bit boundary, the page numbering system being drafted as a way to identify different character encodings. So in Windows for instance, Unicode can be referred to as either code page 1200 or 1201 depending on the orientation of the bits, even though it really isn’t a code page. This can be very confusing when reading contemporary documentation.

It could be thought then, that things had been finally solved, at least for us in the west. After all, if you had to express a new language, you could just create a new code page and tie it to a new font. Unfortunately, although 128 character code pages are fine for German or French which only have a few extra characters in their alphabet, or even Cyrillic which has a clear alphabet, they were not fine Japanese and Chinese that had tens of thousands of commonly used characters. And then there are also problems with just what the definition of a character is.
3.4 The Breadth of Written Language

Take the case of the Arabic alphabet. There are issues involving the position of the symbol on the page, context within the phrase, it is read right to left, it has optional symbols, it is cursive which is to say that the letters are connected together like English handwriting, and the character set has been used to express more than 25 languages in addition to Arabic. Typeset Arabic letters come in four different forms, Initial, Medial, Final, and Isolated, the choice depending on the next character. It's important to remember that different computers have different abilities with some being able to handle right to left sequencing and others not. Arabic text for this reason is considered to be "bidirectional" and the font must support cursive in both directions. Some letter groupings are written as ligatures, special shapes created by combining letters into a single character. From a technical standpoint this would be multiple code values, or numbers within memory, combined into a single glyph or image on the screen. It also has a diacritic symbol called hamza, which is added to other characters to denote a glottal stop. With multiple code values possibly mapping to multiple code points, the standard 128 character code page begins to look rather cramped and complicated.

To make things worse, we have to decide just what constitutes Arabic. It has many variations. In Arabia itself there are the literary Classical Arabic, Modern Standard Arabic, East Arabic, and West Arabic, some with regional variations. The differences between Arabic varieties are such that one speaker sometimes cannot understand another.

This is a huge problem in Internationization, the forcing of a single standard on multiple cultures, and the first point I want to make with all this dusty old history. Computers are pervasive in all levels of human society. Since so much of our communication is handled
by computers, the creation of a code page can be considered a de facto standard for written communication for an entire nation or people. Their printed word is channeled through that standard. Limitations within that standard can endanger regional variations, which may wither and die simply due to lack of support. It is simply not enough to just say you have an Arabic code page.

Where the imposition of a standard can have great effect on a culture, the lack of a standard can be a problem too. Without an international standard, cultures are forced to create their own local regional standards that can prevent the creation of a common global standard. Japan, by the mid 70's was a growing technical powerhouse, limited by a complex language system with four major systems of writing and many sub systems and regional variations. These systems were often used in mixed contexts, even within the same sentence! Most of these systems used complex ideographic symbol sets that outstripped the capacity of the code page system. Worse, the codepage system was not capable at that time of switching between codepages within the same document, so mixed encoding was difficult to accomplish.

Japan was expanding at a furious clip and could not wait for a solution from western computer makers. In 1978 the Japanese Ministry of International Trade and Industry established the JIS C 6228-1978 Code of Japanese Graphic Character Set for Information Interchange standard, or “Old JIS” as it is often called today. The original JIS standard was a 14 bit code, stored within 2 bytes. It encompassed the Latin, Greek, and Cyrillic alphabets, 83 symbols to cover Hiragana, 86 symbols to cover Katakana, and 6349 commonly used Kanji characters.

Thanks to strongly focused national effort, including large subsidies and tariffs, Japan had the hardware to take advantage of the standard, producing at the time a series of IBM System 360 clones. Direct access to their operating systems and peripheral hardware allowed them to implement the standard and make it ubiquitous within the Japanese marketplace. It’s still ubiquitous today in the form of Shift-JIS, a variant created by Microsoft and ASCII Corporation of Japan, currently seen in most Japanese language web traffic. It’s in direct competition with the current international standard Unicode, preventing its full implementation within Japan.

Encoding standards can become centers of ideological and cultural imperialism. In 1981, Mainland China’s Standardization Administration of China (SAC) created its Guobiao (GB) standard, currently used in most Mainland Chinese Internet traffic. In an effort to prevent its being rendered obsolete by Unicode, the current version, GB18030, has

---

7 Kanji, Hiragana, Katakana, and Rōmaji
8 To name some: Hentaigana, used in ceremonial documents; Man'yōgana, used mostly in place names; Gyaru-moji, a modern evolving form used in informal situations, often in texting.
9 This replaced the code page standard JIS X 201, supported by PC DOS.
incorporated all the code points in Unicode, essentially absorbing Unicode. Unicode in response declared it to be part of Unicode, since it had become in compliance with Unicode standards – UTF-GB18030.\textsuperscript{ix}

There are two other important regional standards in Asia. In 1985 Taiwan, Hong Kong, and Macau set their “Big5” encoding standard.\textsuperscript{x} Another important player in Asia that should be noted is EUC, or Extended UNIX Code, used mainly in Linux systems. It is backwards compatible with ASCII and JIS and comes in Japanese, Chinese, and Korean variants. It often appears in web pages served from Linux servers in Asia. These encoding systems are supported by western versions of our major operating systems, but often only through the installation of additional conversion software. All of these standards are in direct competition with Unicode and are in common usage in the Internet.

3.5 Unicode

It turned out to be a wise decision that Asia didn’t wait for American technology because it wasn’t until the end of the 80’s, 10 years later, that there was any movement towards serious internationalization with the Unicode standard. It was the late arrival of Unicode that has limited its acceptance in Asia and allowed other encoding schemes to become entrenched.

Unicode is going to feature prominently throughout the rest of the paper and it’s important that the reader understand a little of its inner workings. If you are a Unicode guru, if you understand its shortcomings, then you can skip straight to 3.8. But knowledge of Unicode will be important when we discuss the impact of collation on culture.

We begin with a little history. In 1988 Joe Becker from Xerox and Mark Davis from Apple published a proposal for an “international/multilingual text character encoding system, tentatively called Unicode.” Soon they had formed a working group and people from all over the computer industry were joining. By 1991 the first volume of the standard was published and today we use it just about everywhere, sometimes even in Asia.\textsuperscript{xi} It’s a large and growing work that is far from finished.

Unicode 1.0, now UCS-2\textsuperscript{10}, was a big improvement over the existing western standards of its day. It covered many heretofore uncovered languages such as Devanagari, Coptic Greek, and Hangul. It also combined Chinese, Japanese, and Korean, often abbreviated CJK, scripts into a unified character set. This eliminated the duplication of the shared symbols between the languages allowing much more complete coverage, cataloging 20,902 CJK symbols, far more than previous system coverage to date combined.

\textsuperscript{10} Universal Multiple Octet Coded Character Set, 2 bytes.
Cataloging and combining these systems was an amazing work. Problems involving the collation and organization of CJK ideograms are very difficult. It also allowed for expansion, leaving a large portion of the possible character space unassigned. Even though this original version is now officially obsolete, there is a sizable body of text encoded in it and legacy support for it in modern operating systems.

What eventually made it obsolete was its limitation to 16 bits. What had seemed a limitless space when compared to the world’s languages, with room for 65,535 characters, was only a drop in the bucket. Han Chinese has more than a hundred thousand characters in it alone. The world in 90’s was changing rapidly. Cheap computers and the Internet were infiltrating every corner of the globe and literally everybody wanted to participate. Most of all, they wanted to participate in their own languages. All of the world’s character encoding standards had to find a way to break out of the limitations of a 16 bit boundary and expand if they were to survive in the Internet age.

In 1990, recognizing the need for growth, the International Organization for Standardization, or more commonly ISO, set out to create a set of universal standards that allowed for expansion beyond 16 bits – ISO 10646. It consisted of three different protocols:

- **UTC-4**: A fixed 32 bit wide scheme creating a huge point space allowing the encoding of all characters.
- **UTC-2**: A 2 byte scheme in direct competition with Unicode 1.0.
- **UTC-1**: A variable length coding scheme with codes being 1 to 5 bytes, designed for character streaming applications.

Just to be clear, this is the UTC standard, not UTF Unicode. Many influential software companies balked at the idea of the size and complexity of the standard and fought to delay it. They weren’t ready for four byte encoding, UTC-1 had technical problems that made it impractical to use, and UTC-2 was really no better than Unicode 1.0. In the mean time Unicode 2.0 was released in 1993 which extended the original specification to allow for optional 32 bit character encodings. ISO realized they could no longer support their bid and instead sought to unify it with the Unicode standard. The now obsolete Unicode 1.0 became UTC-2. UTC-4 was incorporated into the Unicode standard as UTF-32 and UTC-1, which as a flawed design that never really worked, was eventually replaced by the popular UTF-8. By the end of 1996 Unicode looked like this:

- **UTF-7**: An unpopular 7-bit encoding, now considered obsolete.

---

11 Chapter 4
12 Windows up to Windows 2000 and Java up to J2SE 5.0 support only UCS-2. Python uses UCS-2 internally and decodes it to UTF-16.
UTF-8  A very popular 8-bit variable-width (1 to 4 bytes) encoding.
UTF-16  A 16-bit variable-width (2 or 4 bytes) encoding.
UTF-32  A fixed 32-bit encoding, gaining in popularity for use within key fields in database collations and operating systems.

Today there is also the Chinese UTF-GB18030 and UTF-EBCIDIC for compatibility with the old EBCDIC standard.

### 3.6 A Few More Dreary Details

In Unicode the code value is called a code unit sequence. Code unit sequences are encoded using one of the three current encoding specifications or schemes, UTF-8, UTF-16, or UTF-32. The choice of scheme can be dependent on a number of factors. First—the language or languages being encoded. UTF-8 encodes English very well since its 8-bit form is standard extended ASCII. UTF-16 encodes CJK languages very well since it’s rooted in the Unicode 1.0 standard which encoded the most common ideograms within its 16-bit boundary. You are more likely to see UTF-16 in use in Asia. Second—the purpose of the encoding. UTF-32 is useful for the encoding of text that will need to be converted to other encoding schemes, since its fixed width format lends itself to simple character mapping algorithms and it works better with fast sorting algorithms. UTF-8 is potentially useful in the case of network traffic, especially in PDA device applications such as cell phones, where storage space is always a consideration. It potentially can use the least amount of space depending on the language. Third—compatibility. UTF-8 is byte oriented and includes extended ASCII in its true 8-bit form and is 100% compatible with older systems. These are all important factors to remember when designing international applications and databases. xii It’s not enough just to say we should use UTF-8.

<table>
<thead>
<tr>
<th>Code Unit Sequence</th>
<th>Code Point</th>
<th>Glyph</th>
</tr>
</thead>
<tbody>
<tr>
<td>UTF-8</td>
<td>41</td>
<td>U+0041 A</td>
</tr>
<tr>
<td>UTF-16</td>
<td>0041</td>
<td>U+0041</td>
</tr>
<tr>
<td>UTF-32</td>
<td>00000041</td>
<td></td>
</tr>
<tr>
<td>UTF-8</td>
<td>DB 80</td>
<td>U+0430 a (Russian)</td>
</tr>
<tr>
<td>UTF-16</td>
<td>0430</td>
<td>U+0430</td>
</tr>
<tr>
<td>UTF-32</td>
<td>00000430</td>
<td></td>
</tr>
<tr>
<td>UTF-8</td>
<td>F0 90 8C 82</td>
<td>U+10302</td>
</tr>
<tr>
<td>UTF-16</td>
<td>D800 DF02</td>
<td></td>
</tr>
<tr>
<td>UTF-32</td>
<td>00010302</td>
<td></td>
</tr>
</tbody>
</table>

Figure 9, examples of Code Unit Sequences
Code points are abstractions of a Glyph, the actual representation of the code point. The glyph can have many attributes that deal with its portrayal such as serifs, italics, or bold, but code points are not fonts. What’s more, what the user sees as a single glyph may be composed of multiple code points. As well, a single code sequence may invoke multiple code points.

