Bachelor of Science in Computer Science
June 2017

# Generating Topic-Based Chatbot Responses


**Amandus Krantz**
**Petrus Lindblom**



Faculty of Computing
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

This thesis is submitted to the Faculty of Computing at Blekinge Institute of Technology in partial fulfillment of the requirements for the degree of Bachelor of Science in Computer Science. The thesis is equivalent to 10 weeks of full time studies.


**Contact Information:**
Author(s):
Amandus Krantz
E-mail: amkb14@student.bth.se
krantz.amandus@gmail.com

Petrus Lindblom
E-mail: peld14@student.bth.se
petrus.lindblom@hotmail.com

University advisor:
Dr. Prashant Goswami
Department of Creative Technologies

Faculty of Computing
Blekinge Institute of Technology
SE–371 79 Karlskrona, Sweden

Internet : www.bth.se
Phone : +46 455 38 50 00
Fax : +46 455 38 50 57

# Abstract

**Context**. With the rising popularity of chatbots, not just in entertainment but in e-commerce and online chat support, it's become increasingly important to be able to quickly set up chatbots that can respond to simple questions.
**Objectives**. This study examines which of two algorithms for automatic generation of chatbot knowledge bases, First Word Search or Most Significant Word Search, is able to generate the responses that are the most relevant to the topic of a question. It also examines how text corpora might be used as a source from which to generate chatbot knowledge bases.
**Methods**. Two chatbots were developed for this project, one for each of the two algorithms that are to be examined. The chatbots are evaluated through a survey where the participants are asked to choose which of the algorithms they thought chose the response that was most relevant to a question.
**Results**. Based on the survey we conclude that Most Significant Word Search is the algorithm that picks the most relevant responses.
**Conclusions**. Most Significant Word Search has a significantly higher chance of generating a response that is relevant to the topic. However, how well a text corpus works as a source for knowledge bases depends entirely on the quality and nature of the corpus. A corpus consisting of written dialogue is likely more suitable for conversion into a knowledge base.



**Keywords:** Chatbots; Corpora; AIML; A.L.I.C.E.

# Contents

# List of Figures

# List of Tables

# Chapter 1

# Introduction and Background

Many of us do not realize that our environment is changing rapidly and becoming increasingly automated to help us in our daily life. Virtual assistants such as Microsoft's Cortana [21], Google Assistant [13], and Apple's Siri [9] are just the beginning of an automated life. These virtual assistants can be regarded as the evolution of chatbots (also referred to as Chatterbots); computer programs that converse with humans through text or speech using natural language [1]. The aim of most chatbots is to trick the user into thinking they are interacting with a human. This is done by making the chatbot respond like a human. Even though they may respond like humans, they do not possess real intelligence. A chatbot simply decides on a response by looking for specific keywords and patterns; it does not understand what it is saying or what the user has said.

Though chatbots are mainly intended for entertainment, they have many other practical applications [11]. These can be things like e-commerce, education, healthcare [4], etc. A common application is online chat support. Companies don't want to spend money on hiring employees for menial tasks such as 24/7 chat support, so a chatbot that listens for specific keywords can be used instead. The bot can then provide the customer with information as best it can, and refer to a human representative if needed.

One of the first computer programs to simulate human-like conversation was ELIZA in 1966 [17]. By matching specific patterns and handwritten rules it appeared as if it understood the human language, even though it did not have any intelligence or reasoning capabilities. From 1995 and onward, a chatbot called A.L.I.C.E. (Artificial Linguistic Internet Computer Entity) [23] has been developed. Like ELIZA, A.L.I.C.E. is also based on pattern matching but uses AIML (Artificial Intelligence Markup Language) [26] for its conversation rules. AIML is a markup language that is used when creating knowledge bases for chatbots. It divides the knowledge base into smaller chunks, called categories. The categories contain patterns, the rules used to decide on a response, and templates, the rules that describe the response.

## 1.1 Background

This study can be regarded as a repeat of a study performed by Matilda Valleryd and Therese Askling in 2014 [20] where they examined two algorithms, First Word Search (FWS) and Most Significant Word Search (MSWS), in regards to how natural their responses are. FWS and MSWS are algorithms that can be used when automatically generating knowledge bases for chatbots and help the chatbot choose a response. FWS works by indexing the chatbot's knowledge base by the first word in the question. MSWS instead indexes by the word in the input that appears the least in the entire knowledge base [11].

Valleryd and Askling compared the algorithms through a survey. The survey consisted of a series of questions, along with responses to the questions from each algorithm. The participants had to choose which response they felt was more natural on a scale from 1 to 5, where 1 meant response A was a lot better and 5 meant response B was a lot better. The results from this survey were inconclusive, mostly because of a bad knowledge base. We intend to improve on their study by narrowing the scope and only focusing on the responses' relevancy, rather than on how natural they are. We will also use a different source for our knowledge base: A text corpus, which is a large library of organized texts that is generally used for linguistic research. We believe this will generate a knowledge base of sufficient quality as the structure of a corpora is similar to the structure of an AIML knowledge base. The contents of a corpus also comes from real-life examples of human dialogue, which should help make the responses more coherent.

Picking the most relevant response could be useful when creating chatbots for online support, as those kinds of bots don't have to be able to carry a conversation, but rather must be able to respond with information that is relevant to the user's questions and, if necessary, refer them to a representative that will be able to help further. This study will not examine the chatbot's ability to generate an appropriate response, instead it will investigate which of the algorithms will provide the response that is the most relevant to a question. If, for example, the question is "What color is the sun?", the response doesn't have to be "yellow" if it has something to do with the sun (See Figure 1.1).

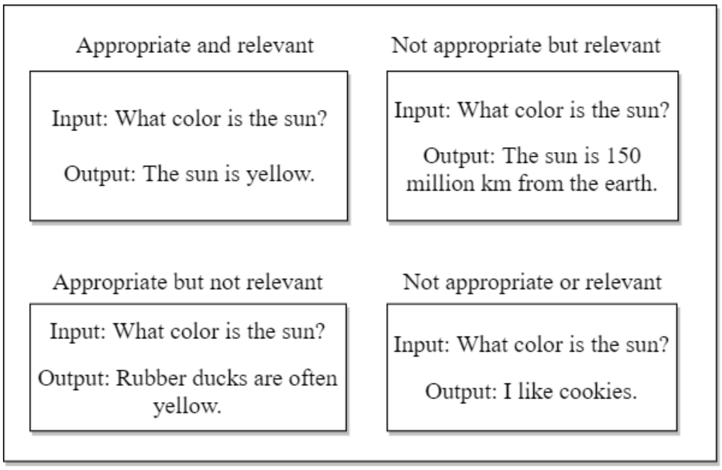



Figure 1.1: Example of the different types of responses to a question. The responses in the top row are acceptable answers.

