

# Chatterbots Go Native: Considerations for an eco-system fostering the development of artificial life forms in a human world

*Dr. Richard Wallace – Alice Foundation*

*Dr. Hideto Tomabechi – Cognitive Research Labs.*

*Dr. Doubly Aimless – Pandorabots*

*(January 12, 2003)*

*Abstract: AIML - A recently developed standard for describing knowledge for Robots using XML data structures is introduced.*

*A very large English-based robot knowledge base developed using the standard is described. Many other knowledge-based projects based on other international languages are described. The English-based knowledge base is currently interacting with people at a rate of 100,000 inquiries per day.*

*A unique web-based computer architecture ([www.pandorabots.com](http://www.pandorabots.com)) supporting the on-going development of the knowledge base is described. The architecture is distinguished from other architectures by providing a very low-cost environment for the development and distribution of multi-lingual robot knowledge bases. The architecture is highly scalable - during the last eight months the system has added more than 6000 different Robots. Currently, the number of new robots is doubling at the rate of every 60 days. The system handles more than 3 million inquiries a month. The architecture supports rapid acquisition of new knowledge while interacting with "unskilled humans". The architecture is secure and able to dynamically develop effective defenses against large-scale attacks while continuing to run.*

*An eco-system supporting the natural evolution and rapid development of these artificial life forms is described.*

## Software Robots – What are they and why are they here?

Software Robots – or Chatterbots – are encountered daily in a variety of forms. Unhappy recipients of Spam e-mail advertisements would like to eliminate the robots generating the messages. Automatic telephone answering systems (another form of a Software Robot) are sometimes difficult to avoid. Yet Software Robots are here with us to stay. Interesting varieties are emerging world-wide daily. America On Line's Instant messenger service (AIM) is very popular in portions of the world, especially among teenagers. And already Software Robots are making an appearance. We know of one teenage girl who created a Chatterbot and used it as a kind-of proxy for herself (*available for viewing and interacting with at: <http://www.international-lisp-conference.org/Competitions/Chatterbots/ILC02-chatterbots.html>*)

Teenagers often have multiple screens open with AOL and spend time switching between screens chatting with friends. This teenager started cutting

and pasting inquiries into her Chatterbot and redirected the Chatterbot's responses to her friends – without them realizing they were interacting with a Chatterbot. She spent more than 2 hours cutting, pasting and laughing hysterically. She went on to enter her Chatterbot in a software robot beauty contest and it won. herself (*The Chatterbot is available for viewing and interacting with at: <http://www.international-lisp-conference.org/Competitions/Chatterbots/ILC02-chatterbots.html>*)

Meeting singles on-line has rapidly mushroomed into a large industry in the US. One enterprising and eligible young man has created his own Chatterbot which now asks a prospective date a number of questions - and acting as a pre-screening proxy for him.

Business uses include online sales representatives and help desks, and advertising. Imagine sending a Chatterbot into anonymous chat rooms. The Chatterbot talks to some one for a few minutes and then suggests they see a movie.

Yet perhaps the biggest markets are the Entertainment markets.

We can imagine Chatterbots acting as talking books for children, Chatterbots for foreign language instruction, and teaching Chatterbots in general.

Now each of you have an idea of what the future will certainly bring and we want to ask you to hold that in your mind while we consider various alternatives for implementing these Chatterbots.

We need methods of describing and installing knowledge in these Chatterbots. We need ways to speak with Chatterbots. We need ways to visually interact with them. And we need a way to describe how we touch and are touched by these Chatterbots.

Common to all these implementation issues is how we describe the Chatterbot interactions.

Dr. Richard Wallace, Director of the non-profit foundation - the Alice Foundation - at [www.alicebot.org](http://www.alicebot.org) - has been working for years on exactly these issues. The foundation has developed an open standard XML-compliant language called AIML (Artificial Intelligence Markup Language). AIML is used to structure a Chatterbot's knowledge – so in effect, AIML codes the *knowledge portion* of the Chatterbot. The foundation offers a the Alice Chatterbot (see:

<http://www.pandorabots.com/pandora/talk?botid=f5d922d97e345aa1>)

based on a very large (English language-based) knowledge set (using AIML) available under a GPL License. Numerous translations to other international languages are underway and most are also freely available. AIML and Alice are implemented in a variety of computer languages: including versions in C, C++, Java, SETL, Lisp, etc. All are freely accessible at locations described at [www.alicebot.org](http://www.alicebot.org). Foundation supporters also have access to the Foundation's most current research knowledge set. Additionally, numerous commercial implementations of Alice also are available.

