Ideas on Learning a New Language Intertwined With the Current State of Natural Language Processing and Computational Linguistics

Robin M Snyder
RobinSnyder.com
robin@robinsnyder.com
http://www.robinsnyder.com

Abstract

In 2014, in conjunction with doing research in natural language processing and attending a global conference on computational linguistics, the author decided to learn a new foreign language, Greek, that uses a non-English character set. This paper/session will present/discuss an overview of the current state of natural language processing and computational linguistics intertwined with ideas for using traditional and technology-assisted ways of learning a new language. Included will be practical issues of character set representations, translation technologies, web-based and traditional resources, etc., and ideas on how to immerse oneself in language learning and/or integrate such methods into a class whose goal involves learning a foreign language.

Languages and writing

Language is the means by which humans communicate meaningful and abstract thought, primarily by making a noise that is recognizable and understood by another human. That is, a message that goes from an origin to a destination. On the other hand, writing is a technology. Language is not a technology. According to Alan Kay, a primary inventor of "object-oriented" and graphical user interface fundamentals in the 1970's at Xerox PARC which led to the introduction of the Apple MacIntosh computer in 1984, "Technology is what wasn't around when you were a kid.".

At some point in human history, writing was invented as a technology. Written communication was first done with sticks making marks on mud tablets and then dried. Later, written communication was done on papyrus form by the Egyptians and stored as scrolls. Later yet, parchment (animal skins) was invented at Pergamus in Asia Minor, eventually cut and bound into codices/books. The concept of paper from wood fibers from the Chinese greatly expanded the use of written communication. At the end of the 20th century, the use of physical written paper began to disappear, with large phone books, out of date shortly after being printed, began to stop being printed. [1]

Languages are based on a finite set of words. This finite set ranges from a few thousand words for useful communication to upper limits of hundreds of thousands, based on the source, for good communication. But a finite set of words and grammar constructs allows humans to express a countably infinite number of thoughts. One can, statistically speaking, easily construct a meaningful sentence that has never before been expressed.

That being said, the author is still learning "English" and will never know it completely. The same is true for any language. One needs to decide the meaningful level of proficiency one desires to attain within the resources of time and space available.

Writing was originally on clay tablets, then on scrolls, then on parchments, with the creation and copying of text a manual, tedious, and potentially error prone process.

It took a while for people to realize that one can read silently without reading. Augustine wrote in the 4th century that he was amazed when he realized that one could read silently without reading aloud and listening to what one was reading. The audit process was originally a process by which the text of records was read aloud. The legal process still uses the terms hearing, and hears evidence.

To make the "paperwork" easier to run his empire in the 800's in Europe, Charlemagne invented lower case letters to same space on valuable and costly parchment. Italic type originated in the same manner, to save printing space. The growth of the written word was greatly increased by the "invention" of movable type by Gutenberg. A half century later, Martin Luther posted some "academic" discussion points on a door. A friend, with the help of a printing press, reproduced them and distributed them all over Europe. Within a month, Europe was into a religious war that wold go on for quite some time, showing the potential power of information. As the use of the printing presses increased, the cases in which the type was stored led to the terms "upper case" and "lower case" letters, in contrast to the terms "magniscule" and "miniscule".

Learning languages

According to Steven Pinker [2], a leading authority on acquisition and use of human language, the ability to learn and communicate via language is built into humans (i.e., in the DNA code) though much is not understood about language as it continues to be an important research area. Children have a built-in mechanism to learn any language to which they are exposed until a certain window of opportunity is closed, somewhere between 8 and 12 years old. Thereafter, learning a new language is much more difficult. Learning a language as an adult is not easy.

The brains of children are designed to learn and construct syntax, grammar, and meaning from what they see, hear, and learn from their parents. How fast do children learn language? Pick a number of words, say 6,000, even though precisely defining words is difficult. Pick an age range to start talking, say 2, and an age to check progress, say age 8. This is 6 years or 6*365 = 2,190 days. To accomplish this, the child must learn about 3 new words every day while integrating those new words, in the proper grammar context, with everything learned so far.

