NLP Analysis of Folksonomies

An examination of the Matukar language

Jonathan Gluck

1 Abstract

Folk taxonomies are powerful cultural tools for the categorization and utilization of the world in which a people live. The English language, for example, has a few folk taxa remaining; including 'fruits', 'vegetables', 'pets', 'farm animals', and 'evergreens'. Folk taxa are categories or logical groupings, usually referring to nature, which may have social and cultural relevance, but not necessarily possessing any scientific relatedness amongst their members. They are useful in day-to-day dealings with the environment, providing a catalogue grouped by salient features. Finding a language's folk taxonomy can often be difficult, with the lines drawn between categories not readily apparent. With this work I examine the theory behind folk taxonomic classification and attempt to devise methods for unearthing folk taxonomies with the help of Natural Language Processing.

The subject language of this inquiry is Matukar. Matukar is an Austroneasian language, spoken by only about 430 villagers on the North Eastern coast of Papua New Guinea. The language is spoken in a rural area and exhibits many onomatopoetic words reminiscent of the ambient sounds of their surroundings (Harrison, Anderson and Mathieu-Reeves 2010). It is a language threatened by the rising popularities of competing languages, such as English and the local creole Tok

Pisin. The folk taxonomy of Matukar has never before been examined, and is the focus of this work.

The job of unearthing a folk taxonomy involves sifting through large quantities of target lexical entries and searching for patterns in word form. Procedures, like these, which make use of large amounts of data are well suited to Natural Language Processing, or NLP for short. NLP is the subfield of Computer Science most concerned with language and its use. With the help of NLP it is possible to process quantities of data that might otherwise be prohibitive for hand analysis.

It is often the case that members in a folk taxon have similar names, or exhibit internal patterns. (Berlin, Breedlove and Raven, 216) One such example is the use of 'fish' to group marine life in 'jellyfish' and 'goldfish'. In order to find such examples, I use the NLP tool of string similarity. This involves comparing the similarities between any two words and selecting for those that pass a certain threshold. This tool should provide a list of similar words in a target language, revealing similar folk taxa.

While members in a given folk taxonomy may not directly map to English's professionally influenced taxonomy, many of the borders between folk taxa are influenced by their members' higher level categories (E.g. bird, insect.) (Hunn, 830-831) Imposing the English taxonomy onto a target language might provide an overlying structure within which to search for orthographic similarities. In order to do this, I implement automatic semantic tagging using WordNet , pairing English and Matukar entries by the English gloss for each Matukar word

With the assistance of Natural Language Processing the examination of folk taxonomies may be streamlined, providing linguists with a starting point with which to theorize folk taxa. I apply these tools to the Matukar Language, and examine the results.

2 Introduction

The range of human interaction, both in natural and social spheres, is vast. Even so, we humans are able to wrap our minds around the complex world in which we survive. The catalogue of discrete objects maintained in the human mind is of astounding length, so much so that the mere listing of a subset of this catalogue, for example names of familiar games, is rendered impossible. Access to this entire list at once is not possible. Yet, if 'Hop Scotch' or 'Mother May I' are referenced, the audience, so long as it has met with these games before, knows immediately not only that they are games, but also the environment in which they might be played and a myriad of other details. Accessing this knowledge is possible by the human process of categorization. Humans observe the dynamics of their surroundings and file away their daily experiences for later use.

One specific, useful type of categorization is the Folk Taxonomy, or Folksonomy for short. Folksonomies are cultural methods developed over time for the classification and compartmentalization, of the day-to-day experiences of human life. They are traditionally biological, although they are not confined to biology

only.[1] They allow an understanding of species and how they relate to one another. They are culturally relevant tools, and though they are not necessarily standard throughout a culture, they are a powerful tool to allow for the organization and control of the surrounding environment.

The goal of this project is to examine the theory behind folk taxonomies, and then analyze one language, Matukar, for clues pointing to possible folksonomies. In addition to the above goal, we desire to test the effectiveness of these NLP tools in automated semantic tagging. The search for folksonomies will be undertaken with the help of Natural Language Processing tools operating on the Matukar Online Talking Dictionary.

3 A Survey of Matukar

Matukar is an endangered language of Papua New Guinea, spoken in two villages in the Madang Province[2]. The language, at current count, has about 430 speakers, including both "experienced elders and children" (Harrison, Anderson and Mathieu-Reeves 2010). Matukar is an endangered language because of the continual rising popularity of English and of Tok Pisin (the local creole) and most common language of Papua New Guinea (The Central Intelligence Agency 2009).