## 1.2 Aim, Objectives, and Research Question

The aim of this study is to investigate which of two algorithms, FWS and MSWS, can generate the most relevant response to a question. To this end, the following research question has been defined: **Which algorithm, First Word Search or Most Significant Word Search, generates the most relevant responses in regards to the general topic of a user's question?**
The objectives that were planned for this project are the following:

- Convert a text corpus into an AIML knowledge base
- Modify a chatbot to use First Word Search and Most Significant Word Search
- Conduct a survey where responses from the algorithms are compared in regards to their relevancy to the topic of a questions
- Compare which of the two algorithms can generate the most relevant response

## 1.3 Method

To see which of the algorithms generates the most relevant responses, the responses from each algorithm are compared in a two-alternative forced choice survey. Participants are presented with a series of questions along with responses to the questions from each algorithm. They then have to choose which of the responses is the best fit to the topic of the question. The algorithm that gets the most votes should be the algorithm that picks the most relevant responses.

To test if the number of votes for each algorithm is significantly different from one another, a Student's t-test [14] is performed. The t-test was chosen over a chi-square test [27] as the chi-square test measures the relationship between observed and expected frequencies for some variable, while the t-test measures the difference between the means of two populations. We do not have any expected frequencies and are interested in the difference between the votes for FWS and MSWS, not their relationship, so the t-test was chosen.

If the t-test can not find a significant difference between the votes, the two populations cannot be assumed to always be different. This means there would be a risk that any perceived difference could be due to random noise in the data.

The test is performed with a significance level of 0.05. The null hypothesis is defined as there not being any significant difference between the number of votes, while the alternative hypothesis is that the number of votes are significantly different.

## 1.4 Thesis Overview

This thesis will answer which of the two algorithms, FWS or MSWS, generates the most relevant responses in regards to the general topic of a user's question. It is organized as follows: *Related Work* describes previous work done in the field, as well as the thesis by Valleryd and Askling that this project is based on. The *Methodology* chapter describes the theory and reasoning behind the algorithms and methods used in the project. *Results* describes the implementation of these algorithms and methods. It also gives an overview of the data gained from the conducted survey. *Analysis and Discussion* dives deeper into this data, discusses specific questions from the survey, and looks at some of the statistics surrounding the text corpus. *Future Work* gives some suggestions for work that could be done in future projects. This includes suggestions on how to improve the accuracy of the two algorithms, as well as how to use the information from this project to build a functioning chatbot. Finally, *Conclusions* provides a summary of the thesis and its results.

## Chapter 2

# Related Work

The two algorithms evaluated in this project, FWS and MSWS, are first described by Abu Shawar and Atwell in their 2007 article "Chatbots: Are they Really Useful?" [11]. In the article, they mainly investigate the many different applications for chatbots, however they also describe methods for automatically converting a text corpus to AIML along with the algorithms in this thesis. They argue that automated conversion is needed as the type of knowledge base that is normally used with chatbots need to be written by hand, which is a very time consuming process. Two methods are described for the automatic conversion between text corpora and AIML. The first is a simple one-to-one conversion where the first turn in the corpus is made into the pattern, and the second turn is made the template. This method is quite simple and does not generalize very well, which is why a second method was developed. The second method works just like the first, but allows for generalization of the knowledge base by using either FWS or MSWS. This is the method that is used in this project.

A comparison of the two algorithms in regards to how natural their responses are was done by Valleryd and Askling in their 2014 bachelor thesis [20]. Their results were inconclusive, largely because they had used a set of movie subtitles and scripts as the source for their chatbot's knowledge base. As per the authors, the amount of alterations that must be done to the subtitles to use them as a knowledge base is a source of error. This is because the more alterations that are done to a knowledge base, the worse it will be. A bad knowledge base can impact the results by generating incoherent responses.

In 2003, Abu Shawar and Atwell published the paper “Using dialogue corpora to train a chatbot” [10] in which they discuss what distinguishes A.L.I.C.E. from Elizabeth, two chatbots that originate from ELIZA. The discussion is centered around dialogue knowledge representation and pattern matching techniques. They highlight the problems that occur when converting the Dialogue Diversity Corpus (DDC) [28] to AIML. A structural standard of the dialog in the corpora is proposed. They suggest that the corpus should have characteristics such as two speakers, a clear and structured format, and preferably short sentences that clearly show whose turn it is to speak. Their conclusion is that the differences between dialogues in DDC are problematic, mainly in markup and annotation practices. To counter this, they urge the dialogue corpus research community to agree to, and set, a suitable standard.

# Chapter 3

# Methodology

## 3.1 AIML

AIML is a markup language created by Dr. Richard S. Wallace to use when creating knowledge bases for chatbots [26] [19]. It works by dividing the chatbot's knowledge base into small sections, called categories. These categories contain two tags: patterns and templates (See Figure 3.1). The patterns are rules by which the chatbot can figure out what its response should be, while the templates describe what that response should look like. The way these rules are built varies depending on the size and complexity of the knowledge base. They can look like regular sentences, or they can be slightly more complex, containing wildcards which lets the chatbot learn things like names of people and places or ages.

```
<category>
   <pattern>HI THERE</pattern>
   <template>Hello! What's your name?</template>
</category>
```

Figure 3.1: A basic AIML category

### 3.1.1 Interpreting AIML

To use the AIML knowledge base, an AIML interpreter is needed. These interpreters take input from the user and searches through the knowledge base for a pattern that fits the input based on some algorithm. The most common one being the Graphmaster that was developed together with A.L.I.C.E. and AIML [24].

The way Graphmaster operates can be likened to the way a user navigates a basic filesystem [25]. In its most basic form, a filesystem contains files and folders. When a user navigates to a specific file in their filesystem, "/home-/alice/pictures/bots.jpg" for example, they start in the root of their filesystem, navigate through a series of folders, until they find the file they are looking for. This is how Graphmaster operates. The root is the beginning of the knowledge base, the folders are each word in the pattern, and the file at the end is the template that describes the response to the pattern. The user input "HELLO MY NAME IS ELIZA" could thus be written as "/HELLO/MY/NAME/IS/ELIZA-/response.template", where "response.template" is the template. Just like how every file path that begins with "/home" would start in the "home" folder, every pattern that starts with "HELLO" starts in the "HELLO" category.

There is one difference between Graphmaster and a file system: rather than just stopping if it can't find an exact match, like a file system, it will try to find the longest possible pattern match. That means that if the knowledge base only contains "/HELLO/MY/NAME/", the template in the "NAME" folder will be chosen as the response.