Unicode stores attributes that describe properties of its code points called semantics. Semantics are very helpful when we are rendering compound pictographic characters composed of several glyphs. They are not pictures or bitmaps. Semantics help in character rendering, collation, dynamic composition, and text direction both internally and when portrayed. You could consider them to be hints about how the character is to be portrayed. Semantics are especially important in the portrayal of composite glyphs which occur in CJK type languages.

<table>
<thead>
<tr>
<th>Character</th>
<th>Code Points</th>
<th>Linguistic Usage</th>
</tr>
</thead>
<tbody>
<tr>
<td>ch</td>
<td>0063 0068</td>
<td>Slovak, traditional Spanish</td>
</tr>
<tr>
<td>th</td>
<td>0074 02B0</td>
<td>Native American languages</td>
</tr>
<tr>
<td>x</td>
<td>0078 0323</td>
<td></td>
</tr>
<tr>
<td>å</td>
<td>019B 0313</td>
<td></td>
</tr>
<tr>
<td>ã</td>
<td>00E1 0328</td>
<td>Lithuanian</td>
</tr>
<tr>
<td>i</td>
<td>0069 0307 0301</td>
<td></td>
</tr>
<tr>
<td>ト</td>
<td>30C8 309A</td>
<td>Ainu (in kana transcription)</td>
</tr>
</tbody>
</table>

Figure 10: User perceived characters with multiple code points

Figure 11, code points verses glyphs
You can have multiple code values mapping to multiple code points mapping to multiple glyphs. Remember the Arabic hemza? Diacritics are just one way this can occur. Ligatures are another way.

All of this is really only the coarsest of summaries. The actual Unicode specification comes in a surprisingly readable book form, with sage advice on how to deal with problems with interpretation and mechanics of encoding, problems encountered with different languages, and conversion between different encoding schemes.

3.7 Fonts

In early computers, the question of font was a mechanical one. Devices generally had only a single font, the one they were built with. Characters were lined up in neat rows, each in its own uniformly fixed size box or character cell. This suited English fairly well, but in other languages with non-Latin character sets characters had to be shoehorned in to fit.

In Asia, special half width fonts had to be designed in order to fit their glyphs into western character cell technology. You can still encounter these fonts today. They were revived for early cell phone technology. It was the arrival of windowed operating systems and dot matrix printing technology that brought graphical capability to computers. It allowed us to include graphical elements in our documents, such as pictures, proportional spacing or characters of varying width, and different sorts of typefaces.

Our first fonts where simple bitmaps, pictures of each letter, that were strung together into words and paragraphs. They could be scaled, made larger and smaller, with mixed success. They usually contained several different versions of the same character, such as italic, bold, and intermediate sizes, and they mapped to ASCII in a simple one to one manner. But bitmap fonts, although still in use today where speed or complex graphics are needed, fell short for two reasons. The first was scaling. Computer screens are divided into a grid of dots called pixels. In scaling bitmaps, making them larger or smaller, the pictures don’t always fit exactly within the pixel boundaries which made edges look blocky and irregular.

To solve this problem a new type of font specification called outline or sometimes vector fonts was created. The first outline font system was developed by Adobe Systems in 1985. It was called the Postscript Type 1 Font Format. Its adoption by Apple, helped make the Mackintosh computer the choice for graphic design applications. Outline fonts use a series of lines and curves to define a glyph, which can be mathematically scaled to
any size. After the lines are defined, then the center is filled in to complete the glyph. It only takes a few points to define a line or a known type of curve so outline fonts take up much less space. Font modifications such as italics and boldness can be simulated mathematically as well.

But there was a second problem with bitmap fonts. The problem was culture. Once we left the 256 character boundary of code pages we encountered a problem of scale. Languages having thousands of glyphs required tens or even hundreds of thousands of bitmap images to support a single font. Not only did all those bitmap images have to be stored on your computer, each possible image of all the possible ways the glyph could be displayed, such as italics or bold, had to be drawn as well. Outline fonts helped here, but didn't entirely solve this problem either, especially with CJK glyphs.

It takes a lot of curves just to define a single brushstroke and CJK glyphs can contain quite a few brushstrokes. The problem was especially acute in handheld devices where storage space can be limited. Enter stroke-based fonts, developed primarily by Agfa Monotype and Type Solutions. Glyphs are composed of a series of vertices or alternatively, as a composite of the TrueType stroke glyphs. You could think that composing Chinese as composites of individual strokes might save a great deal of space, but it has proven difficult to consistently

Figure 14, CJK strokes defined in Unicode which can be used to portray compound CJK glyphs.
accomplish, because the symbols are not as consistently drawn as they might seem. In
addition, these compound font systems require special software in order to use them.
Compound glyphs like this show up commonly as names and in the portrayal of obsolete
words and older text.

Font files contain a table that maps code points to glyphs. It’s impossible for any font to
map all the code points in Unicode directly to particular glyphs, because in languages
such as Han or Arabic a code point can map to different glyphs depending on the dialect.
In Western languages font coverage is not an issue. However, in many Eastern languages
this can be an issue. Common words will probably be covered by a viewer’s font, but
technical jargon, place names, or other obscure words may not be. It is sometimes better
to use phonetic or anglicized writing systems such as Romanji or Pinyan, which we will
discuss later, for these words unless you know the viewer’s computer will support them.
Of course, this is fine if you are just trying to convey information, but if you are trying to
sell something you definitely don’t want to look ignorant. People will laugh at you in
just the same way we laugh at their attempts to communicate.\(^\text{13}\)

Of course too, we’ve just been discussing basic glyphs. The Internet isn’t composed of
just modern text. Glyphs come in many styles. It’s not uncommon to come across fonts
for free in CJK that are designed for specific purposes, such as titles and banners, which
may only contain a few hundred specialized glyphs. Take a minute to review the
coverage of the font before you use it.

3.8 Internationalized Domain Names

With all modern operating systems being able to support it, Unicode is quickly replacing
ASCII in all aspects of the Internet. We don’t generally think of objects such as
filenames as being capable of being named in non-roman characters, but they can be, in
any language and any character set you choose to use. However, the reason we haven’t
seen them generally in the US is that in order for them to be requested across the Internet,
the Internet itself has to recognize the filename in terms of a URL and we built the
domain name system to our standards.

\(^\text{13}\) Take a look at the humor site http://www.english.com
However, starting in May of 2011 the first top level domain (.asia) began recognizing domain names in Chinese, Korean, and Japanese. Soon, all the world’s languages will be supported. In order for this to be able to occur, older ASCII based DNS servers must still be supported. This is accomplished by converting the Unicode characters into ASCII codes called Punycode, using two algorithms called `toUnicode` and `toAscii`. The actual algorithms are somewhat arcane, but to summarize: URL’s are processed one section at a time, a hyphen is appended to the end of the word, non-ASCII Unicode characters are stripped out, converted to character code and appended to the end, and an “xn--” is added to the front to denote that the word contains Unicode. For instance, “bùcher” would translate as “xn--bcher-kva”\(^\text{14}\). The resulting string is pure ASCII and leads to a unique DNS entry.

This may seem like it might pose a problem with existing domain names containing hyphens, which are allowed. In DNS hyphens are converted to spaces prior to lookup so if you wanted to register as “xn--bcher-kva.com”, you would look like “xn bcher kva.com” in the database. One big problem not yet solved in this scheme is spoofing. A Russian “н” looks exactly like an English “n” in most fonts, but could take you to very a different site.

### 3.9 Conclusion

Unicode is fast becoming the world standard and it is hoped that it will pull the world together simplifying communication, but in many sections of the world, such as China, Japan, and Taiwan, national standards and tradition still prevail. Unicode has fallen short

---

in many areas, which include problems with collation of different languages that will be covered in the next section. For instance, no character encoding scheme yet includes the ability to use color directly in characters, which could be helpful in encoding some pictographic languages such as Egyptian or Mayan. Do not make the mistake that ancient texts will not be available over the Internet nor that people will not want to research and quote them. And Unicode has not kept up with the dynamic daily expansion of CJK characters. Stop and think about it for a second. When we find or create a new thing, we make up a word for it. In English this is just a new combination of existing characters. In pictographic languages though, this often means the creation of a completely new character. In Mandarin for instance, a new character is added every week. How does this impact Unicode?

As great swaths of the world are added to the Internet, more and more cultures are exposed to computers. More and more different kinds of written text are becoming available. It is essential in the 21st century for the ultimate survival of any language that it be capable of digitization, so that its works can be joined to what is fast becoming the great world repository of information, the Internet. And it is important for us as 21st century engineers to foster and maintain a global view. A culture that can interface with many is far stronger than one that must wait for others to come to it.

Training Engineers to deal with these problems will lead to solutions and success for businesses within the Internet. What version of UTF should be used for CJK languages? Why is UTF-32 better for fast indexes? Why shouldn’t you use that fancy banner font for Japanese? How does the inclusion of a hemza change the value of a search field in an Arabic index? This last question involves finding order within language and leads us to the next section.
4.0 Typing Chinese

Understanding the nature of character data is only the first step. A third of the world uses character data that is incompatible with the data entry systems that we designed, and yet they manage to get along. For instance, not only can the Chinese text on their phones, it takes them less keystrokes per word than it does us. How this is accomplished is not only useful in terms of internationalization, but makes use of novel design patterns that have uses throughout the tech industry and leads us into ways in which language can be collated. How do people outside the US input data into our applications?

4.1 Finding an Order

You don’t necessarily have to use single keystrokes to invoke a particular glyph. You can use the software in the computer as an aid, trying to find natural orders within the language to break the character set apart so you can look up required symbols. Chinese is a mix of phonetic and ideographic symbols and presents many problems in data entry. These are problems which the Chinese have been struggling with for thousands of years. Chinese is spoken by over one fifth of the world’s population with the Mandarin dialect being the most spoken language in the world. There are 13 major dialects, some of which are actually considered to be sub-languages and are unintelligible to other Chinese speakers. The venerable Kangxi Dictionary recognizes over 47,000 characters, with one or two new ones being created each week. How did the Chinese fit all this into the QWERTY keyboard?

Unlike America and Europe, China doesn’t have a long history of typewriter use to draw on when designing their keyboards. Building a Chinese typewriter was a difficult problem that was constantly set back by war and political upheaval. The evolution of Chinese typing took a very different path from ours. The first problem they had was in finding a way to order their symbols.

Moveable type was invented in China around 1040 CE, the first being made out of carved wood. It was later perfected by the Koreans with cast metal type. With so many characters in their language, the filing of type when not in use became an immediate

Figure 16, a revolving table like type case for individual movable-type characters arranged primarily by a rhyming scheme, 1313 CE.