Children will fill in gaps and extend the language using rules of other constructs. My son would sometimes inadvertently annoy his older sister by making statements such as, "Can we go tennising today.". That is how new language is created, but enough people must agree and start using the new terms until they become established.

An adult must do the same thing as a child, but with the burden of already knowing a language, and having some of the switches for learning in the brain, as theorized by linguists, turned off.

Years ago, the author switched from the QWERTY keyboard layout to the Dvorak keyboard layout. Although one can become productive in a week or so, it took months for the typing to become automatic without reverting to learned and common QWERTY keystroke patterns.

Years ago, the author spent time learning Classical and Biblical Greek. Unfortunately, at least in the United States, this is mostly done using a contrived Greek pronunciation that goes both ways (reading to speaking, and listening to writing), which is good for academic study, but the modern Greek pronunciation only goes reading to speaking. It took months to automatically change to the new reading and speaking patterns.

The Greek language has a long history. An early form of Greek, called Linear B, uses an older alphabet. The current alphabets of all alphabetic languages (pretty much by definition) derives from the Phonetician/Hebrew alphabet. The Hebrew form changed during during the Babylonian captivity. The Greeks adapted the original alphabet.

Originally, the alphabet was used for a syllabic script where there are no spaces between words and the letters are like a recording of the spoken word. When listening to a foreign language, one notices that there appear to be no spaces between words. The brain learns to put spaces between words where, in actuality, there are none. Over time, spaces between words made the written word easier to work with with fewer scribal copying errors.

English words from the Greek come from various sources. Words can come directly from Ancient Greek, through other languages such as Latin, as terms borrowed starting with the Renaissance and the rediscovery of the Greek language. In modern times, Greek has borrowed many English (and other) words.

In languages such as English, seeing a word may not mean one can pronounce the word, and hearing the word may not mean that one can write the word. In German, the process pretty much goes both ways. In Modern Greek, if one can see the word, one can pronounce it, but if one hears it, there may be many possible ways to write that word. In Chinese, hearing the word gives no clue as to how to write it, and seeing the word provides no clue as to how to say it, and dialects of Chinese can pronounce the same written word in vastly different ways.

Linguists make the distinction between a pidgin language (note, not the same word as the pigeon bird) and a creole language. In general, a pidgin language is a greatly simplified language created by adults in order to communicate for practical (e.g., business) purposes. Pidgin languages lack the ability to express precise and abstract thought, being a practical language compromise for immediate purposes. A creole language tends to be a pidgin language, created by adults for practical purposes, that is extended and filled out by children who are in an environment where they must deal with a native language at home but another language with other children. In order to more fully communicate, the language is made more complete.

Note that, in practice, what people call pidgin languages, creole languages, and language dialects, may not fit the general pattern described. For example, there are many "dialects" of Chinese, but although they share the same writing system, they are very different spoken languages. The concept of precisely identifying a language can be confusing. As example, the verbal language of Serbo-Croatian is called Serbian when written in the Cyrillic alphabet but called Croatian when written in the Latin alphabet.

A dialect of a language is not easy to define but tends to be a variation of a language that is still understood by another group but yet differs in the way it is pronounced, written, structured, etc. Most languages have what is called a high language, or official language, that is the way in which written and formal communication is done. Then the low language, or local language, or language dialect, is the way that local people communicate. In high school or college, a student is almost always taught the high language, so they can communicate anywhere that language is spoken. But when they travel to a country that speaks that language, their speech is immediately recognizable as a non-native speaker since no one speaks the high language in practice. In English, the high and low languages are less recognizable but there are some traces. But even English has many dialects throughout the world. See Pediatric for a long list of English dialects.

Immersion

It takes a lot of time and a lot of effort to learn a new language as an adult. To effectively and efficiently learn a new language as an adult, it helps greatly to organize one's efforts in time and space. In essence, one must learn to immerse oneself in the language.