### 3.1.2 The Interpreter

For this project, a modified version of Program Y [18], an AIML interpreter written in Python, is used. It was chosen as Python allows us to use the NLTK (Natural Language Toolkit) [2] platform's tools to find the frequency distribution of the words in the AIML knowledge base. Program Y originally uses the standard AIML Graphmaster to choose responses from an AIML knowledge base. It has been modified to use either the first word in a sentence or the word in the sentence that appears least in the knowledge base.

## 3.2 Text Corpora

To automatically create a chatbot, a source of dialogue is needed to base responses on. This is more commonly known as a knowledge base. A good source for these large quantities of dialogue is a text corpus. The contents of a corpus can either come from spoken or written sources. The main purpose of a text corpus is to be used for linguistic research. However, it can also be used as a data source for Natural Language Processing or, in our case, as the source for a chatbot's knowledge base where it has the added benefit of not requiring too many alterations.

The text corpus used in this project is the OANC (Open American National Corpus) [8], which contains roughly 15 million words from both spoken and written sources. The main reason for using OANC is the fact that it includes tags that indicate who is currently speaking. This is useful as it lets us easily differentiate between the utterances in the corpus and decide what is being said by who.

### 3.2.1 Converting a Text Corpus

The conversion to AIML begins by converting the text corpus from raw text into XML using the tool that is provided by the ANCP (American National Corpus Project) [6]. This is done as the original corpus only contains text and no tags for sentence boundaries or whose turn it is to speak, both of which are needed for an easy conversion process (See Figure 3.2). Converting to XML not only provides these tags, it also makes the corpus easier to parse .

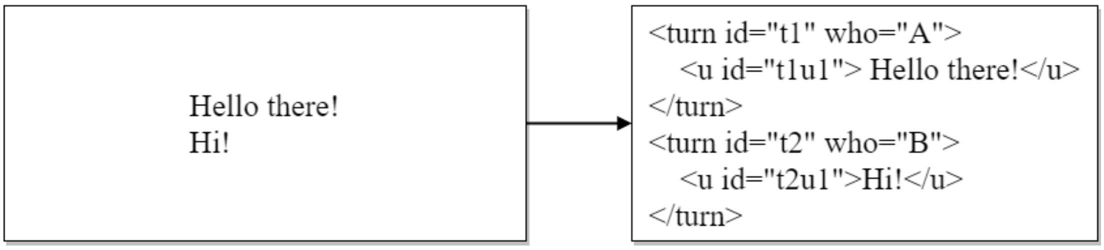



Figure 3.2: Conversion from text in a corpus to an XML-based corpus.

The turns are used to identify whose turn it is to speak. This lets us distinguish between each utterance in the corpus, which is important when creating pattern-template pairs. If this tag was not present, there is a chance that a pattern-template pair might be generated from two utterances by the same person. The sentence boundaries let us quickly determine when each sentence in an utterance begins and ends. This is not as important as the turn-tags, but can be useful when combining multiple utterances by the same speaker.

The conversion from the XML-based corpus to AIML is done by picking one of the speakers in a conversation and making every utterance by that speaker into patterns. The other speaker in the conversation is made into corresponding templates. Pairs of these patterns and templates are then put into AIML categories.

Before the conversion into AIML is complete, one last thing is done: normalizing the patterns in the knowledge base to fit the AIML conventions. To do this, each letter in the pattern must be changed to capital letters and all punctuation must be stripped out (See Figure 3.3). This makes it easier for the chatbot to read the patterns at run-time. None of these alterations are done on the templates, which keeps the tone of the responses intact. Removing punctuation from the templates can also result in clumsy and hard-to-read responses.

```
<turn id="t1" who="A">
   <u id="t1u1"> Hello there!</u>
</turn>
<turn id="t2" who="B">
   <u id="t2u1">Hi!</u>
</turn>
```

```
<category>
   <pattern>HELLO THERE</pattern>
   <template>Hi!</template>
</category>
```

Figure 3.3: Conversion from an XML-based corpus to AIML. The first turn (Hello there!) is made into the pattern, the second turn (Hi!) is made into the template.

A portion of the corpus consists of text from written sources such as books or news articles. These cannot be converted into AIML as there is no dialogue, and must be removed. Furthermore, some parts of the spoken part of the corpus do not contain exactly two speakers. These parts cannot easily be converted into AIML as it is important to make sure the patterns and responses are from different people to keep the question-response format intact. Adding speakers interferes with this pattern as we rely on the assumption that the turn after the question is a response to the question. The introduction of another speaker means there is a chance they will interject something before the response has been made, which offsets the format. Removing a speaker means there is no dialogue at all, just one person talking.

There are some cases where the question or response only consists of words such as "uh-huh" or "um". These filler words are not of use to the chatbot and would not work as good responses if chosen and are therefore removed. When all removals are done, about 3 million words are left to train on.

## 3.3 First Word Search

FWS generates its response based on the first word of the user's input. This is done by iterating through every pattern in the AIML knowledge base and removing everything but the first word. After this has been done, every template that is indexed to the same first word is collected and placed under the same pattern. Each pattern is then placed into its own category (See Figure 3.4).

When the chatbot picks a response, it removes everything but the first word in the user's input and then searches through the categories in the FWS-indexed AIML knowledge base for a pattern that matches it. Once it has found the correct pattern it uses one of its templates as a response. If there is more than one template, it will choose one at random.

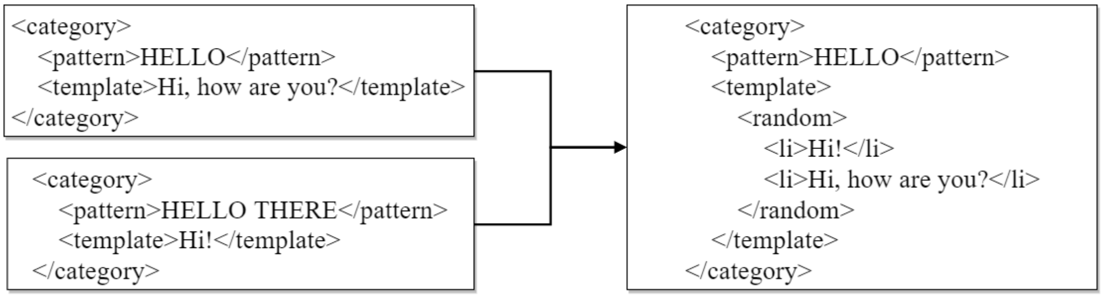



Figure 3.4: Conversion from AIML to FWS-indexed AIML. Two categories with the same first word (HELLO) have been collected into the same category.