---

15 Mandarin, Wu, Cantonese, Min, Hakka, and Gan are the top 5.
16 Kangxi Dictionary, the standard dictionary for Chinese in the 18th and 19th centuries.
problem. Some type sets numbered as high as 60,000 pieces\textsuperscript{17}. The first ordering of the
type pieces, so they could be found when needed was accomplished by a system of
rhyming:

“For each character there were several types, and for certain common characters there
were twenty or more types each, in order to be prepared for the repetition of
characters on the same page. When the characters were not in use he had them
arranged with paper labels, one label for each rhyme-group, and kept them in wooden
cases.”\textsuperscript{xiii}

Rhyming is a method, one of three, used in the ordering of Chinese characters in Chinese
dictionaries. The first surviving “rhyme dictionary” is the Qieyun\textsuperscript{18}, published 601 CE.
In rhyme or rime dictionaries, characters are first divided into four groups according to
“tone names”: level, rising, dark, and low. These categories are divided into sub-groups
according to their rhymes or tones. In the first early dictionaries they were mostly
intended for the composition of poetry, but they worked as traditional dictionaries as well.

Another system of ordering, used in what were probably the first dictionaries ever created,
was first published starting in the 3\textsuperscript{rd} century BCE. These were specialized dictionaries
that ordered words on particular subjects, such as plants. They ordered words into
semantic categories, such as grasses, brush, or trees, much like our plant recognition
books do today. In this case, it was necessary to have some knowledge of the subject
being discussed in order to understand the semantic divisions being used in order to reach
the explanation of the symbol in question.

4.2 Radicals and Strokes

The last method, appearing around the 2\textsuperscript{nd} century CE, used a graphically organized
system of “radicals” and “strokes.” Radicals in English are the roots of words, after the
prefixes and suffixes are removed. These are not to be confused with sememes, which
we will discuss later. In Chinese, symbols are often composed of lesser elements. For
instance 米 cāi ‘to pick, pluck’ is composed of two elements, hand 手 (zhāo or zhua) and
tree 木 (mù). Neither of these characters is an actual root in the English sense, but many
Chinese characters break down into simpler terms. Because of their limitations of space,
the width of the character, and the limitations of design elements, the brush strokes,
Chinese characters can also be characterized by design elements as well. For instance 木
jué or juē in 了 liǎo. Jué is not a real word, but a design element, a visual root. Together

\textsuperscript{17} The “Records of Jingde County”, a 100 copy edition contains more than 60,000 characters.

\textsuperscript{18} Cutting (spelling) Rhymes
these basic words and design elements comprise a set of characters collectively considered to be radicals in Chinese. They are used as section headers in dictionaries and generally number about 200. They are the first division of the Chinese character set.

Strokes are just what they sound like, brush strokes. There are a set number of brush movements that can be used in the creation of a Chinese character. In mathematical terms, you could think of these as eigenvalues, the most basic common components of a set of objects that can be used to create all of the objects in the set. Just about all Chinese characters are composed of these basic brush strokes. Stroke type and stroke count, were and still are used to further subdivide the dictionary. Combined with radicals they form an important system of collation that is used as well as the basis for several systems of character input.

4.3 Four Corners 四角號碼検字法

An important method of organizing Chinese came about in the 1920’s, The Four Corner Method. It was developed to help telegraph operators look up codes for characters in their code tables. A number is used to “encode” the shape of each of the four corners of the Chinese symbol in order to produce a four digit number, which can then be used as an index into the telegraph encoding table.

| 1 | Horizontal stroke               |
| 2 | Vertical or diagonal stroke     |
| 3 | Dot stroke                      |
| 4 | Crossed strokes                 |
| 5 | Three or more intersecting strokes |
| 6 | A box                           |
| 7 | A corner stroke                 |
| 8 | 八 or 八 inverted               |
| 9 | 小 or 小 inverted               |
| 0 | Empty corner                    |

This system has been adapted for many uses including dictionaries, library catalogs, and telephone books. It isn’t precise enough to be used for the language as a whole, but is useful in returning a quick subset of characters from a limited list that can then be chosen from to arrive at the desired input. It’s not in common usage except in older indexes and modern reprints of older books. I mention it to note the practical possibilities of this design pattern and for historical context.

4.4 Wubi 五筆字型輸入

The Wubizixing Input Method, or Wubi for short, is a method of keyboard input that has some similarity to four corners in that it encodes visual elements of the character into four codes, letters instead of digits. Instead of four fixed points, encoding progresses in the order in which strokes are applied when writing the character. If there are more than
Figure 18, the Wubi keyboard layout.

Wubi is a popular method for touch typists as it doesn’t require selection from lists on the screen, and if you can remember how a character is supposed to be written then it can be very fast. Unfortunately, thanks to computer assistance, many Chinese are forgetting how to brush characters and there are too many strokes to fit on a cell phone keypad. Another similar system is Boshiamy, which uses about 300 radicals.

4.5 Cangjie 倉頡輸入法

The most popular input method for personal computers, although rapidly being replaced by Pinyin, is Cangjie. It was first developed for the Apple II system. It uses elements similar to the Chinese dictionary system, but reduced and tailored to fit common keyboards.

Cangjie starts with a set of 24 non-standard radicals mapped to the normal Latin alphabetic keys. X is mapped to the difficult character key. To help the user remember the radicals, they are divided into four arbitrary groups: philosophical, stroke, body parts, and character shapes groups. Radicals are selected with the keyboard and assembled on the screen in an order defined by a series of rules such as: left to right, top to bottom, or outside to inside. There are some characters that do not fit the decomposition rules and punctuation marks that must be input using fixed codes. There is another less popular system Dayi, which uses 46 radicals rather than 24.
Cangjie is difficult to learn, but once mastered can usually produce typing speeds of 25 characters per minute. Experts can produce as high as 60. It has the drawback though of requiring a QWERTY keyboard which means it cannot be used with cell phones. There is a more automated version called Simplified Cangjie, which only requires the user to enter the first and last radical code. The user is presented with a list of possible characters to choose from. Although easier to use, it is difficult to reach high typing rates.

4.6 Bopomofo 注音符號

Bopomofo is very popular input method used in Taiwan and like Cangjie, it often comes standard on Chinese versions of operating systems. It is one of two popular phonetically based alphabets, the other being Pinyin, which can be used to describe Chinese characters by how they sound when spoken in Mandarin. It uses 37 characters and 4 tone marks, a just small enough character set that it can be used on telephone keypads.

Bopomofo was developed in the 1920’s by the Republic of China as an aid for showing pronunciation in dictionaries, text books for children, and books for foreigners. The Bopomofo character set is based on syllable onsets and rimes rather than consonants and vowels and is similar to the system used in early rime dictionaries.

4.7 Pinyin 拼音輸入法

Pinyin is a Romanization system developed by the People’s Republic of China in the early 1950’s for use in computers and phones and is the most common text input system in Mainland China. It replaced older Romanization systems such as Wade-Giles and superseded Bopomofo on the mainland. Its function and purpose is similar to Bopomofo in use in books and street signs, except that it uses Latin characters instead of stroke based characters.
There are two, maybe three common forms of Romanized Chinese, Hanyu Pinyin in mainland China, Tongyong Pinyin in Taiwan, and Zhuyin Fuhao. Zhuyin Fuhao is an older system developed at the start of the 20th century and has great inertia within the Chinese language, with forests of felled trees devoted to existing text. Hanyu was developed in the 50’s and Tongyong in the 90’s. Tongyong was an attempt to differentiate Taiwan from the mainland, but was officially dropped in deference to Hanyu in 2008. There is also Wade-Giles, a system developed by Westerners in the 19th century that appears in street signage and older texts. Hanyu is the version of Pinyin commonly used in computer input.\textsuperscript{xiv}

Using pinyin is fairly simple, as long as you, most importantly, know Mandarin. You begin spelling out the word using Western characters. If the input system you’re using has any intelligence to it, it will begin displaying lists of probable words. As you enter more letters the list narrows until you see the word you want. In some cases it may even suggest phrases. This feature makes it very popular on cell phones and PDA’s as it greatly reduces typing. It helps too that Pinyin input is ubiquitous on just about every device and operating system produced throughout the world.

Unfortunately there is no standard method associated with Pinyin input, with great variance and debate existing between different systems. With more capacity, most modern devices offer a choice of several different input systems in order to cater to user preferences. This allows room for regional versions that cater to specific dialects, such as Cantonese.

The factors that seem to distinguish different systems involve the positioning of spaces, whether between syllables, words, or phrases, the treatment of two non-English Latin characters, ü and ê, capitalization, predictive/auto complete capability, and opinions on usage statistics. Pinyin is based on Mandarin pronunciation and this can create problems in the numerous sub-dialects of Chinese where pronunciation differs from Mandarin. And there is the treatment of other languages embedded in text, especially English, which is usually accomplished by spelling the foreign word in all capitals, something to think about when collating strings of Chinese words.

There are Pinyin input services which you may want to investigate, that can be embedded for free in your web pages:

Languages with an alphabet were at an advantage in the development and dissemination of typewriter technology, allowing the transition to computer keyboards to be fairly straightforward. Phonetic and ideographic languages however were at a disadvantage when computer technology arrived. It was this technology though that allowed them to overcome their disadvantage through the adaptation of existing writing systems to computer input. It's an evolutionary process that is still ongoing.

These new systems of input are based on divide and conquer solutions, built around common elements within the languages themselves, such as visual elements in their writing or phonetic elements in their speech. They sometimes use an intermediate form of writing such as Bopomofo or Pinyin that requires fewer total numbers of keys, used in combination in compound expressions, to identify single glyphs. With these systems they can reach input rates that are higher than their traditional Latin based counterparts.

The use of pinyin in computers and most importantly cell phones is a driving force in making Mandarin the dominant force in the Chinese language. We support this by adding pinyin as standard equipment in all of our operating systems. By catering to the larger market, we are inadvertently becoming involved in a war of cultural imperialism. That is the social element, the implications to cultures our decisions make. The information we choose to store and not store can have a huge impact on societies around the world. Not servicing a language or culture can threaten its very existence. And those societies will fight back. And so we see in our standards cultural conflict playing out. Politics, conquest, market pressures, all this just to input data into our web pages.
5.0 Ordering Things

The ordering of information, its collation, is very important in databases and of course databases are the center of our applications. Look for the Three of Diamonds in a shuffled card deck and you’ll have to look through the entire deck to find it. Look for it in a deck sorted by number and then alphabetically by suit, and you will know that it’s just this side of deck center. That labeling and ordering of the cards is a system of ordering called a collation. Collations are not just how we file and find things, but how we make sense of things as well and when it comes to the Internet, where we have trillions of things to be filed and found, it’s the foundation of ours and every other culture’s intellectual future.

For the last few decades, cultures have been scrambling to establish and standardize systems of collation for their languages. It’s not just so they can find their documents in Yahoo or Google. We saw some of this in the last section where we were organizing language input with the four corners system, radicals and strokes, and rhymes. Collation is important wherever there is data in any form. It’s finding products on shopping sites, hotels in tourist destinations, and movie show times.

There is far more to this than just the putting things in alphabetical order. It’s the way we label things as well. When you have a picture of blue flowers, is it “flowers”, is it “blue”, or is it the “meadow” where they were growing? It’s the question of how we describe the Three of Diamonds as a card as well as its relation to the rest of the deck. These are tricky questions even in Latin based languages, but how are they managing in Myanmar and Laos where their written languages have no word boundaries, or with Urdu, 70 million speakers who until recently had no clear system of alphabetization?

5.1 Collation Defined

Collation is the ordering of written information into a standard order. The most common form of collation in the west is alphabetization. Most alphabetic languages have a set alphabetic order although sometimes it can seem to be a bit arbitrary. For instance some, like German which has phone book and dictionary order, have more than one. French uses the same alphabetic order as English; however some letters have extra marks such as é, à, and ô, but are not considered distinct. In Swedish they are just the opposite, quite distinct and are tacked on after “z”. Sometimes groups of letters are counted as single letters as in “sch” in German, which comes just after “s”. They are, as long as they are consistent, all valid systems of collation.