## **A short history of AIML**

AIML and the Chatterbot implemented with AIML - Alice - began in 1995 inspired by an earlier Chatterbot called Eliza – a Chatterbot loosely modeling a Psychiatrist. Alice and its implementation language AIML is based on the notion that while human thinking is quite complex, it might be just “good enough” to simulate thinking by providing "enough" response patterns to potential inquiries. Whether this minimalist approach will indeed be “good enough” is still hotly debated. Yet while this debate rages Alice and Chatterbots based on Alice and AIML have been winning the annual Loebner contest - a contest in which Chatterbots try to fool judges into believing they are human for the last several years.

Alice's knowledge and software support systems along with the AIML language are open source and supported by the non-profit Alice Foundation.

## **Zipf's Law and Alice's knowledge**

Before we get to ALICE, we need to visit another unusual figure in the history of computer science: Professor George Kingsley Zipf. Although he was a contemporary of Turing, there is no evidence the two ever met. Zipf died young too, at the age of 48, in 1950, only four years before Turing, but of natural causes.

There are many ways to state Zipf's Law but the simplest is procedural: Take all the words in a body of text, for example today's issue of the *New York Times*, and count the number of times each word appears. If the resulting histogram is sorted by rank, with the most frequently appearing word first, and so on ("a", "the", "for", "by", "and"...), then the shape of the curve is "Zipf curve" for that text. If the Zipf curve is plotted on a log-log scale, it appears as a straight line with a slope of -1.

The Zipf curve is a characteristic of human languages, and many other natural and human phenomena as well. Zipf noticed that the populations of cities followed a similar distribution. There are a few very large cities, a larger number of medium-sized ones, and a large number of small cities. If the cities, or the words of natural language, were randomly distributed, then the Zipf curve would be a flat horizontal line.

The Zipf curve was even known in the 19th century. The economist Pareto also noticed the log-rank property in studies of corporate wealth. One only need to consider the distribution of wealth among present-day computer companies to see the pattern. There is only one giant, Microsoft, followed by a number of large and medium-sized firms, and then a large tail of small and very small firms.

Zipf was independently wealthy. This is how he could afford to hire a room full of human "computers" to count words in newspapers and periodicals. Each "computer" would arrive at work and begin tallying the words and phrases directed by Zipf. These human computers found that Zipf's Law applies not only to words but also to phrases and whole sentences of language.

```
8024 YES
5184 NO
2268 OK
2006 WHY
1145 BYE
1101 HOW OLD ARE YOU
946 HI
934 HOW ARE YOU
846 WHAT
840 HELLO
663 GOOD
645 WHY NOT
584 OH
553 REALLY
544 YOU
531 WHAT IS YOUR NAME
525 COOL
516 I DO NOT KNOW
488 FUCK YOU
486 THANK YOU
416 SO
414 ME TOO
403 LOL
403 THANKS
381 NICE TO MEET YOU TOO
375 SORRY
374 ALICE
368 HI ALICE
```

366 OKAY  
353 WELL  
352 WHAT IS MY NAME  
349 WHERE DO YOU LIVE  
340 NOTHING  
309 I KNOW  
303 WHO ARE YOU  
300 NOPE  
297 SHUT UP  
296 I LOVE YOU  
288 SURE  
286 HELLO ALICE  
277 HOW  
262 WHAT DO YOU MEAN  
261 MAN  
251 WOW  
239 SMILE  
233 ME  
227 WHAT DO YOU LOOK LIKE  
224 I SEE  
223 HA  
218 HOW ARE YOU TODAY  
217 GOODBYE  
214 NO YOU DO NOT  
203 DO YOU  
201 WHERE ARE YOU  
.  
.  
.

The human input histogram, ranking the number of times ALICE receives each input phrase over a period of time, shows that human language is not random. The most common inputs are "YES" and "NO". The most common multiple-word input is "HOW OLD ARE YOU". This type of analysis which cost Dr. Zipf many hours of labor is now accomplished in a few milliseconds of computer time.