While there is much that is not known about the language, there are some pertinent qualities which may impact its potential folksonomy. Matukar's villages are situated along the coast line; thus common animal categories and species might range from

[1] *Our modern taxonomies may be non-biological in nature, because our surroundings no longer call for biological categorization. One example of a non-biological folk category would be the "chick flick."*
[2] *Map of area attached in appendix*

aquatic to terrestrial to avian in form. An interesting feature of the language is that it contains many onomatopoetic words for living things (Harrison, Anderson and Mathieu-Reeves 2010). It is also important to note that the main agricultural products of the area are: palm, sweet potatoes, shellfish, poultry, and pork (The Central Intelligence Agency 2009). These products bear keeping in mind as we undertake analysis of the language. The more culturally relevant a word, the more likely it is to exhibit some taxonomic import.

The medium through which I explore the Matukar language is *The Matukar Online Talking Dictionary*. This is a dictionary of some 3,045 entries with associated audio recordings. There are no other published corpora of Matukar. It should be noted that this is not a large dictionary, and it was not created with the goal of folk biological elicitation in mind, so results are likely to be incomplete.

4 Three Theories of Folksonomy

The importance of human classification has engendered much debate. How does the human mind structure information? How does this information relate to the concrete biological hierarchy of modern scientific taxonomy? With what mindset should folksonomies be approached? In this section I will examine the arguments of three scholars on these issues and present their proposed folk taxonomic models.

4.1 Extendable Hierarchical Model

Brent Berlin is an American anthropologist most famous for his work on color terms. Berlin outlines a number of points on the subject of folksonomies. It is

his belief that the similarities between folk taxonomies and scientific taxonomies have been ignored, and that this should change. Berlin begins by stating, "In all languages it is possible to isolate groupings of organisms known as 'taxa'" (Berlin, Breedlove and Raven 1973, 214). These taxa are grouped into small ethno-biological categories, which are arranged into a hierarchy. These taxonomic categories are as follows: unique beginner, life form, generic, specific, and varietal. Taxa of the same category tend to occur at the same level, but this is not required. They are diagrammed below with examples for each category in Figure 1.

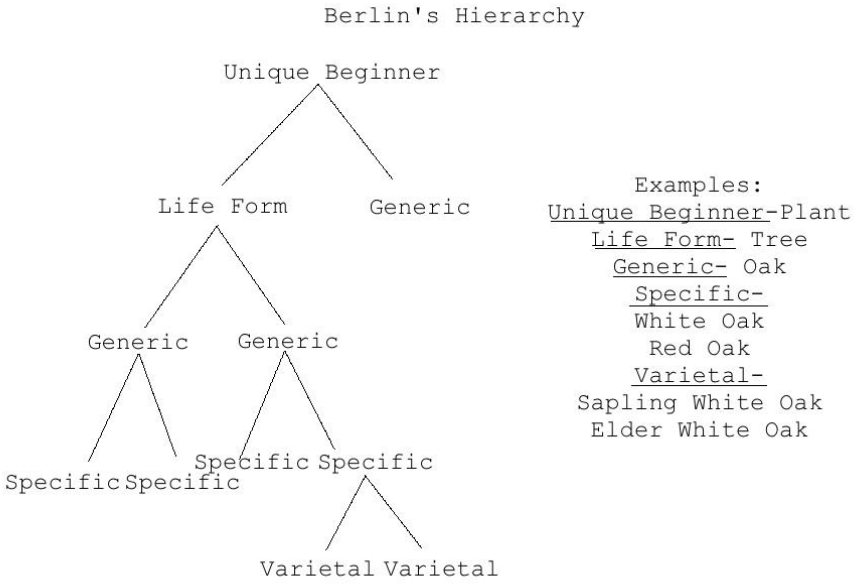


Figure 1: Berlin's Model of Folk Taxonomic levels

According to Berlin, the unique beginner category often goes unnamed in folk taxonomies. This unique beginner is something like "organism," "animal," or "plant." Directly underneath the unique beginner are the life forms. Life forms tend to be few but important. Most taxa fit into one of the life forms. Berlin states, of generics,

that they are more numerous than any other taxon. Most generics are immediately included as a child of some life form. Generics are the most important taxa for daily life. They are the taxa that are most quickly acquired by children. Sometimes generics are found without a parent life form class. In these cases, the generic is usually a borrowed word (Berlin, Breedlove and Raven, 220).