One of the most basic ways of collating things is numeric ordering where lists of numbers such as \{1, 23, -3, 10\} collate into ordered sets like \{-3, 1, 10, 23\} based on some rule, in this case something like \(X = (x, y | x \leq y)\). The numbers can represent anything. In the
U.S. for instance, we have social security numbers that represent US workers. Cars have license numbers. ASCII is a collation sequence for English letters and symbols and has been used itself, usually with some modification to adjust for capitalization, as a collation sequence in ordering databases. Unicode is an even bigger enumeration of human written symbols that could be used. Although assigning things random numbers does constitute a collation, it isn’t a very useful one. It’s better if the numbers reflect some aspect of the data that you’re interested in, such as the ranking of sports teams.

Collation is not to be confused with classification or categorization, which is the grouping of things into categories. Collation is only concerned with the ordering of things, although that ordering can be used to create categories. For instance the Dewey Decimal System used in libraries is a categorization. It divides book collections into categories. It does not innumerate individual books. Collations allow you to search for things and define ranges of things, such as the first five elements of . . . , the largest element in . . . , or the elements between A and B.

In ordering things each element of the set has to compare itself to the set as a whole to find its particular place within the set. The important thing to realize is that it’s the comparison function itself that defines the order. For instance, in numerical ordering each element looks at its neighbors and says “am I bigger than the guy to my left?” If it is, then it swaps places. If they all keep this up, eventually each element will reach a point of equilibrium where none of the elements can rise further. The set will be in ascending order. Change the function, for instance reverse the direction of the comparison, and you will change, after a great deal of jostling as each element climbs downward, the order of the set to descending order.

There are some rules that a function must satisfy in order for the resulting set to be considered a valid collation. In mathematics a collation is considered to be a total order. In order to qualify as a total order it must satisfy some conditions. First, the ordering rule must apply to all elements in the set, it must be mathematically total. If you skip something, then it cannot be considered to be part of the resulting collation. For instance you could create a collation of mammals from the set of animals. The collation is only orders the set of mammals. From the point of view of the collation, it’s as if the other animals don’t exist.

The ordering rule must consistent or mathematically transitive. If A comes before B and B comes before C, A must therefore come before C. The rules can’t change somewhere in the middle of the set. Lastly, the most abstract property it must have is antisymmetry. Antisymmetry means that the ordering rules must apply to A and B or just to A itself, but it can’t apply to B and A. What this means is that the order can only go in one direction, it can’t go backwards. 1 before 2 and 2 before 3, but never 3 before 2 or 1. Satisfy these conditions and you have a certified bona fide collation.
5.2 Apples and Apples

When we are setting up a collation sequence, we must think about the different aspects of objects being collated. What do they have in common? There are an infinite number of ways any object can be described, but if we're going to create a collation we must find a type of description that is common to all objects within that sequence. Our choices of aspects of data that we use to build collations are based on the relationship between the information we have at the start of our queries and what we will want when we use our collation. We have a name, we want a phone number. We have a title, we want a book location. These relationships are obviously linear, but sometimes the linearity of the relationship isn't so obvious.

Search engines build their collations based on keywords supplied by web documents or derived from their titles or content. These words are treated as individual atoms of information and are taken together to form a document key, a signature that identifies that document. For instance an auto repair shop might have a set of keywords like: auto, car, repair, lube, tire, oil, etc. It's their association to particular documents that is the key that identifies the auto shop's web page. Having one or more matching keywords in the search is the function that builds the initial collation. The results though are returned sorted by the computed weight, relevance, or rank of the relationship. It's how well the keywords being searched for fit with the keywords associated with the document. Did you match one, two, or more keywords? We have two collation sequences at work here, the first producing the data for the second.

5.3 Apples and Oranges

Anyone who uses the web encounters keyword searches almost every day. They've become the ubiquitous method of finding things on the web. But sometimes the data we have isn't exactly compatible with the data we are searching through. Can we enter “apple” into Baidu, a Chinese search engine, and get meaningful results? The Chinese word for apple is 食果. Are they the same? The simple answer is currently no. 食果 and “apple” are different sets of code points. Chinese speakers and English speakers both visualize the same thing when they see their written words,
but to a computer they are just strings of code values and are completely different. For a
database to service both languages, it would have to store both words. For a database to
service the entire world’s languages it would have to store thousands of versions of each
word.

This is a huge problem. Currently we avoid it by dividing up the world by major
languages using different sites with different databases so we don’t have to store ten
versions of each word. For instance there are English, French, German, Japanese, and
Chinese Amazon.com sites, the dominant versions of the languages that are spoken in
the areas of the world that Amazon has chosen to focus its marketing. But this is hardly
comprehensive. By various counts, there are at least 165 different spoken languages in
current use within the United States alone. In Papua New Guinea there are 820 And
what does this do to less spoken languages and regional dialects when the only way they
can participate in the Internet is to adopt one of the dominant languages?

So we segregate the world’s data into separate sites or limit the content of our sites to a
limited word set that we can keep multiple translations of. But how can this technique
cover the thousands of languages humans speak or store the thousands of versions of each
word or phrase and still deliver proper service to each? To be blunt it doesn’t. There is
no current solution to this problem. However, where there are problems there is also
opportunity.

Where the big companies try to service as many as they can, often poorly, regional
providers can target the specific needs of specific locales. Localization is the counterpart
of internationalization, the focus on the individual rather than blanketing the whole.
There are, for instance, more than 40 Arabic search engines based in more than 14
countries each catering to regional needs not serviced by the global search engines. As
you should know at this point, Arabic is not a homogenous language but has diverse
regional dialects that are very different from one another. Although these search engines
are Arabic, they service distinct populations.

The more local you become the more specific your key terms can be in your collations
and the more usable your database can be. On the other hand, you may have fewer
people who can use it. Local search engines may lead you to local delicacies, but then
not all your customers may understand that they found them when they see them. In
some ways this is sad, because the world is a far richer place than we are presently
capable of seeing, all because we don’t know how to find it.

19 Amazon.com, Amazon.fr, and Amazon.de
20 Top 20 countries by number of languages. Vistawide World Languages and Cultures.
http://www.vistawide.com/languages/20_countries_most_languages.htm
This problem has been mitigated to a small extent by the marriage of translation programs and search engines. Google's translation services claims to be able to translate 135 languages, including Klingon. If you've used these services though, you'll know that they are not very accurate, and on occasion completely unreadable. The problem is that different languages do not have a one to one relationship. Words in one language are missing in others or have synonyms with subtle differences. Euphemisms often based on jokes or shared experiences often cannot be translated to other cultures without a great deal of thought and explanation.

Problems of localization become especially acute in regions like Europe, where there are many cultures close together without a single dominant language. Bureaucrats in the European Union sometimes encounter great difficulty in obtaining translations of important documents in their preferred language. Human translation is time consuming and expensive. Machine translation is a seductive alternative and research is always ongoing.²²

However, in the case of collations, exact translation may not always be necessary. A "lossy" query may be acceptable. Keyword description of information is not very accurate to begin with. Think of the last novel you read and then try to think of all the keywords someone might use to look for it. It's almost certain that you won't think of them all. Libraries focus their queries by limiting the indexes that you can search against to certain types of words such as author or title. General purpose search engines though, do not have that luxury.

What if you could create a kind of a common key, one not tied to any particular language? You could translate your search terms to that intermediate language before the search is performed and then you could create a non-language specific search. The problem is that most people want documents to be in their own language. Sites in arbitrary languages are often not always very useful, so language itself must be an important part of the search itself.

5.4 Unicode Collation Algorithm

It should be obvious now that there is more to sorting things than just putting them in alphabetical order. Ligatures and compound characters are composed of multiple code values that count as single characters so you can't sort by just code value. Multiple characters, such as "ch" in Spanish, count as single characters in sorting so you can't just sort by how the word looks on the screen. The value of case varies from language to language and, as we will see in the next section, the order of characters within a word can alter their collation sequence. And clearly the most difficult of all, you may need

²² MOLTO, Multilingual Online Translation. http://www.molto-project.eu/
collations based on mixed languages where you must reconcile two or more sets of language rules. Clearly we cannot, for instance, rely on the ordering of characters within Unicode to sort our data.

The Unicode collation algorithm is being developed to address this problem. The algorithm is as complex as the problem it seeks to solve. It breaks the assessment of characters or character groups down into weighted “layers”, generally at least three deep, using language specific information from the Unicode Common Locale Data Repository or CLDR. CLDR entries are written in Local Data Markup Language and are loaded by the operating system depending on languages selected. Characters are given a computed weight based on data from the table, which can be used in sorting. In cases where multiple languages are involved, because it is their weights that will be sorted instead of the characters themselves, the data will be interleaved independent of the language.

For example, Unicode defines a Default collation ordering the, Default Unicode Collation Element Table (DUCET)\(^23\), however this is not presently sufficient to address the needs of the European Union and the European Free Trade Association, and their needs are immediate and acute. In response they are developing a more advanced rule standard, the European Ordering Rules (EOR)\(^24\). It has four layers. The first, which takes precedence, orders the basic letters:

- **Latin:** 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 Ѳ
- **Greek:** α β γ δ ε ζ η θ ι κ λ μ ν ξ ο π ρ σ τ υ χ ψ ω Ѳ
- **Cyrillic:** а å ä é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ š ñ ç é ø ë æ ш Ѳ

Layer two deals with diacritics:

- å â ä å â š â å ô å ö å ŋ å ŋ å ŋ å ŋ å ŋ å ŋ å ŋ å ŋ

The others are ordered as variants of their base letters. Level three deals with the treatment of capitalization. Level four concerns punctuation and the treatment of whitespace. What you should see here is that the standard DUCET is still not adequate for all situations.

UCA only works with Unicode data. It doesn’t handle Braille for instance. Some databases will incur conversion costs from other character encoding schemes; others will

\(^{23}\) ISO/IEC 14651:2007

\(^{24}\) EN 13710
just throw errors if non-Unicode data is used. What is important is that the encoding format of the data must be considered when selecting a collation scheme. Check the specifics of the collation scheme, what algorithm it uses, what language rules it supports, and does it support mixed languages, before selecting it.

5.5 Cultural Standardization

Up until now we’ve talked about standards expanding and evolving in an attempt to fit the world. This is only possible up to a point after which the cultures themselves must make an effort to fit in or be left behind. Up until the end of the last millennium Urdu, one of the two official languages of Pakistan and one of the 22 official languages of India, had no set dictionary order. There were at least eight, with many Urdu dictionaries being hand copied or even typewritten. Urdu typewriters and printing have been common for over a century and several Urdu keyboard standards have been in place for over a decade, but the effort to set a collation order, so important for information storage and retrieval, didn’t start until the turn of the millennium.

Part of the problem with any standard is that, like most major languages, Urdu itself has many variants. These are caused by geographic divisions or mixing with other local languages. While most Urdu speakers live in Pakistan or India, Urdu is spoken widely in the Middle East, Europe, and North America. Written Urdu uses Arabic characters in its alphabet, but is linguistically closely related to Persian and Hindustani. It’s because of the geographic spread of the language, with no true central homeland, that it has been difficult to create a single standard. But with the new millennium, Pakistan has stepped forward to bring Urdu to the Internet.