Estimates vary, but a concerted effort of about six months time, more or less, is needed to attain a decent pidgin language proficiency, where one knows hundreds of simple words, some simple grammar rules, common phrases, and is comfortable using them in normal conversation. It takes a year or more to fill in more details and refine the language. It takes many years to attain great fluency in the language, and one may never lose the accent so that native speakers will usually always be able to detect that non-native accent.

In essence, to learn a new language, one must immerse oneself in the language to the greatest extent possible. Fortunately, with the available resources of the Internet, this is much easier than at any point in the past.

Author background

In early junior high school, the author discovered a book an linguistics that was fascinating. At the same time, that school provided a half year of Spanish, French, and German, from which the student, if so inclined, could pick one of the three for study in 9th grade. The author spent 3-1/2 years in high school and 2-1/2 years in college studying German and had become quite proficient in it, to the point where the author could think in that language. In 2014, the author decided to learn modern Greek, which uses a very different character set than English or German. The author has had a lifetime interest in languages, both human languages, ofter called natural languages, and programming languages which have many similarities but also many important differences. With a PhD in computer science in programming languages, and working in areas of applied research in natural language processing and computational linguistics, the study of learning the Greek language and processing Greek text fit in with both personal and professional objectives. In addition, the Greek language has an almost 3,000 year history of use so changes over time can be studied, many English words have their origin in Greek, and in doing sophisticated text processing, the Greek character set provides challenges to getting programs and features to work properly.

To be really proficient, one needs to get to the point where one can think in the language without actual translation between native and non-native languages. The same kinesthetic aspects are required for sports, music, etc. In music, another lifelong interest of the author, one gradually learns to recognize patterns and use those patterns in both reading, playing, and improvising music. From a language perspective, the author has found, often, that when trying to get the Greek word to pop into mind, if it is not readily available, the German word pops instead into mind, not really desirable. Switching between languages while speaking (e.g., words in one language, grammatical structure in another language) is what linguists call code switching.

Getting started

How exactly does one get started learning a new language? There are a large number of products that claim to help one get started in learning a language, and they vary greatly in price and value. The author has tried many of them. The general approach used by the author was as follows.

1. Learn some basic words and some useful grammar, including some common irregular constructions.
2. Learn basic phases for travel, etc., using any of the many options available.
3. Long term fill-in with more vocabulary, nouns, verbs, idioms, constructions, etc.

Many native speakers speak the language very fast. In addition, and, as in many languages, if the vowel sound at the end of one word is the similar to the vowel sound at the start of the next word, those words are just combined together in speech, but not in the written words, which have the artificial spaces between them.

The best method found for #1 by the author was the Michel Thomas method, available on CD, in three versions. The "Start" method (1 CD), the "Total" method (4 CD's, including the beginning CD) and the advanced method (3 more CD's). The Michel Thomas approach assumes that the teacher is teaching two students, and you are the third student. The quality of the teachers and sound quality vary, so one may want to try the beginning CD before purchasing more. The Greek was excellent. The Spanish was much less so, and the sound quality was not that good. The languages supported include: French, Spanish, Italian, German, Dutch, Greek, Portuguese, Japanese, Polish, Russian, Arabic, Mandarin Chinese.

Another author requirement was for the system to be mp3-based so that the time spent learning could be combined with other activities (e.g., working out, riding bicycle on the trail, commuting, etc.). There is a lot of marketing for Rosetta Stone. But their approach used pretty much requires the user to sit in front of a computer and run a copy-protected program that could possible interfere with the normal working operation of the computer. So Rosetta Stone appeared to be very expensive, not geared towards deep grammar and written communication understanding, and so did not meet the basic requirements of the author for a language learning system.

If one does #2 first, one hears and repeats many phrases, but it is often very unclear where the phrases come from and how they are constructed. Doing #1 first, when one hears phases from #2, one has a much better idea of the actual construction properties such as gender, case, irregular parts, etc.