After Berlin lays out his taxonomic hierarchy, he undertakes a short explanation of the formation of these words. He shows that, in his system, all taxa, with the exceptions of specific and varietal, are denoted by "primary lexemes." Specific and varietal taxa are denoted by "secondary lexemes" (Berlin, Breedlove and Raven 1973, 216). Primary lexemes tend to be single words and can be either analyzable ('blueberry') or un-analyzable ('spruce'.) Secondary lexemes tend to be made up of two words, a descriptive word and a primary lexeme from another taxa, for example 'blue spruce.'

Berlin's arguments are compelling. The true utility of his hierarchy stems from its flexibility. He attempts, through his arguments, to find a model that is a compromise of several older models. In doing so he creates a truly extensible system.

4.2 Central Decentralized Model

Eugene Hunn is an American anthropologist who has a special focus on the cognitive aspects of ethno-biology. He believes that ethno-biology as a field has lost sight of the importance of examining the utility of folk taxonomies. He exhibits a strong belief that folk taxonomies are products of necessity and thus intrinsically utilitarian. In this vein, he gives a nod to Berlin who acknowledges that

folksonomies are often affected by “cultural significance” (Berlin, Breedlove and Raven 1973, 839). Hunn explains that one reason for the utilitarian basis of folksonomies is that there is an information processing limitation that is imposed by the sheer number of possible items to classify. Thus, humans must process those species that are the most useful first.

Hunn forgoes the hierarchical model for a centralized/decentralized model. He explains that the central categories are the easiest to recall. They are polythetic, determined by several optional characteristics. Non-central categories are both artificial and monothetic; members of these sets must subscribe to strict properties. This system is diagrammed, with examples, in figure 2.

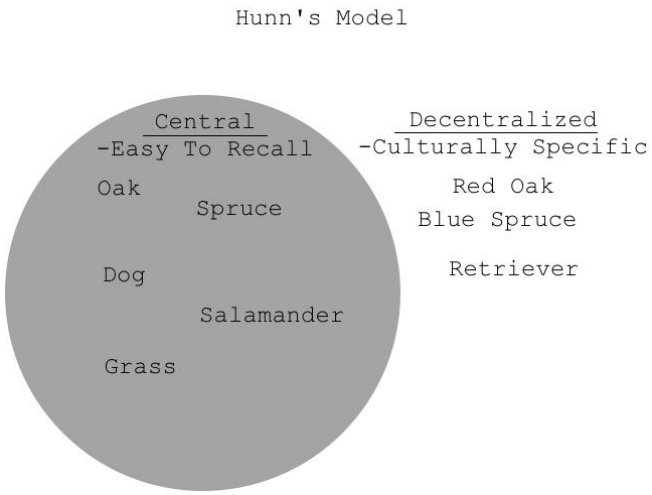


Figure 2: Hunn's Model of Folk Taxonomic Groupings

Hunn believes that Berlin might be attempting to jam these "central" categories into his generic taxa. This, to Hunn, seems “awkward” (Hunn 1982, 836), as the generic class, in Berlin’s hierarchy, is often found at several different

locations, superordinate and subordinate to the generic taxa level. Hunn also highlights an issue with Berlin's parallels between scientific and folk hierarchy, that the folk taxon, "bird," might be entirely different from the scientific taxon of the same name. The folk taxon, for example, might refer to organisms with "environmental or aerial habitats," (Hunn 1982, 838) while the scientific taxa are concerned with biological relatedness.

Hunn's central/decentralized model is an appealing alternative to Berlin's hierarchy. Hunn is concerned by the overwhelming focus on folk taxonomies as examples of "classification for its own sake" (Hunn 1982, 831). Hunn proposes that the utility of each word in a given taxonomy must be examined closely before attempts are made at compiling a model of that folk taxonomy.

4.3 Concrete Hierarchical Model

Scott Atran is a French American anthropologist. He is concerned with universal concepts in human thought and society. He currently studies biological classification in the mind. Atran believes that the system of classification present in folk taxonomies is "a cognitive mapping that places living-kind categories in a structure of absolute levels, which may… correspond to different levels of reality" (Atran 1995, 141). Based on this statement, Atran's theory is more akin to Berlin's hierarchical model than to Hunn's central/decentralized model. Additionally, it suggests that Atran believes folksonomies have a basis in reality. Atran states that the concept of folk taxonomies is hinged on the belief that variation not only exists in nature, but that it divides down salient lines (Atran 1995, 135). Humans develop taxonomic classes and imbue them with qualities learned from "naturalness" (Atran

1995, 137). 'Naturalness', in this case, refers to the quality of an object, which belongs to a category, being associated with the rules governed by that category. (e.g.: even a pygmy elephant is cognized as a huge animal, simply by being an elephant.) Atran points out that folk biological taxonomies are special in that they have this quality of naturalness. Taxonomies of artifice do not exemplify this naturalness. Atran provides the following example. A no-legged table, suspended from the ceiling is considered a perfectly good table; but a three-legged tiger with a prosthetic leg is considered deficient (Atran, 137).