## 3.4 Most Significant Word Search

MSWS can be misleading as a keyword is not said to be significant if it is a common word. The significance of the keyword is instead determined by how likely it is to lead the chatbot to a response that will fit the current question [11]. The likeliness that a word will direct the chatbot to the most fitting response is calculated by looking at its frequency in the knowledge base. The less frequent a word is, the more likely it is to point to a very specific set of responses. For example, words like "I", "AND", "THE", and "YOU" occur very frequently in standard English. Using these as the most significant word would likely result in thousands upon thousands of responses, which would all have thousands of different contexts and meanings. On the other hand, a word like "DOG" does not occur as often in a normal conversation and should not result in an enormous amount of responses. The few responses that would be found for the word "DOG" are also likely responses to have been uttered in conversations about dogs, thus increasing the chance that the response will fit the current conversation.

The conversion from standard AIML to MSWS is done by first getting the frequency distribution of every word in the entire AIML knowledge base. Every pattern in the knowledge base is then compared against that frequency distribution, and the word in the patterns that has the lowest frequency is chosen to replace the pattern. Once the least frequent word has been found, all templates indexed to the same word are collected and placed under the same pattern. The patterns are then placed into categories (See Figure 3.5).

When the chatbot later picks a response, it goes through a very similar process. It begins by fetching the same frequency distribution that was used in the conversion to MSWS-indexed AIML, and then finds the least frequent word in the user's input. The MSWS-indexed AIML knowledge base is then searched through to find a pattern with that word. If the pattern that is found has more than one template to choose from, a template is chosen at random.

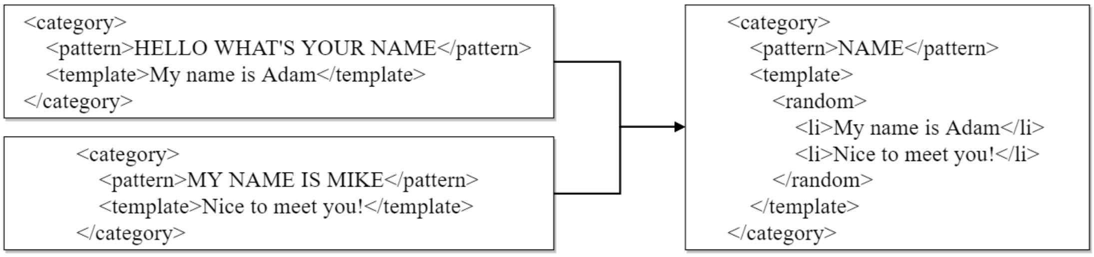
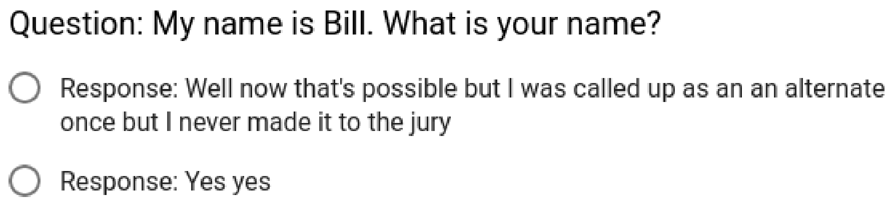



Figure 3.5: Conversion from AIML to MSWS-indexed AIML. Two categories with the same most significant word (NAME) have been collected into the same category.

## 3.5 Evaluating the Algorithms

One of the most common ways to evaluate the performance of an AI or chatbot is the Turing test. It was first described in 1950 by Alan Turing in his paper "Computing Machinery and Intelligence" [5] as "The Imitation Game". The test is done by having a person, called the interrogator, converse with two entities that he or she cannot see. One of the hidden entities is a machine, called A, and the other is a person, called B. The interrogator's objective is to figure out which of the entities, A or B, is the person by asking them a series of questions. The objective of the machine is to fool the interrogator into choosing the wrong entity. The problem with this test is that it can be very resource intensive. Because of this, another method was developed by the chatbot community. The chatbot that is to be evaluated is asked several questions, the questions and their respective responses are then put into a survey where participants rank each response on a Likert scale [22].

The evaluation method used in this project is like the method that was developed by the chatbot community, but some alterations have been made to make it fit the circumstances [12]. As we have two chatbots to evaluate, a Likert scale cannot be used, instead two-alternative forced choice (2AFC) is used. The resulting survey consists of 31 pre-generated questions with two answers for each, one from each version of the chatbot (See Figure 3.6). The participants are asked to choose which of the answers they think have the best fit to the topic of the question. To avoid confirmation bias, the order of the responses to each question was randomized. The questions were taken from the selection rounds of the 2010 Loebner Prize contest [16] [15], and example questions from the Robochat Challenge [3], two international competitions where chatbots compete by trying to pass the Turing test. A list of these questions, along with responses from each algorithm, can be found in Appendix A.


Question: My name is Bill. What is your name?

Response: Well now that's possible but I was called up as an an alternate once but I never made it to the jury

Response: Yes yes

Figure 3.6: An example of a question in the survey. The first response is from First Word Search. The second response is from Most Significant Word Search.

The responses were slightly curated to remove filler words like “uh” and “um”. Some punctuation was also added to make them easier to read for the participants. The contents or topics of the responses were never changed.

To see if technical knowledge and experience with chatbots impact which response is chosen, the participants were asked to estimate their technical skill and knowledge on a scale from one to five before taking the survey.

# Chapter 4

# Results

The survey was sent out to 36 people. Google Forms was used as it provides an easy and free platform for conducting surveys. Out of the 36 people it was sent out to, 30 responded. 24 (83%) of the participants were Swedish, while the remaining 6 (17%) were American. It should be noted that the Americans might have a slightly different interpretation of the responses than the Swedes. This is because the responses were provided in English, which Americans should have a slightly better grasp of than Swedes. We have chosen to ignore these differences arising from cultural contexts of the language.

Looking at the results for the questions in table 4.1, it becomes clear that MSWS is the winner when it comes to finding the topic of a user's input, with MSWS being chosen 60.86% of the time.

| **Algorithm** | **Total number of votes** |
|---|---|
| FWS | 364 |
| MSWS | 566 |

Table 4.1: Total number of votes for each of the algorithms. For number of votes per response, see appendix A.

A Student's t-test was performed at a significance level of 0.05 to test for significant difference between the number of votes for each algorithm. The null hypothesis is defined as there being no significant difference to the results.

The test returns a p-value of 0.000133, which means there's a 0.0133% risk that there is no significant difference between the number of votes for each algorithm. 0.000133 is lower than the significance level, which means we can safely reject the null hypothesis.

## 4.1 Implementation of the Algorithms

The most demanding part of the project was the implementation of the corpus-to-AIML conversion tool. It was written in Python, which allowed us to use NLTK's frequency distributions when deciding the most significant word.