Some languages are spoken at a very great speed. Some systems account for this, but others do not. It as a basic property of most learning that if one learns to do something slowly, one can always increase the pace, but starting at a higher speed does not work well if at all. The author has found this true in music, athletics, and other areas such as learning languages . In earlier years, the author spent many years in musical endeavors. One jazz band director explained many times (paraphrased here), "If you are not good at music, play loud and fast. It is much harder to play slow and soft". Needless to say, we had to learn to play slow and soft, at which point, we could then easily play loud and fast, and vary the volume and pace as needed.

In the absence of people speaking the desired language, recognizing language can be facilitated with audio CD's, podcasts, DVD's, etc.

The web site at http://www.50languages.com allows one to "**Learn more than 50 languages for free online or with our Android or iPhone AP**". There are more than 8 hours of audio recordings of basic phrases from any one of 50 languages to any of the other remaining 49 languages, though at some point more languages may be added. Part of their business model is to sell a $10 book showing the written form of the audio. The author found the German once and Greek twice (i.e., two speakers) to be useful in reviewing German and learning Greek.

A daily calendar in the language of interest, if available, can be useful.

The site at http://www.innovativelanguage.com offers many ways to learn a language. Their free word a day email is useful. The author subscribed to their advanced model for a few months. They tend to send a lot of marketing emails, and their audio is not perfect, but it can be very useful approach to learning a new language. The word a day tends to repeat itself after a few months.

YouTube has a huge collection of video's in many languages, but one may need to search around for them. One needs to sift through the beginning ones of basic words, to advertisements for purchasing other systems, etc. Once native language videos are found, there are usually links to other such videos. The author found many YouTube videos in Greek for learning programming, solving high school math and science problems, science and technology, lectures on various topics, etc., as well as radio, TV, and children's shows in Greek. One just needs to spend some time searching. To get started, it may help to use the transliterated name in the search. If one searches for "Greek" on gets advertisements and beginning Greek words. Searching for the transliterated name "Ellinika" provides better results. Better yet, change your Google account to the desired language - which can then take effect on every device used with that account. The same strategy works well for searching in, say, the iTunes Apple Store for podcasts, Amazon.com for books, etc.

Devices and software

Some easy places to immerse oneself are by changing the language of the phone and/or the computer. Warning: Whenever you change a language in a phone, computer, program, etc., document exactly how you did it so that if you need to, you can change the language back. This is especially important when faced with important and long messages, user agreements, etc. If possible, having a similar device that is in one's native language can be useful if one gets lost in the new language. Be aware that

your device language has changed. At a wedding rehearsal dinner, my brother asked to use my phone. No problem. I handed it to him. He looked at it for a while, then said, "How do I use this?".

For some reason, the Amazon.com music program will not work if one sets the language of the phone to Greek, or a host of other languages. No problem. Just use the Google music player which works just fine.

Operating systems such as Windows may need a special version to switch the language. Windows 7 Professional can only be installed in one language. The author upgraded one machine to Windows 7 Ultimate so that any of more than thirty languages could be used. Be aware that some programs may alter their behavior when the language is changed. Microsoft Office reset some of the default features that caused existing VBA macros to not be found, so they needed to be referenced again.

Keyboards to get alphabets into the computer are straightforward. Keyboards to get any one of thousands of Chinese characters into the computer are not. The Greek language is based on 26 characters, the vowels of which can have accents, and most of the letters map/transliterate to English characters.

Learning ideas

This section mentions learning ideas that the author has found useful, or, in some cases, less than useful.

Amazon.com has stores in Germany, France, Spain, and, by the time this is read, perhaps in other areas. These are a great place to get books in other languages. Since the author had learned German in earlier years, German books that teach Greek were/are a good way to review German while learning Greek. For some of the audio CD's, the Greek speakers tended to use a German accent.