Starting in 1996, a national effort was initiated in Pakistan to implement a set of national coding standards for computing in Urdu, including collation. In 2001 the Urdu Zabta Takhti standard for character encoding and collation was adopted. UZT was based on a 256 character code page at a time when code pages were well on their way towards being replaced by Unicode. As such, the first efforts lacked the depth necessary to become a comprehensive standard. And collation in Urdu is highly complex and a simplistic
numerical ordering was clearly inadequate. Work has continued on a replacement standard.\textsuperscript{xv}

Urdu has many features that make it difficult to sort. It has many diacritics, such as the hemza from Arabic, which must be dealt with in the same manner as accents and umlauts are in European languages. However there is still debate as to which characters are letters or diacritics. In some cases this may even depend on their placement within the word. Dealing with the many diacritics requires Urdu to use a two pass sort. The first pass sorts the root characters, the second identifies and sorts the diacritics. Because Urdu uses the Arabic alphabet, Urdu code points are included in the Arabic section of Unicode. There is Urdu collation support in the CLDR and the Unicode Collation Algorithm does currently work, but debate continues about the details of the rules and continues to be updated to this day.\textsuperscript{xvi}

And now here is the point. These decisions about collation standards were made in Pakistan and accepted by the Unicode consortium. But only 14% of Urdu speakers live there. Pakistani Urdu contains many elements of Pashto, Punjabi, and Sindhi that other speakers of Urdu find foreign. Despite this, Urdu speakers in the future will be seeing this new sequence in their dictionaries and search engine results without any question of consent being raised. The ultimate effect will be to make Urdu and the perhaps the people who speak it as well more homogeneous, but at a loss of regional dialect.\textsuperscript{xvii}

5.6 What is a Word?

There are conventions unconsciously adopted by the Western creators of the Internet which do not apply the world over. Most of the alphabetic languages of world use a space to denote the boundary between words. The ability to isolate and use individual words is for most Westerners an unconscious assumption. The design of our databases is founded on it. However, the Myanmar Language, spoken by over 32 million people and formerly known as Burmese, does not separate their words. To be clear, you may see spaces sometimes in modern text because some authors will insert them as a convenience; it is often in a haphazard manner. Burmese does not naturally use whitespace, which poses a big problem in a word-oriented Internet. In older text you will not see spaces at all.\textsuperscript{xviii}

Burmese is classified as a mono-syllabic or isolating language and does not yet have a system of Romanization. An isolating language has a low morpheme, the smallest component in a word, to word ratio. Burmese words tend to be composed of single morphemes. An English example might be “boy”, which is composed of a single morpheme “boy”, a 1:1 morpheme ratio. The Russian word “protivopravitelstvennyi” consists of six morphemes, a 6:1 ratio.\textsuperscript{xix}
When the Four Eight Uprising occurred, approximately three thousand people died.²⁵

Figure 25, Burmese sentence structure

To make things complicated, it’s not a pure isolating language. It does have compound words:

\[
\text{head} \begin{array}{c}
\text{head}\end{array} + \text{pack} \begin{array}{c}
\text{head}\end{array} = \text{hat} \begin{array}{c}
\text{head}\end{array}
\]

\[
\text{language} \begin{array}{c}
\text{language}\end{array} + \text{look, see} \begin{array}{c}
\text{language}\end{array} + \text{building} \begin{array}{c}
\text{language}\end{array} = \text{library} \begin{array}{c}
\text{language}\end{array}
\]

These make it difficult for a programmatic parser to differentiate between the words “head” and “pack” or “library” because without word boundaries it could be read either way. The meaning is implied by the context. It also has loan words, foreign words which are spelled out phonetically:

\[
\text{computer} \begin{array}{c}
\text{computer}\end{array}
\]

\[
\text{sub-committee} \begin{array}{c}
\text{sub-committee}\end{array}
\]

Interestingly, Burmese doesn’t have adjectives. Instead it has verbs that carry the meaning “to be X”, where X is an English adjective. These verbs can modify a noun by means of the grammatical particles \begin{array}{c}
\text{particle}\end{array} or \begin{array}{c}
\text{particle}\end{array} in literary Burmese. They take the form be + verb + \begin{array}{c}
\text{particle}\end{array} + noun.²⁶

The Myanmar Language, according to one dictionary, has over 32,283 different cataloged syllables and it is around these that efforts to create a collation sequence have progressed. The language has excellent alphabetic ordering, so if it can be broken into atomic syllables, then these could be ordered and would make a reliable key for collation. However, this is a problem that has not yet been satisfactorily solved. There are several approaches being researched. The most of them rely on the fact that syllables in the Myanmar Language follow a certain pattern, generally diagrammed like this: C(G)V((V)C). “C” is a consonant, each syllable starting with an onset consonant. “G” is
a glide or semi-vowel. They’re like vowels, only shorter. “V” is a rhyme, the last part of the word that we usually latch on to when we rhyme words. Burmese also can in some cases have an extra rhyme and certain particular consonants tacked onto the back of the syllable.

There are three classes of approaches to slicing sentences apart: pattern or rule based, syllable matching, and neural net algorithms. All of these approaches have so far achieved levels of correctness above 95%. One complicating factor is that the Burmese character set is still in a state of flux with 9 new characters, mostly punctuation, being added in 2006.

At best, systems like this can only generalize. There are always exceptions in any language and misspellings pose problems. Wouldn’t you hate it if your business name happened to have one of the syllables that couldn’t be parsed and no one could find you on the web? One of the important metrics used in judging these algorithms is how well they recover from unintelligible, misspelled, slang, new and made up words, all precisely the kinds of things marketing people love to add to every language.

Luckily, most form input, predominantly used in the web, is self segregating, using different input fields for different information such as city and name. In this type of situation being an isolating language has an advantage. You actually get to remove all whitespace, erasing all arbitrary boundaries that might have been inserted by the author.
and simply deal with the text fragment as a single block. The problems start when you must deal with streams of text, such as in translation and spell checking.

5.0 Conclusion

Having a definite collation order for your language is necessary for its inclusion in the databases that drive the Internet. Accomplishing this is not always straightforward, as we have seen both with CJK languages and Burmese. Many languages, such as Urdu which is widely spoken throughout the world, did not have a collation order before the Internet arrived.

The requirement of collation for entry to the Internet can be seen as the setting of a world standard for language if they are to be included in any media beyond printed books. Use of electronic media is replacing older print media throughout the world. Telephone books are being replaced by smartphones. If a language is not included and easily available in electronic media, then the users of that language will use the nearest dominant language instead, to the detriment of their native language. Lack of a collation order can mean that the literary works, crafts, and products of a culture will not be as easy to include in Internet indexes as well. They risk being sidelined.

Information engineers are working feverishly in many of these cultures to find ways to standardize their languages so they can be included. Sometimes outside agencies such as the Unicode Consortium and Microsoft will step forward champion a standard for a particular language or set of languages, but not always with success as we saw with the European Union who had to create their own better standard.

It's still difficult to collate multiple languages in a single index, although the Unicode Collation Algorithm is a promising solution. Because of this, we segregate languages in our databases. This means that generally only a few of the world's dominant languages are supported, as we saw with Wikipedia and Amazon.com. This can be viewed as a sort of cultural imperialism, especially in cases where it goes hand in hand with other forms of cultural repression, such as in China.

The use of language can also limit sales, limiting access to subgroups who do not speak one of the dominant languages. But this can also be an opportunity to smaller regional vendors, who service these cultural regions. These smaller vendors though, can limit our access, because it's less likely that we will speak their language and we will not have access to their products.

The finding of uncatalogued and unenumerated sections of the world is the work of the entrepreneur and the bureaucrat, but the clever art of finding that point of commonality in any set of data in order to build a collation and then expressing that data in a usable form.
is the work of the information engineers as they work to piece the disparate parts of the world together. We are the foot soldiers in the creation of global unifying standard.
6.0 Universal Meanings

Earlier we discussed how 蘋果 and Apple are not the same even though to their readers they invoke the same image. Would it be nice if they could be linked to the same meaning? What if we could identify their common aspects? This is a hot subject in computer science these days and when it is discussed in schools, it is usually only discussed in terms of local culture. This robs the subject of its usefulness. It’s the cross cultural application of this research that makes it important. Relating apple to red or green is not as important as relating 蘋果 and apple. It’s the bridging of cultural divides that make this subject potentially profitable. It’s why Europe is investing so many Euros in research in this area.

6.1 Microformats

The collation of documents, especially in the Internet, use keywords provided in document metadata or derived from document content. These provide only a single one-dimensional collation, a general description of content. Sometimes though, documents are more than just a single thing. Often they are containers, composed of many elements that are themselves worthy of collation.

In traditional web documents, information is portrayed as pictures and text. There is no actual meaning in the data itself. It’s up to the viewer to decide if what they’re looking at is an address or a laundry list. You couldn’t expect to find bread recipes buried in blog posts because the blog site only provides generic blog site oriented keywords to the search engine, not blog post content information. The blog posts themselves could be almost anything.

However, if you could identify your posts as recipes or movie reviews in a systematic way, then the search engine could catalog them. Those blog posts would carry meaning as well as content. Then the search engine could create collations that just deal with recipes, news, resumes, or even human relationships. These collations would be smaller and faster, and could be kept in separate machines spreading the workload out, increasing reliability and response time. And they would make available information normally buried. Best of all, this context would be culturally abstract. An address would be an address, regardless of where it was.

This is giving the information we see a context, a place within the larger scheme. XHTML and HTML, the language used to define web pages, allows for the embedding of metadata within their tags. Metadata is data that describes data. Microformats take
advantage of this to insert their descriptions of the data being presented. For instance, contact information might be presented like this in a normal web page:\footnote{26 From Wikipedia, Microformat. http://en.wikipedia.org/wiki/Microformat}

```html
<div>
  <div class="fn">Joe Doe</div>
  <div class="org">The Example Company</div>
  <div class="tel">604-555-1234</div>
  <a href="http:/example.com/">http:/example.com/</a>
</div>
```

With microformat markup this becomes:

```html
<div class="vcard">
  <div class="fn">Joe Doe</div>
  <div class="org">The Example Company</div>
  <div class="tel">604-555-1234</div>
  <a class="uri" href="http://example.com/">http://example.com/</a>
</div>
```

The browser can identify this information as an address which can be selected and added to the user's address book should they so choose.

At present there are a sizable number of microformats defined, however browser plugins are required to recognize them. Microformats are not a standard but are instead defined by a community of users led by Commerce.net, a non-profit organization, and they are not the only effort working along these lines. There is ContextObjects in Spans, or COinS and the Microdata proposal in HTML 5. What is important here is that there are major efforts in the works to identify and group data within documents in a standardized way. By classifying data we no longer have to rely on arbitrary keywords dependent on the nationality, spelling ability, and literacy, of the authors. If we can link documents and the objects within them to some abstract meaning, then we may be able to begin to create collations that are based on relationships between them rather than just arbitrary subjective labels. Of course the rub here is that it's up the authors to give their data meaning.

### 6.2 Ontologies

In the beginning of the Web, Web 1.0, we saw documents described in HTML. They were static, connected together in a web using links which, when selected took you from page to page. The next generation web, our current generation, saw the introduction of
user participation to the web. The web page became an interface with which the user could interact. Social-networking sites, blogs, video sharing sites, mashups, and hosted services are all examples of Web 2.0 sites. But what will the Web 3.0 look like? Many feel it will be the “Semantic Web”.