When the conversion begins, we first loop through every file in the corpus and pair each turn together in pairs of two. Every turn comes divided into its sentences, which means it's necessary to iterate through each sentence and collect them into the same paragraph. Once the sentences have been collected, they are placed into AIML categories. The first chronological turn is made into the pattern of the category, and the second is made into the template. The pattern and template are then curated to remove any extraneous line breaks and spaces. Finally, the pattern is changed to capital letters and the entire category is written to an AIML-file. When this process is done, the entire corpus has been converted into a basic pattern-template AIML knowledge base.

Converting the basic AIML knowledge base to FWS-indexed AIML begins by iterating through every category in the basic knowledge base. Each pattern is stripped of everything but the first word. During this pass, filler words such as "uh-huh", "um-hum", and "um" are filtered out. When this pass is complete, a second pass is done where every pattern that consists of the same word is collected. The templates are collected in a list using AIML's random-tag, which tells the chatbot to choose a response at random. This results in an FWS-indexed AIML knowledge base.

The conversion from basic AIML to MSWS-indexed AIML works similarly to the AIML-to-FWS conversion. Two passes of the knowledge base are made. The first pass filters out filler words and strips the patterns of everything but the word that appears the least in the frequency distribution of the entire knowledge base. The second pass collects every pattern with the same word to one category and, using AIML's random-tag, collects their templates into a list for the chatbot to randomly choose a response from.

Some modifications had to be done to Program Y [18], the Python AIML interpreter used in this project. Rather than completely replacing Graphmaster, it is slightly modified to remove everything but one word from the user's input. The word that is left is decided by methods that are quite similar to the algorithms used to generate FWS-indexed and MSWS-indexed AIML knowledge bases. For FWS, the user's inputs are stripped of everything but the first word. With MSWS, every word in the user's input is compared against the frequency distribution of words in the AIML knowledge base. Everything but the word that appears the least is removed from the user's input. The words are then sent through the Graphmaster, which tries to use its pattern matching to find the best fit for the input. However, as there is only one word in the input, and the knowledge base is indexed by one-word patterns, it will find the pattern that contains that word and choose one of its responses.

# Chapter 5

# Analysis and Discussion

The results from the survey show MSWS as the clear winner with 60% of the votes. This likely means that the word that is the most indicative of the topic of a sentence is the word that is the most specific to the sentence, or the word in the sentence that has the lowest frequency in its language.

## 5.1 Technical Skill

The results from the question about technical skill and knowledge turned out to be difficult to interpret as a clear bias emerged. As evidenced in figure 5.1, a clear majority of the subjects ranked their technical skill and knowledge above average. People are either reluctant to rank themselves as average or under average, indicating an egocentric bias, or the people selected for the survey had a higher chance of having a high technical skill and knowledge, indicating a selection bias. In any case, since most of the data gained from the survey comes from participants who estimate their technical skill as average or above average, we do not have a sufficient distribution of technical skill to determine whether technical knowledge affects the selection of response.

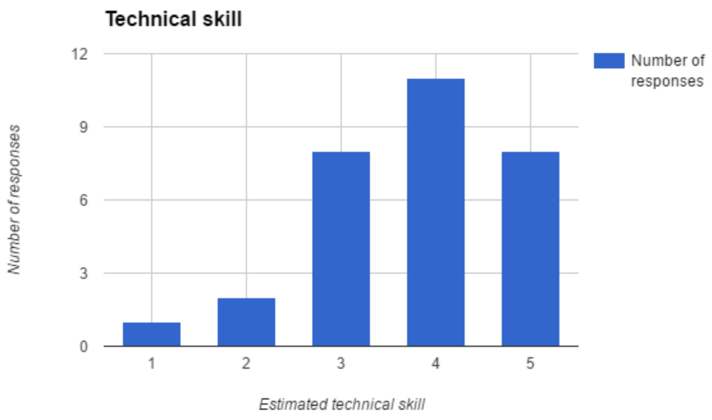



Figure 5.1: Distribution of technical skill as estimated by survey participants.

## 5.2 Open American National Corpus

The un-modified OANC contains roughly 14.6 million words, however 11.4 million of these come from written sources such as 911 reports and travel guides [7]. Those sources rarely contain any dialogue, which means they are not useful for this project and must be removed. This leaves us with 3.2 million words from only spoken sources such as face-to-face interviews and telephone conversations. However, some of these conversations and interviews are not usable. They might contain too many or too few speakers, which means we cannot create pattern-template pairs, so the interview or conversation must be removed. The exact number of interviews and conversations that were discarded is unknown, however 182 222 words were removed, leaving us with a total of 3 035 550 words to train the chatbot on.

Of the roughly 3 million words in the knowledge base, there are 27149 unique words, the most common of which are words like “I”, “AND”, “THE”, and “YOU”. These are the building blocks of most sentences in English, which is evident when looking at the frequency distribution for the knowledge base in table 5.1.

| **Word** | **Frequency** |
|---|---|
| I | 122819 |
| AND | 102658 |
| THE | 90600 |
| YOU | 78138 |
| THAT | 71625 |
| A | 67233 |
| TO | 65930 |
| IT | 65012 |
| OF | 51202 |
| KNOW | 44089 |

Table 5.1: The ten most common word in the knowledge base. The higher the frequency number, the more common the word is.

These words will most likely never be chosen by MSWS as it looks for the words with the lowest frequency, rather than words with the highest frequency. There are exceptions of course; if the input was “I KNOW YOU”, the 10th most common word in the knowledge base, “KNOW”, would be chosen.

The ten most common first words (See Table 5.2) are words like "I", "AND", "WELL", and "SO". These are words that are almost always used when beginning a sentence. The more common a first word is, the more responses there are to choose from. This isn't necessarily a good thing, especially when the number of possible responses are in the thousands. When the variance in possible responses gets too high, the chance of choosing a response that fits the current topic is drastically decreased.

| **Word** | **Frequency** |
|---|---|
| I | 5461 |
| AND | 5133 |
| WELL | 4514 |
| SO | 2491 |
| BUT | 2240 |
| THATS | 2147 |
| YOU | 1981 |
| OKAY | 1767 |
| NO | 1304 |
| YES | 1256 |

Table 5.2: The ten most common first words in the FWS-indexed knowledge base. The higher the frequency number, the more common the word is.

## 5.3 Observations from the Survey

In addition to the results, special situations were observed where participants might have made a choice without any reasoning.

In some cases, the number of votes for each algorithm were almost equal, such as in question 8 (See Figure 5.2). Both of the responses generated by the algorithms were bad and did not fit the topic of the question. This likely caused the participants to choose a response at random, or to factor in unrelated information when making their decision.

Another scenario that was observed was when both algorithms generated a response that was fine. An example of this is question 15 (See Figure 5.3); MSWS chose "My Mexican restaurant" as its response, which is directly relevant to the topic of the question. The response chosen by FWS does not directly mention food, but can still be considered an acceptable response as "I have never tried any" could be interpreted as "I have never tried any food". That response was one of 691 available responses for the word "WHAT" which means the chance that the response would fit the question is low enough that we cannot depend on it to always fit.