Atran's model is divided into four taxa in a hierarchy. The taxa in descending order are: folk kingdom, folk life form, folk species, and folk subspecies. This model is diagramed below, with examples, in figure 3.

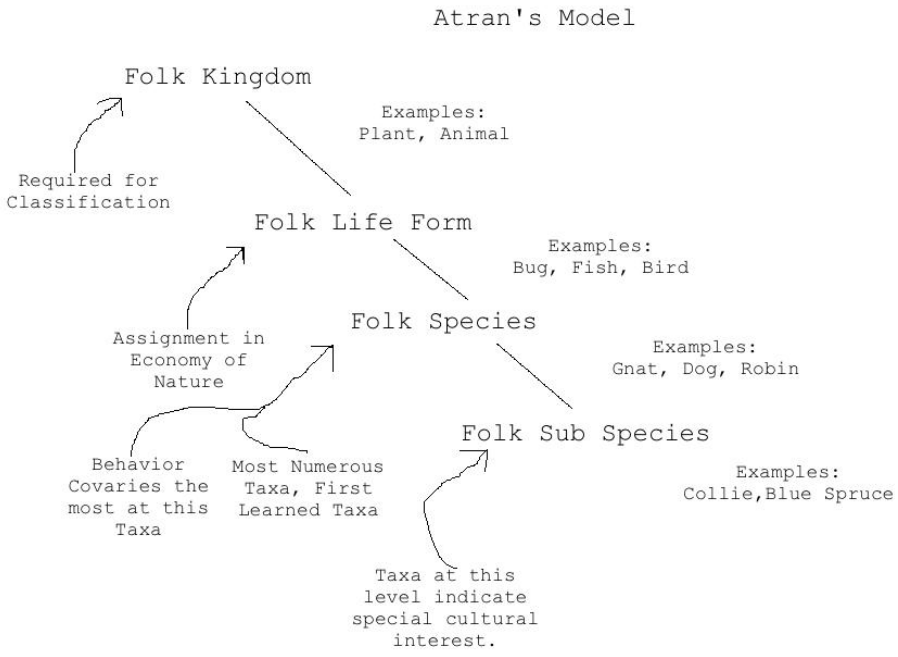


Figure 3: Atran's Model of Folk Taxonomic Levels

Atran makes some observations about particular taxa in this system. Of the folk kingdom, he explains that any observation must be classified into a folk kingdom first if it is to be classified at all. This is a sensible requirement of classification. In order to classify some actor at a low level, we must first understand where that actor fits into a higher level. Additionally, it provides some insight into why scientists are disturbed by the uncertain kingdom of viruses. Of folk life forms, Atran explains that this class is responsible for the assignment of a classification in the "economy of nature," (Atran 1995, 142) that is to say, how a particular plant or animal fits into its surroundings. He says of folk species, that they make up the most numerous level in the hierarchy. They are the point at which individual behavior differs the most. Folk species are the first taxa learned by children. They are also the most culturally relevant (Atran 1995, 143)[3]. This suggests that folk species are akin to the central terms in Hunn's model, and to the generic level in Berlin's model. Of folk sub-species, Atran explains, that this is the level of cultural interest. Taxa at this level, for example different varieties of corn, exist because they are of particular interest to a given culture.

While Atran's model is more similar to Berlin's than it is to Hunn's, he shares Hunn's belief that the field examining ethno-biological classification is too focused on scientific parallels. He states that natural kinds are determined by necessity (Atran 1995, 164).

[3] *Atran shows this with an explanation of how children in the western world recall folk species the most quickly only in cases of mammals. When a non-mammal was elicited, the children produced folk life form terms.*

An additional feature unique to Atran's model is the inclusion of intermediate taxa that often go unnamed. He provides the example of an intermediate taxon in English with 'mouse' and 'rat 'as children. This taxon accepts no other small rodent (Atran 1995, 140). Atran believes that, although unnamed, these taxa deserve inclusion in a complete ethno-biological model. This possibility of intermediate taxa is mentioned in Berlin but, because intermediate taxa often go unnamed, Berlin argues against their inclusion as an ethno-biological category (Berlin, Breedlove and Raven 1973, 216).