In microformats we saw information being grouped and identified, tied to common elements of abstract meaning. The vCard class for instance groups information such as city, name, and phone number together as an address. A city, in addition to being a city, has the additional property of potentially being part of an address, a “has a” relationship. The Symantec Web takes this a step further. It supplements the simple markup text with descriptive data stored within the document in its own separate markup languages called RDF and OWL.\(^\text{27}\) A document may contain more descriptive data content than it does flat content. It’s from this descriptive data that collations are formed.

Elements of knowledge, like the city in an address or the address itself, are called “atoms”. They’re kernels of fact called classes, entities, or concepts, the term depending on the background of who you are reading. Relating these atoms to each other is “fact”. “The sky is blue”. Sky and blue are atoms and “is-a” is a class of relationship. This is called predicate logic. You can ask the data questions by stating a fact, but replacing one of the atoms with a variable such as “the sky is X”, which will return everything with the atom sky and an is-a relationship, including “the sky is blue”.

Because different relationships can connect to the same atoms, “the sky is blue” and “my VW is blue”, they form a web – a “universe of discourse” or “ontology”. Does the computer know what these relationships and atoms mean? No, it just keeps lists of them. Any meaning they have comes from human readable text, “definition” or “annotation”. Those definitions are stored in the text of the document and in the text of documents out on the web.

All the descriptive markup code in the document is there to identify the atoms and relationships within the document and to relate them to atoms defined outside on the web. Atoms out on the web are identified using a URI, or Uniform Resource Identifier. URL’s are subset of URI’s. So “blue” might have its own definition out on the web and when you talk about blue, you will be relating your knowledge to all the other discourse on blue. For instance, the definition for integers is at http://www.w3.org/2001/XMLSchema#integer and there is a definition for “iron” is at http://www.daml.org/2003/01/periodictable/PeriodicTable#Fe. If you’re into philosophy, you might say they are a bit like Platonic forms. Documents themselves can become definitions themselves as new documents link to them and then other documents link to

\(^{27}\) Resource Description Framework and Web Ontology Language. OWL instead of WOL for ascetic reasons and because Owl in Winnie the Poo spell his name “Wol”.

47
those. This forms a web of references that define documents in terms of the environment around them.

In current semantic web implementations these relationships have to be flat lists or hierarchical tree structures, otherwise you can have circular paths which can cause problems with search functions. In the real world though, without some yet to be discovered very strict control, circular paths are inevitable and natural. For instance you might have relationships like “blue cars”, “pictures of cars”, “pictures with blue”. A query like “I want pictures of blue cars” could cause a logic loop and a Star Trek style explosion!

You might notice as well that the expression “pictures of cars” is in English. All current ontology mechanisms, like web 2.0 keywords, are language centric. They rely on keywords in a particular language. There can be French ontologies, Russian ontologies, and Chinese ontologies. Other than creating a more complicated way of keeping keywords, have we really improved anything if we can’t relate them? Isn’t blue the same as blau or 藍 or اللون الأزرق؟ And even if we both speak English, how can we know that the blue I’m referring to is the same as yours?

6.3 Natural Language

The semantic web is still in a very primitive state. Ontologies are discrete from one another, a bit more like individual XML databases. OWL isn’t really any better than the language Prolog, invented 40 years ago. Current implementations don’t use natural language, although they intend to eventually. They use complex query languages like SPARQL, a query language for ontologies, which is based on predicate logic with queries that look like:

```sql
PREFIX table: <http://www.daml.org/2003/01/periodictable/PeriodicTable#> SELECT ?name FROM <http://www.daml.org/2003/01/periodictable/PeriodicTable.owl> WHERE {
  ?element table:name ?name;
  table:atomicWeight ?weight.
} ORDER BY DESC(?weight)
LIMIT 10
OFFSET 10
```

You can import ontologies into ontologies to create huge compound ontologies, but this is not conducive to quick searches.
This is not a query method that can be commonly used. But, one of the promises of semantic web is the ability to use natural language, or at least something like natural language, to pose questions, as you would your local librarian, and expect back lucid answers.

In order to accomplish this, ultimately ontologies will have to be connected together into webs with references that branch outside themselves. Instead of just saying “blue”, you might reference another blue somewhere else on the web, a blue that’s like the blue you’re describing. It might be a color swatch of a standard blue in a color library, a picture of a field of blue flowers on a clear day, or a description of a “blue moon,” which isn’t really blue at all. This is where the “semantic” in the semantic web comes from. And blau or اللون الزرقاء can all refer to common references because in the mind of their authors, they all mean the same thing and thus we can relate them to each other through their common references.

One approach being tried in order to find commonality in expression is to break sentences into simple logical fragments. Consider the sentence below:

“Yesterday I saw Beowulf. I liked it very much. It’s a fantasy movie with lots of actions and good special effects.”

By following a few rules, such as “and” results in several statements and “much”, “good” translate as subclasses, this could be written in a pidgin form:

“Yesterday I see Beowulf. I like it. It is fantasy movie. It has much action. It has good special effects.”

Although somewhat ugly, this form is still readable by humans and could translate into most languages. It’s also more capable of being processed by computer. Of course words like “a few rules” only work for this one sentence. In the face of human language
the task of creating a comprehensive library of parsing rules has so far proven to be difficult. And they wouldn’t be the same for any two languages. Remember the lack of adjectives in Burmese with verbs being used as adjectives? Even identifying the individual words, not just parsing them into simple sentences that can be evaluated has proven to be very difficult. And then consider the ambiguities posed by the words themselves. “Action” for instance can mean many things. What did the author mean by it? Each human word is a blur of linked meanings.

Despite the difficulty, the promise of creating cross language collations that can account for the meaning of documents as well as the normal arbitrary keywords is important, especially in areas of the world where large numbers of languages overlap. The European Union is the biggest investor in semantic web research. They spent over €300 million in 2010 for semantic web research and are expected to top €700 million in 2014. This is not just for development of the web itself, but the all important applications to exploit it as well.

Creating a document in the semantic web will involve explaining it in terms of its place within the web. You will have to relate it to similar items in order to explain its purpose and meaning. This could double or triple the cost of creating a properly indexed document. But this is the cost of creating the precise meanings that can help us cross national boundaries.

6.4 Conclusion

In chapter 5 we discussed the ordering of language. In chapter 6 we discussed the ordering of meaning. Different languages express similar ideas in different ways. Only by relating these similar meanings will people of different cultures be able to use the same index to find the same information. If this can be accomplished, then language will no longer divide us. You will be able to use any device using the language you’re comfortable with.

The semantic web is a promising possible solution to this problem, but it’s still in a very primitive state. Indexing documents is difficult as well, since creators will have to index their documents in many more ways than they presently do, linking them to all possible related atoms of meaning.
7.0 Quantifying Culture

What kind of solutions can we develop to help us deal with culture? Can we reduce it to just another variable in any world solution? Right now any company these days wanting to expand into a different country has to hire people within that country to advise them, as Google did, people familiar with both the target culture and yours. But what if you wanted to expand worldwide? Hiring someone in every country just isn’t practical. This is why Amazon.com for instance, only has a presence in six countries.

It would be nice if it were possible to quantify and type culture in a general global way. Perhaps then you could alter the output of your web applications depending on the audience. Then, perhaps, everyone would feel comfortable. Maybe it could work like the preferences system in Netflix. To do it though, you would need a baseline start from.

Enter *Cultural Dimensions Theory*, first proposed by Geert Hofstede in 1980. Hofstede was a mechanical engineer who went back to school to complete his Ph.D, graduating in social psychology cum laude in 1967. After graduation he joined IBM as a management trainer and later as manager of personnel research. At IBM he was able to compile a huge database of employee opinion surveys, over 117,000 of them, from over 70 countries. In mining the data he began to notice statistically significant patterns between attitudes and national identity. He later began testing people outside IBM and found correlation, finally publishing his work as a book. Since then, six further cross-national studies have been conducted, the last in 2010 by the World Values Survey (WVS) using Hofstede’s principles, which included 93 countries.

What they found was that culture and values could be quantized across at first four and then later six dimensions:

- **Power Distance** (equality versus inequality). The level of acceptance of unequal power distribution. These scores tend to be high in Latin and Asian cultures and low in Anglo and Germanic cultures.
- **Individualism** (versus collectivism). The value of personal achievement over group achievement. North America tends to score high, Japan low. However Japan has been rising quickly.
- **Uncertainty Avoidance** (versus tolerance). The extent to which society attempts to cope with anxiety by minimizing uncertainty. This tends to be high in Eastern Europe and low in China.
- **Masculinity** (versus femininity). The distribution of emotional roles between genders. High in Japan, low in Nordic countries.
- **Long Term Orientation** (temporal orientation). The tendency towards persistence and saving. High in Asia, low in Anglo countries.
• **Indulgence** (versus restraint). The allowance of hedonistic behaviors. High in Latin America, parts of Africa, the Anglo world and Nordic Europe. Low in East Asia, Eastern Europe, and the Muslim world.

The 2010 WVS Study focused on two larger macro dimensions: Traditional values versus secular-rational values and survival values versus self-expression values. Traditional versus secular represents the difference between religious and secular societies. Survival versus self-expression represents the difference between economic and physical security vs. long term management and tolerance, such as ecological management and gay marriage. The questionnaire used in these surveys was designed by social scientists from all over the world and has been in use, evolving for over 30 years.\(^{xxvii}\)

A few of the results:

- Societies become more liberal as they become more secure, the largest increase in security occurs with the transition from agrarian to industrial
- Over the last 30 years results have been converging towards a more homogeneous world
- The greatest increase in personal power and self expression occurs in the transition from industrial to knowledge based society most often expressed in Northern Europe
- Individual social values are dwarfed by a factor of five to ten by societal values
- Emancipation and empowerment occurs on three levels: *socio-economic*, which increases with resources, *socio-cultural*, socio-economic increase increases aspirations towards freedom, *legal-institutional*, democratic rights and a sense of entitlement to exercise them
- These empowerment factors can spiral both upwards and downwards and feed back on themselves giving momentum in one direction or another
- Societal groups enter modernity through different paths which can affect their sense of security and individual empowerment

These are metrics which we can use to quantitatively generalize and gauge our approach to different regions. How we apply them really depends on the product and your company’s expectations.

### 7.1 Usability

Massively Multiplayer Online Games or MMOG’s have become popular the world over. These are digital universes in which players move about and interact within through
**avatars.** These universes generally mimic ours, having gravity, ground, and light that mimic the physical laws that govern our universe. The avatars tend to be humanoid, giving the player the ability to project themselves into the storyline of the universe. Players are given tasks to perform which often draw the player further and further into the universe.

The largest MMOG by far is the sword and sorcery style game World of Warcraft, but there are dozens of other active universes as well. Another popular universe is Second Life, in which their world is entirely created by the users themselves. These universes can have their own economies, social hierarchies such as nations and guilds, wars and even diseases. This makes them popular mediums of social study and a logical place for the application of Cultural Dimensions Theory, CDT.

In order to apply CDT, we need a world-spanning multinational user base. Large games like World of Warcraft segregate their user base by region to ease language problems and server lag due to distance, but smaller games can’t afford to do this, choosing instead to let users solve their own communication problems. One such game is EVE, a space based universe. EVE’s parent company, Crowd Control Productions, CCP Games, is based in Iceland with ties to Europe giving it a strong mix of nationalities; which makes it a great place to perform experiments using CDT.