Considering the vast size of the set of things people could possibly say, that are grammatically correct or semantically meaningful, the number of things people actually do say is surprisingly small. Steven Pinker, in his book *How the Mind Works* wrote that:

Say you have ten choices for the first word to begin a sentence, ten choices for the second word (yielding 100 two-word beginnings), ten choices for the third word (yielding a thousand three-word beginnings), and so in. (Ten is in fact the approximate geometric mean of the number of word choices available at each point in assembling a grammatical and sensible sentence). A little arithmetic shows that the number of sentences of 20 words or less (not an unusual length) is about  $10^{20}$ .

Fortunately for chat robot programmers, Pinker's combinatorics are way off. Our experiments with ALICE indicate that the number of choices for the "first

word" is more than ten, but it is only about two thousand. Specifically, 1800 words covers 95% of all the first words input to ALICE. The number of choices for the second word is only about two. To be sure, there are some first words ("I" and "You" for example) that have many possible second words, but the overall average is just under two words. The average branching factor decreases with each successive word.

531 WHAT IS YOUR NAME  
352 WHAT IS MY NAME  
171 WHAT IS UP  
137 WHAT IS YOUR FAVORITE COLOR  
126 WHAT IS THE MEANING OF LIFE  
122 WHAT IS THAT  
102 WHAT IS YOUR FAVORITE MOVIE  
92 WHAT IS IT  
75 WHAT IS A BOTMASTER  
70 WHAT IS YOUR IQ  
59 WHAT IS REDUCTIONISM  
53 WHAT IS YOUR FAVORITE FOOD  
46 WHAT IS AIML  
38 WHAT IS YOUR FAVORITE BOOK  
37 WHAT IS THE TIME  
37 WHAT IS YOUR JOB  
34 WHAT IS YOUR FAVORITE SONG  
34 WHAT IS YOUR SIGN  
33 WHAT IS SEX  
32 WHAT IS YOUR REAL NAME  
30 WHAT IS NEW  
30 WHAT IS YOUR AGE  
30 WHAT IS YOUR GENDER  
28 WHAT IS YOUR LAST NAME  
27 WHAT IS HIS NAME  
27 WHAT IS YOUR SEX  
26 WHAT IS 2+2  
26 WHAT IS MY IP  
25 WHAT IS YOURS  
24 WHAT IS YOUR PURPOSE  
21 WHAT IS YOUR FAVORITE ANIMAL  
20 WHAT IS 1+1  
20 WHAT IS YOUR HOBBY  
19 WHAT IS THE WEATHER LIKE  
19 WHAT IS YOUR PHONE NUMBER  
18 WHAT IS ALICE  
18 WHAT IS GOING ON  
18 WHAT IS THAT SUPPOSED TO MEAN  
18 WHAT IS WHAT  
17 WHAT IS A SEEKER  
17 WHAT IS LOVE  
17 WHAT IS THE OPEN DIRECTORY  
17 WHAT IS YOUR FAVORITE TV SHOW  
16 WHAT IS JAVA  
16 WHAT IS THE ANSWER  
16 WHAT IS YOUR ANSWER  
16 WHAT IS YOUR FULL NAME

15 WHAT IS AI  
15 WHAT IS THAT MEAN  
15 WHAT IS THE WEATHER LIKE WHERE YOU ARE  
15 WHAT IS TWO PLUS TWO  
15 WHAT IS YOUR FAVORITE BAND  
14 WHAT IS CBR  
14 WHAT IS ELIZA  
14 WHAT IS GOD  
14 WHAT IS PI  
14 WHAT IS THE TURING GAME  
13 WHAT IS 2 + 2  
13 WHAT IS A COMPUTER YEAR  
13 WHAT IS IT LIKE  
13 WHAT IS MY FAVORITE COLOR  
12 WHAT IS 2 PLUS 2  
12 WHAT IS A CAR  
12 WHAT IS A DOG  
12 WHAT IS ARTIFICIAL INTELLIGENCE  
12 WHAT IS IT ABOUT  
12 WHAT IS LIFE  
12 WHAT IS SEEKER  
12 WHAT IS YOUR NAME  
12 WHAT IS YOUR FAVORITE  
12 WHAT IS YOUR SURNAME  
11 WHAT IS 1 + 1  
11 WHAT IS A CHATTERBOT  
11 WHAT IS A PRIORI  
11 WHAT IS SETL  
11 WHAT IS THE TIME IN USA  
11 WHAT IS THE WEATHER LIKE THERE  
11 WHAT IS YOUR FAVORITE FILM  
10 WHAT IS A CATEGORY C CLIENT  
10 WHAT IS A PENIS  
10 WHAT IS BOTMASTER  
10 WHAT IS MY IP ADDRESS  
10 WHAT IS THE DATE  
10 WHAT IS THIS  
10 WHAT IS YOUR ADDRESS  
10 WHAT IS YOUR FAVORITE MUSIC  
10 WHAT IS YOUR FAVORITE OPERA  
10 WHAT IS YOUR GOAL  
10 WHAT IS YOUR IP ADDRESS