DVD's are a great way to get improve a language. Greek DVD's are mostly require a Zone 2 or Zone free player, but these are not difficult to obtain. Since the author had watched many animated movies with his son many many times, in watching the same movie in Greek, the author already knew the English lines which made it easier to concentrate on the Greek lines. Note that the subtitles appear to be direct translations of the English into Greek. The Greek actors, however, use the Greek that best approximates the semantics expressed by the characters. In that sense, the subtitles are not very useful. However, one can obtain the text of the subtitles for many movies in many languages online. These can be useful in studying the foreign language as some words can be difficult to hear and not easy to get off the screen when one needs them.

Audio and music CD's are great for helping learn the language. Children's books that have an audio CD or mp3 that allows one to listen while following along in the book are very helpful.

There is a huge market for children's books to help in the learning process. Those same books, with stories, pictures, etc., can help adults with the same basic vocabulary and terms.

At some point, one will want books in addition to the ones in the local bookstore or from Amazon.com, whose selection of Greek books is very limited. The Internet can help in finding a bookstore in the country here the desired language is spoken. The author eventually located the Greek bookstore at

http://www.malliaris.gr located in Thessalonica. The shipping costs are high, of course, but the credit card company provides a reasonable conversion from the Euro into dollars. It takes about 3 weeks from order time until delivery time. One can then order the English version of the book, if available, from Amazon.com. However, it is very useful to order some books written by authors in the native language. This author found a great number of children's books and books with audio CD's that were very useful.

Google has news available in many languages, including Greek. Search for Google news in the desired language.

Research articles in the desired language can be useful. One can use the Google advanced search with, say, a file extension of "pdf" and a language of "**Greek**".

Character representation

When the first working programmable computer was developed in the 1950's, the language was English (primarily United States and Great Britain). A small character 7 bit character set called ASCII, character codes 0 to 127, later 8 bits for character codes 0 to 255 for extended ASCII, sufficed (ignoring the IBM 9 bit EBCDIC code). Western languages such as Spanish, French, German, etc., which use the Latin alphabet were added using special characters in the extended range 128 to 255, using a technique called code pages (still used in Windows by default). The oriental languages, referenced as CJK (for Chinese Japanese Korean) required thousands of character glyphs each. The Unicode representation was developed, of which there are many variant encodings, the most common of which today is a UTF-8 encoding that allows a more compact encoding/storage of characters in 1, 2, 3, or 4 bytes. UTF-8 is backward compatible with the original ASCII, codes 0 to 127, while avoiding issues of large or small endian encodings (i.e., high to low end ordering of bytes, or low to high ordering of bytes). Each language, even ancient languages, have a space in the Unicode code space.

A useful program to explore and work with Unicode is BabelMap for Windows. A Linux alternative for exploring fonts, etc., is **`gucharmap`**. Note that to see the characters one needs a font installed that supports those characters. A fairly complete, but huge, font that does this for modern languages is the Microsoft "Arial Unicode MS" font. Unicode fonts such as "**Egyptian Hieroglyphics**", "**Linear B Ideograms**", "**Cuneiform**", etc., need other fonts.

In programming systems with ASCII, Unicode, UTF-8, etc., it can become very problematic if the program processes text as if it were in the wrong encoding.

Computational linguistics

Computational linguistics is the study of using computation to perform natural language processing. Natural language processing is inherently difficult, the common example being the following.

- Time flies like an arrow.
- Fruit flies like a banana.

Another being the following.

John saw Mary with a telescope.

Who has the telescope? If a human cannot tell, how could a computer? An upper bound for recognizing parts of speech in a language is about 96% since human experts in the field tend to disagree on the assignment of parts of speech tags.

But just like weather prediction, though chaotic and not exactly predictable, it is useful. If the forecast is pretty good out to 48 hours, natural language processing and computational linguistics can provide useful results.

Expert systems have two primary means implementing expertise. One is a rules-based model whereby rules are used. The other is a statistical pattern based model where patterns, and lots of data are used. In practice, both methods are used. In machine learning, a lot of effort is put into what is called feature extraction (e.g., data scrubbing, picking out parts of an image, etc.), and then the statistical models are used on the extracted data.