How do different nationalities approach the game’s user interface? Is there a difference in the problems that they encounter? If there are, how can we explain them and predict future ones? Just such a study was performed by Greek researchers. In this study, EVE players were given two questionnaires: a background questionnaire asking questions about the user such as age, gender, nationality, and gamer experience, and an EVE interface usability questionnaire assessing navigation, learnability, accessibility, consistency, visual design, quality of help, and many other factors. These questions were all derived from similar studies involving the usability of games. The original questionnaire had over 70 questions, so an initial mini-study was performed to assess the usability of the questionnaire itself resulting in its reduction to 57 questions.

A total of 307 gamers from 19 countries responded, the bulk of which were from Netherlands, Canada, Greece, Sweden, and France, however there were respondents from as far away as Hong Kong, Singapore, Argentina, and the United Arab Emirates. The authors of the study used Hofstede’s original four cultural vectors: power distance index (PDI), individualism collectivism index (ICI), masculinity/femininity index (MFI), and uncertainty avoidance index (UAI).

<table>
<thead>
<tr>
<th>Respondent Country</th>
<th>PDI</th>
<th>ICI</th>
<th>MFI</th>
<th>UAI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Argentina</td>
<td>49</td>
<td>46</td>
<td>56</td>
<td>86</td>
</tr>
<tr>
<td>Australia</td>
<td>36</td>
<td>90</td>
<td>61</td>
<td>51</td>
</tr>
<tr>
<td>Austria</td>
<td>11</td>
<td>55</td>
<td>79</td>
<td>70</td>
</tr>
</tbody>
</table>
Questions were grouped in 14 groups by focus of usability such as learnability, consistency, and accessibility, etc. Data analysis was performed in two phases: a straight correlation between the four cultural variables and question area groupings, and then grouping respondents into low and high cultural variable groups for each question and then looking for differences between the corresponding high/low groups.

The first phase of analysis revealed several statistically significant correlations, some of which seem obvious and some which are difficult to interpret on their own.

### Figure 29, cultural variables per respondent country along with Hofstede’s values for each country

<table>
<thead>
<tr>
<th>Country</th>
<th>Power Distance</th>
<th>Individualism/Collectivism</th>
<th>Masculinity/Femininity</th>
<th>Uncertainty Avoidance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Canada</td>
<td>39</td>
<td>80</td>
<td>52</td>
<td>48</td>
</tr>
<tr>
<td>Denmark</td>
<td>18</td>
<td>74</td>
<td>16</td>
<td>23</td>
</tr>
<tr>
<td>Finland</td>
<td>33</td>
<td>63</td>
<td>26</td>
<td>59</td>
</tr>
<tr>
<td>France</td>
<td>68</td>
<td>71</td>
<td>43</td>
<td>86</td>
</tr>
<tr>
<td>Greece</td>
<td>60</td>
<td>35</td>
<td>57</td>
<td>112</td>
</tr>
<tr>
<td>Hong Kong</td>
<td>68</td>
<td>25</td>
<td>57</td>
<td>29</td>
</tr>
<tr>
<td>Japan</td>
<td>54</td>
<td>46</td>
<td>95</td>
<td>92</td>
</tr>
<tr>
<td>Netherlands</td>
<td>38</td>
<td>80</td>
<td>14</td>
<td>53</td>
</tr>
<tr>
<td>New Zealand</td>
<td>22</td>
<td>79</td>
<td>58</td>
<td>49</td>
</tr>
<tr>
<td>Norway</td>
<td>31</td>
<td>69</td>
<td>8</td>
<td>50</td>
</tr>
<tr>
<td>Portugal</td>
<td>63</td>
<td>27</td>
<td>31</td>
<td>104</td>
</tr>
<tr>
<td>Singapore</td>
<td>74</td>
<td>20</td>
<td>48</td>
<td>8</td>
</tr>
<tr>
<td>Sweden</td>
<td>31</td>
<td>71</td>
<td>5</td>
<td>29</td>
</tr>
<tr>
<td>Switzerland</td>
<td>34</td>
<td>68</td>
<td>70</td>
<td>58</td>
</tr>
<tr>
<td>Turkey</td>
<td>66</td>
<td>37</td>
<td>45</td>
<td>85</td>
</tr>
<tr>
<td>United Arab Emirates</td>
<td>80</td>
<td>38</td>
<td>53</td>
<td>68</td>
</tr>
</tbody>
</table>

### Figure 30, phase one results for straight correlation between cultural variables and question types

Breaking the respondents into high/low cultural variable group give the following national groupings:

<table>
<thead>
<tr>
<th>Cultural Variable</th>
<th>Countries Represented</th>
<th>Participants Represented</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Power Distance</td>
<td>Argentina (1), Australia (20), Austria (12), Canada (44), Finland (10), Denmark (22), Netherlands (47), New Zealand (8), Norway (10), Sweden (37), Switzerland (9)</td>
<td>220</td>
</tr>
<tr>
<td>High Power Distance</td>
<td>United Arab Emirates (1), Turkey (5), Singapore (1), Portugal (6), Japan (1), Hong Kong (1), Greece (41), France (31)</td>
<td>87</td>
</tr>
<tr>
<td>Individualism</td>
<td>Australia (20), Austria (12), Canada (44), Denmark (22), Finland (10), France (31), Netherlands (47), New Zealand</td>
<td>250</td>
</tr>
</tbody>
</table>
Correlating the high/low groups and looking for statistically significant differences between them brought the following results:

<table>
<thead>
<tr>
<th>Cultural Variable</th>
<th>Related Question Groupings</th>
</tr>
</thead>
<tbody>
<tr>
<td>Low Power Distance</td>
<td>Feedback</td>
</tr>
<tr>
<td>High Power Distance</td>
<td></td>
</tr>
<tr>
<td>Individualism</td>
<td></td>
</tr>
<tr>
<td>Collectivism</td>
<td></td>
</tr>
<tr>
<td>Masculinity</td>
<td></td>
</tr>
<tr>
<td>Femininity</td>
<td>Navigability, Interactivity, Assessment, Socialization</td>
</tr>
<tr>
<td>Low Uncertainty Avoidance</td>
<td>Consistency, Feedback</td>
</tr>
<tr>
<td>High Uncertainty Avoidance</td>
<td></td>
</tr>
</tbody>
</table>

PDI represents the tendency in which a society reinforces individual or collective achievement. The Media question grouping refers to the appropriate use of media to portray visual and spatial concepts. EVE’s layout is not highly structured and the nature of the game involves a three dimensional work environment, outer space, which requires a highly developed spatial interface. The game play involves both the formation of groups to achieve goals as well as individual play in pursuit of personal goals depending on user preference. You can play alone or in a group as you choose. This pairing could almost represent a summary of the game as a whole, and certainly reflects its ability to fit with players from both rigid and loosely organized society. That the bulk of the gamers polled come from low power distance, PDI, counties shows what drew them to this type of game.
Low PDI people generally prefer shallower hierarchies and do not seek authoritative recognition and guidance. The Feedback grouping refers to the user’s need for information from experts in accomplishing tasks. EVE’s gamemasters appear in the game as robots and do not take active part in the game, delivering information only. Low PDI people appreciate the availability of information without intervention or judgment. This form of help scored high for the rest of the players as well, all except the High PDI people.

MFI, masculinity/femininity, returned the most varied results. Differences in masculinity and femininity highlight an emphasis on social roles, gender, and relationships, as well as differences in goals and achievement. These are the elements that make MMOG’s such good research vessels for social scientists. Important points to note: at the time of this study although roughly 50% of the game characters were female, although only 8 to 16% of the actual game players were. In EVE females are treated equal to males.

Navigability scored strongly in gender areas. Feminine players prefer more choices, such as modes of play (missions, trading, and player versus player) and navigational structures that support mutual cooperation and exchange. And they want a say in the evolution of the game’s storyline. These are all key elements in MMOG’s in general. It has been said often that EVE’s interface almost offers too much choice, tending towards cluttered.

Interactivity and Socialization scored strongly for both Femininity and MFI as well. Interactivity refers to the player’s ability to make use of new knowledge and skill, and the capacity of the game to motivate learning. Feminine players tend to prefer function, cooperation, and team oriented interactivity. Many players spend time in game forums and in-game chat rooms where they make use of log and multimedia video captures of game play to assess new ideas and game performance. Learnability is the mechanics of the layout in support of game play. In EVE, tasks are clear and well specified and it is generally accepted by players that there enough game help to aid in discovery. Much of the game though involves discovery. This is part of the play of the game. In-game chat also allows players to ask for help from their peers when things get confusing or when a player gets stuck in a difficult situation.

Consistency and Visual Design focus on the display and correct positioning of information, and the attractiveness of the interface. EVE’s display is freeform giving the player the ability to organize it as is expedient. The player can also decide what and what not to display. The game itself also offers rich vistas of planets, stars, and nebula.

Much of what attracted EVE players to the game becomes evident in analysis. EVE’s audience is greatest in feminine-rated countries. It’s flexible and generally flat social structure, high level of socialization and discovery, and opportunity for cooperation and

---

29 This is changing. Female/Male participation in MMOG’s in recent years is 40/60% and rising.
exchange make it attractive to these types of people. Cultural Dimensions can be an aid in the systematic assessment of the effects of culture on the popularity of a product.

### 7.2 Acceptance

If culture can be modeled and quantified, how can we apply these models to our software? The most widely used model for analysis of the acceptance of new technology is, naturally, the technology acceptance model (TAM). TAM is an extension of the Theory of Reasoned Action (TRA). TRA suggests that a person’s behavioral intention depends on their attitude and the subjective norms in which they live. $BI = A + SN$. If a person intends to do a behavior then it is likely that the person will do it. TAM, in its simplest form states that when confronted with a new technology, the user’s intention to use it is influenced by two factors: perception of usefulness and perceived ease of use. If it looks useful and not too much trouble, then they will probably use it.

![TAM Diagram](image)

**Figure 32, TAM, Technology Acceptance Model**

That seems obvious, but TAM gives us a standard framework of hypothesis with which we can compare data to create results we can use in studies. If you have figures on the perception of usefulness and ease of use, then you can statistically correlate them and get actual figures on use, in a way that everyone can understand, compare, and rely on.

In cases of culture and the use of technology, it might be nice to combine TAM with CDT. The diagram in figure 33 was developed by Xin Li, et al, in 2009, combining eight years of previous work, integrating individual dimensions and parts of CDT with TAM into a kind of grand unified version. Hybrid models like this have become more common recently in analyzing data involving culture and technology acceptance but as yet there is no settled standard.
The effects of culture are largely the effects of cultural norms and beliefs. To help with factoring these in, the authors tend to combine TRA with TAM, adding the relationship between Normative Beliefs and Intention to Adopt. This is an important one and has generally proven to be significant, especially in relation to culture. Normative beliefs represent the environment we grow up in, really the basis of culture. The relationships between CDT dimensions and TAM relationships were based on prior work as well. These relationships are hardly set in stone nor was this the first study to do this. A different take on this is the work of Srite and Karahana, an earlier study that uses only the original four dimensions: xxxi
Not all of these relationships are considered direct, some were considered moderating. There are relationships that are unexplored. For instance uncertainty avoidance or masculinity femininity and the relationship between perceived ease of use, willingness to accept a challenge, and perceived usefulness or perhaps power distance and the relationship between perceived usefulness and intention to adopt, willingness to accept risk in the face of authority.

7.3 Composition

Modifying TAM as a means to gauge the acceptability of our work, but are there other ways we can quantify culture? Since the user interface itself is the point of contact, wouldn’t it be nice if we could work out a system whereby we could modify it depending on who is looking? What if we could quantify esthetics?