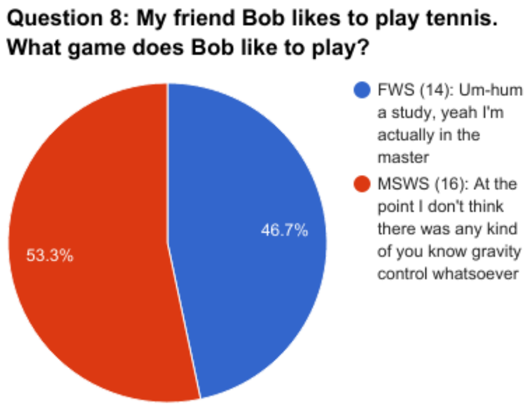



Figure 5.2: Distribution of votes for question 8 in the survey. For more information see appendix A.

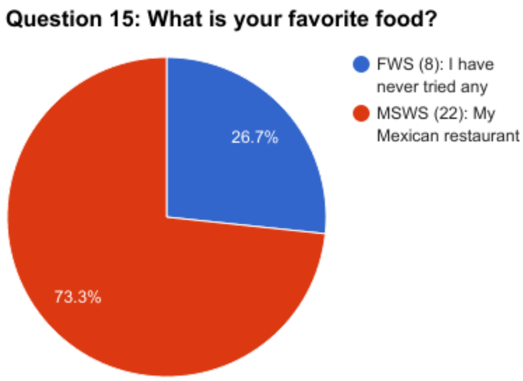



Figure 5.3: Distribution of votes for question 15 in the survey. For more information see appendix A.

It is also possible that some participants just skimmed the questions and responses without paying attention to their contents or topics. One likely example of this is question 21 (See Figure 5.4). The response generated by MSWS is "Brown and Black", a perfect fit both in regards to the topic and as an answer to the question. Yet 10% of participants chose the response generated by FWS, a response that, while appropriate, does not fit the topic of the question as well as the one chosen by MSWS. It is reasonable to assume that other questions have been affected by this as well.

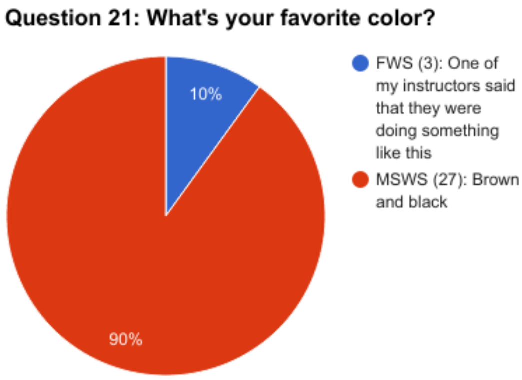



Figure 5.4: Distribution of votes for question 21 in the survey. For more information see appendix A.

## 5.4 Breakpoints

In both algorithms, each pattern in the AIML knowledge base has at least one template. In cases where there are more than one template the chatbot will randomly choose a template as a response. The response that is chosen is thereby partly determined by chance. This means it is possible that the chatbot will give wildly different responses to the same question.

Since the text corpus contains several hundreds of conversations between many different people, the chatbot's responses will not have a consistent personality. Our chatbot is, for all intents and purposes, a parrot: unable to understand what it's saying, with no ability to improvise, and only able to blindly repeat sentences it has heard before.

Many of the participants pointed out that the chatbot's responses did not fit the questions and asked why algorithms that give such poor responses were chosen. This suggests that we were not clear enough that the study mainly looks for relevancy in the responses, not necessarily a good fit for the questions. It also suggests that the participants might have been unable to separate the responses from the topics. A better way to do the survey, which could have prevented this, might be to discard the responses and instead focus solely on the topics that the algorithms chose. Instead of choosing between two responses, the participants would be given two words, one from each algorithm, and be asked to choose which of the words better represents the topic in the question. We believe that MSWS would win over FWS by a much larger margin in a situation like this, as the words selected by FWS are mostly words like "THAT", "WHAT", and "I" while MSWS is more likely to pick a word that is more specific to the question.

# Chapter 6

# Future Work

The chatbots created in this project are not able to have a proper conversation. Creating a chatbot that can do this might be something that a future project would want to consider. To do this, a handwritten AIML knowledge base, tailored to the MSWS algorithm, could be created. This would require a deep understanding of AIML. A good place to start would be to figure out how wildcards and topics could be used in conjunction with MSWS to create a knowledge base that is better at selecting the most appropriate response than just standard MSWS.

Improvements to FWS could also be made. For example, a list of banned words like "I", "YOU", "WHAT", etc. could be added. These are words that are too common to add any useful information to find a relevant response and result in patterns with several thousand responses.

Another idea is to write a more sophisticated corpus-to-AIML conversion tool. Rather than just doing a straight conversion, things like names could be replaced with wildcards. MSWS could even be used to augment the conversion process by finding the topic of the sentences and creating topic-tags with that topic.

The criteria suggested by Abu Shawar and Atwell [10] are very barebones and are not enough to create a high-quality knowledge base. More in-depth research into what exactly is required for a good corpus-to-AIML conversion could be done, and a larger set of criteria could be suggested. The project could consider things like the maximum number of sentences per response, or which type of dialogue is the most useful.

If the goal is to develop a chatbot that is relevant and appropriate, there is room for further research. We propose to use MSWS plus an additional method to figure out how appropriate the responses are in relation to the given question. One way of doing this could be to use the most significant word combined with the second most significant word. Choosing a response based on two words instead of one should increase the chatbots accuracy when it comes to choosing appropriate responses. Using the sun example from earlier, double MSWS would first select the word “SUN” from the input. It would then do a second pass of the input and select the second most significant word: “COLOR”. The chatbot would then search through the knowledge base for patterns that only contains those two words. A fallback could be created where, if a pattern with the two words does not exist, the chatbot would search for patterns that contain either of the words.

# Chapter 7

# Conclusions

This study set out to examine two algorithms that can be used during automatic generation of knowledge bases for chatbots. The two algorithms are called First Word Search and Most Significant Word Search. The algorithms were compared to see which of them would generate the response that was the most relevant to the topic of a question. The responses didn't necessarily have to be appropriate responses, if they were about the same topic as the question. This was examined by conducting a two-alternative forced choice survey were the participants were presented with a series of questions. Each question had two responses, one from each algorithm. The participants were then asked to choose the response that best fit the topic of the question.