Even subsets of natural language, like the example shown here of sentences starting with "WHAT IS", tend to have Zipf-like distributions. Natural language search Chatterbots like Ask Jeeves are based on pre-programmed responses to the most common types of search questions people ask.

## From Eliza to Alice

The story of Joseph Weizenbaum is in many ways almost as interesting as that of Turing. An early pioneer in computer science, Weizenbaum was one of the fortunate few to join the embryonic MIT Artificial Intelligence Lab in the early 1960s. His most celebrated accomplishment was the development of ELIZA, a program so entertaining that it still attracts clients to its web site today. ELIZA is based on very simple pattern recognition, based on a stimulus-response model.

ELIZA also introduced the personal pronoun transformations common to ALICE and many other programs. "Tell me what you think about me" is transformed by the robot into "You want me to tell you what I think about you?" creating a simple illusion of understanding.

Weizenbaum tells us that he was shocked by the experience of releasing ELIZA (also known as "Doctor") to the nontechnical staff at the MIT AI Lab. Secretaries and nontechnical administrative staff thought the machine was a "real" therapist, and spent hours revealing their personal problems to the program. When Weizenbaum informed his secretary that he, of course, had access to the logs of all the conversations, she reacted with outrage at this invasion of her privacy. Weizenbaum was shocked by this and similar incidents to find that such a simple program could so easily deceive a naive user into revealing personal information.

What Weizenbaum found specifically revolting was that the Doctor's patients actually believed the robot really understood their problems. They believed the robot therapist could help them in a constructive way. His reaction might be best understood like that of a western physician's disapproval of herbal medicines, or an astronomer's disdain for astrology. Obviously ELIZA touched something deep in the human experience, but not what its author intended.

From the back cover of *Computer Power and Human Reason* by Joseph Weizenbaum (1976):

"Dare I say it? This is the best book I have read on the impact of

computers on society, and on technology and man's image of himself."  
--- Keith Oatley, *Psychology Today*

"A thoughtful blend of insight, experience, anecdote, and passion that will stand for a long time as the definitive integration of technological and humanistic thought." --- *American Mathematical Monthly*

"Superb...The work of a man who is struggling with the utmost seriousness to save our humanity from the reductionist onslaught of one of the most prestigious, active, and richly funded technologies of our time." --- Theodore Pizsak, *The Nation*

Weizenbaum perceived his own program as a threat. This is a rare experience in the history of computer science. Nowadays it is hard to imagine anyone coming up with an original idea for a software program and saying, "no, this program is a dangerous genie and needs to be put back into the bottle." His first reaction was to shut down the early ELIZA program. His second reaction was to write a book about the whole experience, eventually published in 1972 as *Computer Power and Human Reason*.

*Computer Power and Human Reason* seems a bit quaint today, much the same as Turing's 1950 paper does. For one thing, Weizenbaum perceived his mission as partly to educate an uninformed public about computers. Presumably the uneducated public confused science fiction with reality. Thus most of *Computer Power* is devoted to explaining how a computer works: this is a disk drive, this is memory, this is a logic gate, and so on. In 1972 such a primer may have been necessary for the public, but today it might seem like the content for *Computers for Dummies*.