When we look at Asian or South American advertising you get a definite sense of difference. There is something different in the choice of color or perhaps it’s the way things are spaced on the page. But how do you explain it? What exactly is different? Comparisons can only be antidotal and subjective – or can they? Is it possible to quantify the difference? More importantly, is there a reason for the difference? Do these cultures
have needs that drive these differences, needs that must be met if our web pages are to be successful?

We have seen that there are ways to quantify culture. We can quantify composition as well by breaking compositional elements into categories.\textsuperscript{xxxii} In a study by Bruce Lo and Panqun Gong this was accomplished by breaking site composition down into two categories:\textsuperscript{30}

- Color usage includes the choice of color for webpage background, foreground, frame, image, hyperlink and logo.
- Page layout includes components of the webpage: such as banners, animation, splash windows, rollover-graph/text, buttons, icons, white space and website navigation models.

100 e-commerce sites, 50 from China, 50 from the US were chosen based on popularity at the time. Using their front splash pages, page colors were simply counted grouping by primary and secondary colors with the following results:

![Color Usage](https://via.placeholder.com/150)

Figure 35, color usage differences between China and the US.

The Chi-square test was applied successfully to confirm that figures were significant and the following conclusions drawn:\textsuperscript{31}

\textsuperscript{30} Page 3
\textsuperscript{31} Page 4
• The predominant color for background is white for both countries (Figure 2). Background color usage here appears to show some trend for standardization.

• In terms of the foreground color usage (Figure 3), blue, black, and yellow are the three leading colors in both countries. But it would be difficult not to notice the prominence of red as a foreground color in Chinese sites.

• In terms of total color usage, Figure 5 shows that the dominant colors for US are blue (23%), white (19%), Black (15%) and Yellow (14%); and for Chinese are blue (18%), red (16%), yellow (15%), and black (13%). Blue is the most frequently used for both countries. But the prominence of red and the relative de-emphasis of white in Chinese sites are clearly evident. Traditionally Chinese associate red with good luck, wealth, and marriage (e.g. bridal dress) and is naturally included in business sites. White is associated with death and mourning.

• The US sites show a clear trend of using blue color for hyperlinks, which is the default for most browsers. Interestingly, in Chinese sites three colors, black, blue and red dominate equally.

• With respect to logo colors (graph not shown), black and red are the clear favorites for Chinese sites while white, black, blue, and red have equal frequencies on the US sites, perhaps because of the US flag.

Web page layout components counted were:

<table>
<thead>
<tr>
<th></th>
<th>US</th>
<th>China</th>
</tr>
</thead>
<tbody>
<tr>
<td>Active Banner</td>
<td>0.1</td>
<td>1.14</td>
</tr>
<tr>
<td>Active Animation</td>
<td>0.52</td>
<td>4.92</td>
</tr>
<tr>
<td>Rollover Graph</td>
<td>1</td>
<td>17</td>
</tr>
<tr>
<td>Rollover Text</td>
<td>1</td>
<td>13</td>
</tr>
<tr>
<td>Splash Windows</td>
<td>15</td>
<td>12</td>
</tr>
<tr>
<td>Frames</td>
<td>5.5</td>
<td>9.3</td>
</tr>
<tr>
<td>White Space</td>
<td>41</td>
<td>19</td>
</tr>
<tr>
<td>Buttons</td>
<td>7.3</td>
<td>12.8</td>
</tr>
<tr>
<td>Icons</td>
<td>1.2</td>
<td>0.72</td>
</tr>
</tbody>
</table>

Figure 36, web page component usage counts, China vs. US.

US web sites made far more use of white space, Chinese sites being more densely packed. Chinese sites made more use of active content such as banners and animations. They also had more buttons per page. This denser page composition may have to do with the way Chinese is written, with block shaped characters being able to be more densely packed on the page. This may carry over to Chinese web content. Not listed is a difference in buttons with Chinese favoring text based buttons and US sites favoring graphic buttons.

If we were to build a system of cultural adaptation, much in the way we are attempting to translate language, then factors like these would be a component, much like cultural
dimensions. For instance, some page translation software indulge in a bit of page
reformatting as part of their process. This is necessary because of the differences in text
size and orientation between different languages. Page density could be a factor
dependning on the culture of the viewer. A Chinese customer could have more packed on
each page. It’s a matter of finding factors that balance and unify the two cultures and
expressing them visually in our pages.

7.4 Conclusion

For companies looking to build a global web site that adjusts itself based on culture, it’s
not so much a matter of “can it be done” as it is a matter of “how to approach it” and
whether that approach is cost-effective or cost-prohibitive. There are basically two
approaches: localize and internationalize manually or automatically. Most companies,
with a few exceptions, don’t have the capital and human resources to staff content
managers and developers for every country in which they do business. At the same time,
the technology to automate localization, to create cultural adaptive services isn’t here yet,
mostly because it has yet to be tried.

Language translation services exist, but they are still a ways from being able to create
natural sounding translations. There’s little in the way of cultural adaptive services and
this latter capability can be as important as language translation. We have models that we
can use to successfully analyze problems in localization, analyzing the acceptance of
single sites or products, but nothing that can model culture globally in any sensible
generalized way, to help us create general models that will help us to create universally
acceptable sites. The idea of a universal web site that adapts to not only linguistic, but
cultural differences as well seems possible and research in this area needs to be done.
This is an important area of research because, as we saw with Google, these problems can
have geopolitical implications that can make or break the economies of nations.
8.0 Final Summary

In the last millennia, we were able to ignore our audience because our audience was us and our knowledge of us was complete. We were blind to this facet of our work. Look at what happened to Google! They went in to China with an “us and them” sensibility. But the U.S. is no longer an island. We must compete in a larger aggressive world in which we are no longer the center. We as technicians need to understand our customers even when we don’t directly face them.

Think of our work as the work of an artist. A good artist cares about the stability of pigments, the quality of surfaces, the ability of brushes to keep their points, but that’s

<table>
<thead>
<tr>
<th>Rank</th>
<th>Language</th>
<th>Primary Country</th>
<th>Countries Spoken</th>
<th>Total Speakers</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Chinese</td>
<td>China</td>
<td>31</td>
<td>1,213</td>
</tr>
<tr>
<td>2</td>
<td>Spanish</td>
<td>Spain</td>
<td>44</td>
<td>329</td>
</tr>
<tr>
<td>3</td>
<td>English</td>
<td>United Kingdom</td>
<td>112</td>
<td>328</td>
</tr>
<tr>
<td>4</td>
<td>Arabic</td>
<td>Saudi Arabia</td>
<td>57</td>
<td>221</td>
</tr>
<tr>
<td>5</td>
<td>Hindi</td>
<td>India</td>
<td>20</td>
<td>182</td>
</tr>
<tr>
<td>6</td>
<td>Bengali</td>
<td>Bangladesh</td>
<td>10</td>
<td>181</td>
</tr>
<tr>
<td>7</td>
<td>Portuguese</td>
<td>Portugal</td>
<td>37</td>
<td>178</td>
</tr>
<tr>
<td>8</td>
<td>Russian</td>
<td>Russian Federation</td>
<td>33</td>
<td>144</td>
</tr>
<tr>
<td>9</td>
<td>Japanese</td>
<td>Japan</td>
<td>25</td>
<td>122</td>
</tr>
<tr>
<td>10</td>
<td>German, Standard</td>
<td>Germany</td>
<td>43</td>
<td>90.3</td>
</tr>
<tr>
<td>11</td>
<td>Javanese</td>
<td>Indonesia</td>
<td>5</td>
<td>84.6</td>
</tr>
<tr>
<td>12</td>
<td>Lahnda</td>
<td>Pakistan</td>
<td>8</td>
<td>78.3</td>
</tr>
<tr>
<td>13</td>
<td>Telugu</td>
<td>India</td>
<td>10</td>
<td>69.8</td>
</tr>
<tr>
<td>14</td>
<td>Vietnamese</td>
<td>Viet Nam</td>
<td>23</td>
<td>68.6</td>
</tr>
<tr>
<td>15</td>
<td>Marathi</td>
<td>India</td>
<td>5</td>
<td>68.1</td>
</tr>
<tr>
<td>16</td>
<td>French</td>
<td>France</td>
<td>60</td>
<td>67.8</td>
</tr>
<tr>
<td>17</td>
<td>Korean</td>
<td>Korea, South</td>
<td>33</td>
<td>66.3</td>
</tr>
<tr>
<td>18</td>
<td>Tamil</td>
<td>India</td>
<td>17</td>
<td>65.7</td>
</tr>
<tr>
<td>19</td>
<td>Italian</td>
<td>Italy</td>
<td>34</td>
<td>61.7</td>
</tr>
<tr>
<td>20</td>
<td>Urdu</td>
<td>Pakistan</td>
<td>23</td>
<td>60.6</td>
</tr>
<tr>
<td>21</td>
<td>Turkish</td>
<td>Turkey</td>
<td>36</td>
<td>50.8</td>
</tr>
<tr>
<td>22</td>
<td>Gujarati</td>
<td>India</td>
<td>20</td>
<td>46.5</td>
</tr>
<tr>
<td>23</td>
<td>Polish</td>
<td>Poland</td>
<td>23</td>
<td>40.0</td>
</tr>
<tr>
<td>24</td>
<td>Malay</td>
<td>Malaysia</td>
<td>14</td>
<td>39.1</td>
</tr>
<tr>
<td>25</td>
<td>Bhojpuri</td>
<td>India</td>
<td>3</td>
<td>38.5</td>
</tr>
</tbody>
</table>

Figure 37 languages supported out of the top 25 languages of the world by Amazon.com.
only the start. There’s an ultimate goal to be achieved in any work, ideas that must be expressed or all the technique is wasted. In focusing narrowly on technique we become blind to the bigger picture, our ultimate goals.

We saw Google move in to China with a US cultural bent, a one size is good enough for all attitude. It was an attitude we didn’t know we had, and how could we? We have been the sole dominant force in computing since its creation. But now this is changing. We have discussed the Internet in terms of a global standard, with all of our work being a part of it. It’s unifying the world, sometimes to the detriment of smaller cultures. These cultures will fight back because exclusion can mean death. And we looked at one possible way in which that standard might made to be self adapting with cultural adaptive services. Culture can be approached in an automated way.

There are areas in which we can further research along these lines. Culturally adaptive services can be combined with translation services to produce more acceptable and pleasing page translations, and perhaps eventually even the holy grail of Internet computing – the universal web page. In situations where page complexity is limited, this may not be as difficult to achieve as it seems. Form input, menus, and data output are perhaps the most common forms of documents and have limited cultural scope. We are already tailoring output in some situations to personal preferences.

In order to grasp the Internet of the new millennia we must make the subtle shift, the change from “us and them” to “we”. We must face the truth that the US no longer entirely calls the shots. We must work within a larger evolving world system, a system that can be gamed if we make the effort grasp the entirety of it. Our engineers need to be pointed towards the bigger picture and finding our place in it. After all, your next clients may be in Dubai or Peking and even if they aren’t, your coworker may be.
References

Dr. Khin Marlar Tun Yuzana. A Comparison of Collation Algorithm for Myanmar Language. IEEE, DOI: 10.1109/ICDIM.2008.4746740


http://doi.acm.org/10.1145/1409540.1409578


The Gamer Experience: Investigating Relationships Between Culture and Usability in Massively Multiplayer Online Games. Panagiotis Zaharias, Anthony Papargyris, ACM, Computers in Entertainment (CIE) - SPECIAL ISSUE: Media Arts and Games (Part II) archive, Volume 7 Issue 2, June 2009. DOI: 10.1145/1541895.1541906