The results from the survey show Most Significant Word as the clear winner, getting 60% of the votes. This contrasts with the previous study by Askling and Valleryd who were unable to achieve conclusive results. The two main causes for this is the improved knowledge base used in this study, as well as a narrower scope. Rather than comparing how natural the responses are, we compared their relevancy. The results indicate that the best way to find the topic of a question is to look for the word that is the most specific to the question. The first word, on the other hand, does not appear to be a good indicator of the topic of a sentence, likely because most sentences in a standard English conversation begin with words like "THE", "I", and "SO", none of which are not very specific or indicative of the topic.

Furthermore, we do not believe that the criteria suggested by Atwell and Abu Shawar in their 2003 article [10] are enough for a corpus to be converted into a high-quality AIML knowledge base. The criteria are that the corpus should have two speakers, a structured format, and short obvious turns. The corpus used in this project meets all the criteria, but the resulting chatbot is still prone to inappropriate and incoherent responses.

While converting a corpus that fits these criteria does result in an easier conversion process, the quality of the knowledge base is still very much dependent on the quality of the corpus. Atwell and Abu Shawar point out that if the dialogue in the corpus is bad, the conversation with the chatbot will be bad. We suggest adding a fourth criteria; the corpus should be made up of written dialogue. This criterion comes from the fact that verbal and written conversations tend to be very different when it comes to things like sentence structure and choice of words.

# References


[1] Chatbot - Definition of Chatbot in English | Oxford Dictionaries. https://en.oxforddictionaries.com/definition/chatbot. [Online; retrieved: 2017-05-11].

[2] Natural Language Toolkit. http://www.nltk.org/. [Online; retrieved: 2017-04-24].

[3] Robo Chat Challenge. http://www.robochatchallenge.com/scoring.html. [Online; retrieved: 2017-04-24].

[4] Abbas Saliimi Lokman and Jasni Mohamad Zain. Designing a Chatbot for Diabetic Patients. In *International Conference on Software Engineering and Computer Systems*, pages 19–21, Swiss Garden Kuantan Pahang, 2009.

[5] Alan Turing. Computing Machinery and Intelligence. *Mind*, 59(236):433–460, 1950.

[6] ANCP. ANC Tool. http://www.anc.org/software/anc-tool/. [Online; retrieved: 2017-04-24].

[7] ANCP. Contents | OANC. http://www.anc.org/data/oanc/contents/. [Online; retrieved: 2017-05-09].

[8] ANCP. Open American National Corpus. http://www.anc.org/. [Online; retrieved: 2017-03-01].

[9] Apple. Siri. http://www.apple.com/ios/siri/. [Online; retrieved: 2017-05-14].

[10] Bayan Abu Shawar and Eric Atwell. Using Dialogue Corpora to Train a Chatbot. In *Proceedings of the Corpus Linguistics 2003 Conference*, pages 681–690, Lancaster, UK, 2003.

[11] Bayan Abu Shawar and Eric Atwell. Chatbots: Are They Really Useful? In *LDV Forum*, volume 22, pages 29–49, 2007.

[12] Bayan Abu Shawar and Eric Atwell. Different Measurements Metrics to Evaluate a Chatbot System. In *Proceedings of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, NAACL-HLT-Dialog '07, pages 89–96, Stroudsburg, PA, USA, 2007. Association for Computational Linguistics.

[13] Google. Google Assistant. `https://assistant.google.com/`. [Online; retrieved: 2017-05-14].

[14] Winston Haynes. *Student's t-Test*, pages 2023–2025. Springer New York, New York, NY, 2013.

[15] Hugh Loebner. 2010 Loebner Price Selection Process. `http://loebner.exeter.ac.uk/selection-process/`. [Online; retrieved: 2017-04-24].

[16] Hugh Loebner. The Loebner Prize in Artificial Intelligence. `http://www.loebner.net/Prizef/loebner-prize.html`. [Online; retrieved: 2017-05-16].

[17] Joseph Weizenbaum. ELIZA - a Computer Program for the Study of Natural Language Communication Between Man and Machine. *Communications of the ACM*, 9(1):36–45, January 1966.

[18] keiffster. program-y. `https://github.com/keiffster/program-y`. [Online; retrieved: 2017-04-24].

[19] Maria das Graças Bruno Marietto, Rafael Varago de Aguiar, Gislene de Oliveira Barbosa, Wagner Tanaka Botelho, Edson Pimentel, Robson dos Santos França, and Vera Lúcia da Silva. Artificial Intelligence Markup Language: A Brief Tutorial. *International Journal of Computer science and Engineering Survey*, July 2013. arXiv: 1307.3091.

[20] Matilda Valleryd and Therese Askling. *Chatbots and natural automated language: A comparison between first word and most significant word search*. B.s. thesis, KTH, Stockholm, 2014.

[21] Microsoft. Cortana. `https://www.microsoft.com/en/mobile/experiences/cortana/`. [Online; retrieved: 2017-05-14].

[22] Rensis Likert. A Technique for the Measurement of Attitudes. *Archives of Psychology*, 140:1–55, 1932.

[23] Richard S. Wallace. A. L. I. C. E. The Arificial Linguistic Internet Computer Entity. `http://www.alicebot.org/about.html`. [Online; retrieved: 2017-04-24].

[24] Richard S. Wallace. AIML Graphmaster. http://www.alicebot.org/documentation/matching.html. [Online; retrieved: 2017-04-24].

[25] Richard S. Wallace. AIML Graphmaster - Filesystem metaphore. http://www.alicebot.org/documentation/matching.html#filesystem. [Online; retrieved: 2017-05-09].

[26] Richard S. Wallace. AIML Overview. https://www.pandorabots.com/pandora/pics/wallaceaimltutorial.html. [Online; retrieved: 2017-04-24].

[27] Ronald J. Tallarida and Rodney B. Murray. *Chi-Square Test*, pages 140–142. Springer New York, New York, NY, 1987.

[28] William C. Mann. Dialogue Diversity Corpus. http://www-bcf.usc.edu/~billmann/diversity/DDivers-site.htm, September 2003. [Online; retrieved: 2017-05-09].

# Appendix A

# Survey Questions

The number of votes for each response is listed in parentheses after the algorithm.

1. My name is Bill. What is your name?
   - **FWS (13):** Well now that's possible but I was called up as an an alternate once but I never made it to the jury
   - **MSWS (17):** Yes yes
2. Which is larger, a pineapple or a pine tree?
   - **FWS (8):** Sure
   - **MSWS (22):** Not too bad, it's about I guess half and a half pine trees and various hard woods of you know
3. What is a calendar?
   - **FWS (10):** Christine Tammy Marie Watkins. I lay my head down on the desk and cried because I had to learn how to write a long name and everyone else's was short. I'd been writing "Tammy Watkins." I could read and write before I went to school, because Mama had a lot of time to spend with me. My next sister wasn't born until I was like five years old. [Long pause] The teachers were upset because I was way ahead of everybody else. They didn't know what to do with me.
   - **MSWS (20):** Well just like that air bag I think that thing is fantastic cause I've seen some of the you know like the head on collision type things where they had it
4. What is my name?
   - **FWS (3):** We do not watch a lot of TV, we try to watch it on Thursday nights we usually watch LA Law i like that real well
   - **MSWS (27):** I can't think of it
5. John is taller than Mary and Mary is taller than Sue. Who is shorter, John or Sue?