Two chapters of *Computer Power and Human Reason* are however devoted to an attack on artificial intelligence, on ELIZA specifically, and on computer science research in general. Weizenbaum is perhaps the stereotypical 1960's neo-Luddite. Not only would he slow down the pace of research, he would roll back the clock to a pre-computational era. One reviewer praises Weizenbaum for saving humanity from the "reductionist onslaught" of AI research, driven in those days by generous funds from the military-industrial complex.

Most contemporary researchers did not need much convincing that ELIZA was at best a gimmick, at worst a hoax, and in any case not a "serious" artificial intelligence project. The irony of Joseph Weizenbaum and *Computer Power and Human Reason* is that, by failing to promote his own technology, indeed by encouraging his own critics, he successfully blocked further investigation into what would prove to be one of the most promising and persistently interesting demonstrations to emerge from the early AI Lab.



## Chatterbots go Native – The Web Architecture

Versions of Alice became available to the public through a variety of internet-based programs years ago – yet until very recently, only people with extensive computer skills could host or modify these programs. The Alice Foundation refers to each of the Alice implementations by a capital letter. So, for example the Java version of Alice is known as Program D.

In 2002 a version of Alice that *anyone* – especially useful for non-computer-experts – can modify, develop and deploy became widely available at an experimental and free hosting site at [www.pandorabots.com](http://www.pandorabots.com). This version is a Lisp-based version known as Program Z (so named because it may be the *last* version anyone will ever need). The site [www.pandorabots.com](http://www.pandorabots.com) became operational on May 13, 2002. By January 5, 2003, the site had over 8100 registered botmasters.

Non-computer-literate people have built and deployed over 10,000 separate (and individually different and unique) versions of Alice. [www.pandorabots.com](http://www.pandorabots.com) supports the development and deployment of Chatterbots written in any international language (character set) along with sound and pictures. The URL's listed herein are examples of just two of these 10,000 Chatterbots. Chatterbots hosted on the site were generating between 2,000 and 20,000 interactions per hour. The number of new botmasters registering with the site was growing exponentially - and doubling approximately every 65 days.

Botmasters build and deploy Chatterbots using only their browsers. Botmasters are typically not computer programmers! Pandorabots provides facilities for creating and storing knowledge in the Pandorabots for the non-programmer. Pandorabots also offers virtual faces and speech facilities.

Few robots have won the Loebner Chatterbot contest as many times as Alice has. Alice is currently hosted at:

<http://www.pandorabots.com/pandora/talk?botid=f5d922d97e345aa1>

## The pandorabots implementation and architecture

The implementation is entirely done in Common Lisp and is hosted on Linux-based PC. This architecture was chosen for a variety of reasons, including:

- (1) The ability to change the system while it runs,
- (2) low-cost hosting systems, and
- (3) very low-cost software development and deployment costs.

The entire code base consists of less than 500k of source code. In January 2003, the system ran entirely on one PC with 1 gigabyte of ram and 100 gigabytes of disk. The system's cost was less than \$2000. This system was hosting the 8100 botmasters along with the 10,000 Chatterbots mentioned earlier.

On startup, the Pandorabots' code base requires about 20 megabytes, and grows slowly over time as botmasters add knowledge. The system has been down about 4 times since inception (May 13, 2002) - with the most significant downtime due to machine hardware upgrades. The system has been (and is still) under hacker attack each of which has been successfully repelled as the system can be reconfigured dynamically while continuing to run (which, incidentally is a reason for doing the system in Lisp). We estimate that a similar system done in Java would require 10 times the resources that were expended here - indeed, there are very real questions whether Pandorabots could ever be deployed in Java.

## Open Questions for multilingual use

The Japanese Language, unlike other Asian Languages creates an interesting problem for Chatterbot writers. For non-Japanese readers, below is the example of the use of distinct character sets presents the following dilemma. Each of the following inputs to Alice are equivalent.

わたしはにほんじんです  
私は日本人です  
ワタシハニホジンデス

[wastashi wa nihonjin desu]

*I am a Japanese.*

Even worse, more combinations exist, as arises when Hiragana and Kanji are intermixed - for example: わたしは日本人です is another version. In principal knowledge content authors will have to write AIML for each of the possible combinations. One approach might be to convert all input inquiries into one character set – like Hiragana, and to convert all input into Hiragana through a morphological analyzer. However, we will be losing semantic information associated with Kanji (Chinese characters) in this method. Another approach would be to simply rely on Zipf's law again and consider only what has been typed in across many interactions.