Natural language processing has traditionally been rules-based, with tokenizers, parts of speech taggers (for nouns, verbs, etc.), grammars, both fixed and probabilistic, etc. A few years, Google came in and easily beat every rules-based system by using a probabilistic pattern-based machine learning technique that had millions of documents of translated works (e.g., United Nations transcripts, etc.) on which to train.

Google makes their translation system, based on machine learning techniques, available to everyone for browser web use for free, and for other commercial type uses for a minimal fee. Their free translation system is great for anyone learning a language. They use crowd sourcing to help improve the system, whereby anyone who sees something incorrect can help provide a better translation. But be aware that some phases do not translate perfectly. Idioms may not translate well. And the case, upper, lower, etc., may not match as the translations are based on machine learning matching.

Software systems

Many native speakers speak the language very fast. Software that can slow down the speed without changing the pitch can be useful. Such software, often command line, include Sox. The author prefers command line programs that can be automated for large numbers of files, but Audacity in a GUI program that can be used to do similar things.

As a research and learning tool, the details of which are beyond the scope of this paper, the author has created a system that has the following features.

1. Supports typing in either language based on a mode. The Greek characters are transliterated with the single quote modifying the character before to either have or not have the quote. If in the wrong mode, switch the characters already typed.

2. Hyperlink from a Greek or English word to Firefox and pull up the Google translation. In vocabulary files, the form of a line is as follows.

```
Greek text = English text
```

So if the cursor is to the left of "=" the mode is Greek, to the right of the cursor, English.

3. The vocabulary documents can be processed by scripts to extract the corresponding English and Greek words. If the English is not provided, do a programmed translation via Google. Divide the groups of words into sections based on hints.

4. For each English and Greek word/phrase, use Google to obtain a mp3 file of each translation.

5. Convert each mp3 to wav, decreasing speed if necessary, then combine all of the desired ones together according to patterns, such as English once, Greek three times, low sounding beep, repeated for all words, with three low sounding beeps at the end of each section.

6. Create a Word document for reading the words while listening. A quick version of the mp3 file, just one Greek at a time, is for proofing and reading along.

This system allows the author to put together vocabulary words gradually, and then listen as convenient. Another system hooks in with subtitles for movies as another way to learn vocabulary.

Another related system takes Greek words/text, divides it up, and creates a web page with words not known yet with links that, when clicked, bring up the Google translation for that word in Firefox.

Summary

This paper/session has discussed and/or demonstrated some ideas on learning a new language intertwined with the current state of natural language processing and computational linguistics.

References

[1] Gleick, J. (2011). The Information: A History, a Theory, a Flood. Pantheon Books.

[2] Pinker, S. (1994). The Language Instinct. HarperPerennial.

[3] Clark, A. and Fox, C. and Lappin, S. (2012). The Handbook of Computational Linguistics and Natural Language Processing. Wiley.

[4] Merkouri, A. (2012). Greek: Dictionary & Phrasebook. Hippocrene Books.

[5] Kambas, M. (2003). Greek-English Concise Dictionary. Hippocrene Books.

[6] Uhlig, E. (2003). Beginner's Greek. Hippocrene Books.

[7] Merritt Ruhlen. (1994). The origin of language: tracing the evolution of the mother tongue. Wiley.

[8] Pustejovsky, J. and Stubbs, A. (2012). Natural Language Annotation for Machine Learning. O'Reilly Media, Incorporated.

[9] Kumar, E. (2011). Natural Language Processing. I.K. International Publishing House.

[10] McWhorter, J. (2003). The Power of Babel: A Natural History of Language. HarperCollins.

[11] Berlitz Publishing. (2012). Greek Phrase Book and Dictionary. Berlitz.

[12] Manning, C.D. and Schuetze, H. (1999). Foundations of Statistical Natural Language Processing. MIT Press.