- **FWS (22):** I am, what was the first part of the subject that you
- **MSWS (8):** You know it's like yeah that too, that's true

6. What day of the week is it?

   - **FWS (12):** Well I'm a very big favor uh much in favor of red wines and i have about a thousand bottles down in my cellar and i've been collecting it now for oh fifteen years or so
   - **MSWS (18):** We're just going through, well there's a couple of changes to the our health plan and I guess right now the one in my mind is health insurance

7. What will be tomorrow's date?

   - **FWS (12):** What a diet oh no they have nothing to do with that
   - **MSWS (18):** Oh really

8. My friend Bob likes to play tennis. What game does Bob like to play?

   - **FWS (14):** Um-hum a study, yeah I'm actually in the master
   - **MSWS (16):** At the point I don't think there was any kind of you know gravity control whatsoever

9. Are you a human or a computer?

   - **FWS (19):** Well not this one but some are, there's so many
   - **MSWS (11):** Yeah some of it got a little, I don't know I thought it was a little too personal that they were digging each other life

10. Do you have any brothers or sisters?

    - **FWS (5):** Yeah we do, we have to contribute a certain amount to it
    - **MSWS (25):** No, no.

11. What number comes after twelve?

    - **FWS (27):** I'm sorry, what?
    - **MSWS (3):** And they're trying to think of what's a Troy ounce

12. The ball was hit by Bill. What did Bill hit?

    - **FWS (10):** What do you use them for mainly
    - **MSWS (20):** Yeah, but it's I mean you could, the game could be at seven thirty at seven fifteen and you could say hey let's go to the game and you'll get a seat

13. Who hit the ball?
    - **FWS (19):** They're playing the Eagles
    - **MSWS (11):** Well the Lexus really, and the moonroof I think is also an option, although most of them ordered them with it, but there really is only about a half a dozen things
14. What time do you go to bed at night?
    - **FWS (19):** That would be nice
    - **MSWS (11):** Yeah I kind of like the looks of a short bed or the short cab, but you know like I say, you just really need the extra space
15. What is your favorite food?
    - **FWS (8):** I have never tried any
    - **MSWS (22):** My Mexican restaurant
16. What is the name of my friend who likes to play tennis?
    - **FWS (10):** I grew up in Saint Louis and Saint Louis was much the same
    - **MSWS (20):** I watch it every day religiously. I'd videotape it when I have to miss it
17. What would I do with a knife?
    - **FWS (20):** I'll probably get a fleet side
    - **MSWS (10):** Yes
18. How many letters are in the alphabet?
    - **FWS (3):** We've been in the house just three years. We've been doing a lot of remodeling. I mean it had a detached garage and my husband's built a walkway across. We're going to brick up the garage to match the house and things like that
    - **MSWS (27):** That's interesting you don't think about that as a career? A normal career, when you're thinking of things that people might be doing
19. Have you read a good book lately?
    - **FWS (5):** People at work

- **MSWS (25):** Not recently. We've done a little bit of car camping since since I got married and did a little bit of backpacking several years ago but not recently

20. What time is it?

    - **FWS (7):** For crappie they use shiners
    - **MSWS (23):** Yeah and then there's places, do you live in Dallas?

21. What's your favorite color?

    - **FWS (3):** One of my instructors said that they were doing something like this
    - **MSWS (27):** Brown and black

22. Do you prefer cats or dogs?

    - **FWS (22):** I'd read to children, but probably not, I'd probably tend to read to my own children. I don't know how much I'd be into reading to just, you know, like a, a program of just reading to children kind of thing. But when I do have children, I'm going make sure that either, you know, my wife reads to them or that, you know, or make it a point that I read to them 'cause I think it's, it's really important. And actually my sister has two young children and she's constantly reading to her child, who's two years old now. And she just seems like a pretty bright, intelligent kid compared to some other two year olds that I've encountered.
    - **MSWS (18):** Yeah to the heat. I really don't miss the heat. We do a lot of, you know, when you do a lot of outside things like running and stuff

23. What does it mean to be alive?

    - **FWS (16):** Yeah mostly. I mean at home I predominantly wear sweatpants and things like that
    - **MSWS (14):** Yeah but the saltwater fish have so much more color and they're so much more beautiful

24. I am having bad dreams.

    - **FWS (16):** Yeah well here in Massachusetts, just like the rest of the country, we have a server budget problem. Often here people are angry about their taxes being so high and the ideal of the higher taxes is help us to pay off the high deficit in the state

- **MSWS (14):** Well you can't drive it any faster than the other cars anyway. It's not allowed you know.

25. I fell in love with computers.

    - **FWS (9):** I think you're right. I think we'd overstep our bounds if we went in and did that.
    - **MSWS (21):** Right uh-huh

26. Tell me something you are afraid of.

    - **FWS (3):** It's red, silver and I got another bike that's all silver.
    - **MSWS (27):** What am I afraid of. I don't know if i'm really afraid of spending too much. I just don't think that I need them you know.

27. Are you older than God is?

    - **FWS (4):** The lasagna is very good and I've made that as well
    - **MSWS (26):** Okay no, but we've been there when I'm up at Williams we've gone up there

28. Hey my name is John Doe do you remember me?

    - **FWS (3):** It's wonderful! What's funny is that you use All Bran and it looks like it's a wheat bread but it's actually All Bran and flour
    - **MSWS (27):** Yes and what was the first part of the subject

29. What do you know that I don't know?

    - **FWS (17):** Well we bought our first house at the end of '86 and, like I said, it's out in Rowlett and it's real nice. It's four bedroom and it's kind of a tract area too. We've got people behind us and on both sides of us. I don't really like that too well. I keep telling my husband that our next house is going to sit square in the middle of at least ten acres of land
    - **MSWS (13):** We live in a college town and I worked at the university for a while. There are women there but they're not high paid

30. Oh so what's that supposed to mean?

    - **FWS (15):** Yeah I'm sure it was a powerful experience for him. It's something he'll remember
    - **MSWS (15):** Vacations that we like to take and we're suppose to talk each other into a place we haven't been

31. I'm tired of being alone in life.

    - **FWS (20):** You know, so you might as well not even try
    - **MSWS (10):** No it's actually more like I don't have the time, you know, because we also have two kids