[www.pandorabots.com](http://www.pandorabots.com) converts input inquiries written in Japanese into an equivalent sentence with spaces inserted between “words” (morphemes) after morphological analysis as in:

私は日本人です → 私 は 日本人 です *I am a Japanese*  
Watashi (“I”) wa (Case-marker) Nihonjin (“Japanese”) Desu (Verb head “state of being”).

with the idea of making input inquiry patterns easier to recognize. Another problem lies in the construction of the knowledge bases. Direct language translation fails because the context of inquiries is so different. The English Language version of Alice contains knowledge of many abbreviations favored by users of American On-line Instant Messaging Systems. Lol means “laugh out loud” and these will have no equivalent in Japanese.

Problems arise with AIML – while AIML has been designed so that anyone can add knowledge – it is hobbled due to having arisen from the execrable (in our humble opinion) syntax of XML. While non-programmers use it, none appear to like it. A program called pandorawriter is available at [www.pandorabots.com](http://www.pandorabots.com) which converts sentences to AIML knowledge automatically – and much more could be done to simply knowledge building.

## **An Opinion on Alice Implementation Languages**

We are often asked why Common Lisp was chosen for the development. When we ask what other alternatives we might have chosen we often hear about the virtues of Java and .net.

Here is a quote from Ziff Davis News and Technology (November 15, 2002)

(<http://www.zdnet.com.au/newstech/enterprise/story/0,2000025001,20269968,00.htm>)

*To date, around 70 percent of initial Java implementations have been unsuccessful, according to new research from Gartner Group.*

*"An inordinately large number of large-scale Java projects have been failures," said Mark Driver, Gartner research director for Internet and ebusiness technologies.*

*However, Microsoft shouldn't draw any comfort from those figures as it seeks to promote its .NET technology strategy either. In all likelihood, the failure rate for early implementations of .NET systems will be similar, Driver said.*

*"The only practical way to mitigate the risk [of a failed implementation] is to outsource development."*

Why anyone would undertake to develop any significant internet-based application using either .net or Java will remain a mystery to us. Especially given the recent worldwide debacle and demise of E-commerce sites - over 4000 in the US alone - that were based on these technologies. This question that will undoubtedly be addressed by the pundits and other people who excel at describing the tendency for humans to move in herd-like-lemmings movements and propel themselves off of high-technology cliffs onto the rocky shoals below. Why Universities continue to promote the usefulness of Java or .net is an open question...

Some have suggested the lack of Lisp programmers hampers the development of such sites. And we find even this argument strange. Businesses thrive on the existence of sustainable competitive barriers. The fact that Lisp-based systems can be built at a cost of approximately a factor of 10 less than the same system using other popular technologies (which has only a small chance of being successfully deployed according the aforementioned Gartner Group Research) is about the best competitive barrier possible. Especially in light of the small number of Lisp programmers.

## References

Bogomolny A. *Benford's Law and Zipf's Law*  
[http://www.cut-the-not.com/do\\_you\\_know/zipfLaw.html](http://www.cut-the-not.com/do_you_know/zipfLaw.html) 2003

Pandorabots *About Pandorabots.com* <http://www.pandorabots.com> 2003

Nielsen J. *Zipf Curves and Website Popularity* Nielsen Norman Group  
<http://www.useit.com/alertbox/zipf.html> 2002

Wallece, R. *The Anatomy of ALICE* <http://www.alicebot.org/> 2003

Wentian Li, Center for Genomics and Human Genetics North Shore - LIJ Research Institute,  
<http://linkage.rockefeller.edu/wli/zipf/> 2002

Wentian Li Zipf's Law in Importance of Genes for Cancer Classification Using Microarray Data  
(UPF/IMIM, Barcelona, Spain, April 2002)

Zipf, GK *Psycho-Biology of Languages* (Houghton-Mifflin, 1935; MIT Press, 1965).

Zipf, GK *Human Behavior and the Principle of Least Effort* (Addison-Wesley, 1949).

Zipf, GK *National Unity and Disunity: The Nation As a Bio-Social Organism* (Principia Press,  
Bloomington Indiana, 1941).