5 A Primer on Natural Language Processing

Discerning folk taxonomies from a corpus involves sorting through large amounts of data and searching for patterns or similarities in morphology. Procedures making use of large amounts of data are perfectly suited to Natural Language Processing (NLP for short). Natural Language Processing, also sometimes referred to as Computational Linguistics, is the subfield of Computer Science most concerned with language and its use. There are many tools available to NLP, but the two that I will examine here are: *String Edit Distance* and *WordNet*.

5.1 String Edit Distance: finding string similarity

In Computer Science, any arbitrary arrangement of characters is known as a "string." String Edit Distance is a measure of similarity between two strings. The smaller the string edit distance, the more similar the two strings. If the string edit distance between two strings is zero, then the two strings in question are identical.

One particular implementation of String Edit Distance is known as "Levenshtein String Distance." This algorithm steps through each paring of words and scores that pairing. This score is the minimum number of character transformations that must be made from one string to get to the other. The algorithm understands three operations at any given character, these are: deletion, insertion, and substitution. If any of these three operations is necessary, a point is added to the string edit distance between the two strings. Levenshtein String Distance keeps track of the edit distance of each substring of length *n* in word *a* to the corresponding substring of length *n* in word *b*. At each step, the algorithm adds the distance gained by appending the n+1 letter to both strings. When the last letter of the strings is appended, then the resulting distance is the string edit distance. An example of the computation of Levenshtein String Distance is shown below in figure 4, where the Matukar words for wave, *lalor*, and firefly, *altot* are compared. The distance between each substring of these two words is shown in their respective cells. For instance one might see that the transformation between the substrings 'ALT' and 'LAL' can be achieved in two edits, one deletion 'T' and one addition 'L'.

		A	*L*	*T*	*O*	*T*
	0	*1*	*2*	*3*	*4*	*5*
L	*1*	*1*	*1*	*2*	*3*	*4*
A	*2*	*1*	*2*	*2*	*3*	*4*
L	*3*	*2*	*1*	*2*	*3*	*4*
O	*4*	*3*	*2*	*2*	*2*	*3*
R	*5*	*4*	*3*	*3*	*3*	*3*

Figure 4: Levenshtein String Edit Distance Example

The importance of string similarity may be seen in Berlin's explanation of the morphology of taxa. Berlin states that taxa are made up either of primary or secondary lexemes. Primary lexemes are further subdivided into analyzable and un-analyzable groups (e.g. 'crabgrass' is analyzable while 'grass' is not) (Berlin, Breedlove and Raven 1973, 218). The reason both analyzable primary lexemes and the whole group of secondary lexemes may be analyzed is that they contain embedded orthographic clues. These morphological similarities provide hints at the underlying order of the folk taxonomy. For example, the secondary lexeme 'white rose' is a combination of the primary lexeme 'rose' with the color term 'white.' If we wanted to examine the various varieties of roses in English, we could look for every instance of the word 'rose' in a complete dictionary and the result would be a list containing all roses (as well as some noise, such as 'arose'.) This would give us a window into the English folk taxonomic specific children of the taxonomic generic 'rose.'

The above is only possible because we know that English forms binomials in which the second word is 'rose' for its rose taxon. The question is, how might we find these analyzable taxa without knowing what any of the language specific patterns are to start? At their base, patterns require some similarity. This is where string similarity becomes useful. If string edit distance is run on an entire dictionary, and the words most similar are reported, then words such as 'Colorado Spruce' and 'Blue Spruce' would be reported together due to their second words being identical. String similarity can unearth similar patterns over an entire corpus. Thus, string similarity is a useful tool in an automated taxonomic search.

5.2 WordNet: A Semantic Hierarchy of English

The second of the NLP tools used in this word is WordNet. WordNet is a powerful resource created by Princeton's Computer Science and Linguistics departments. It may be accessed online at http://wordnet.princeton.edu. It contains a relatively comprehensive hand annotated semantic hierarchy for English. WordNet is, in essence, an attempt to provide a solid reference to English's categorization scheme. English words in WordNet are grouped into sets of "cognitive synonyms," known as synsets (Miller 2011). Synsets are linked together by semantic relations. For example, the synset containing 'dog' is a child of the synset containing 'domestic animal' and also a child of the synset containing 'canine.' Children of the synset containing 'dog' include but are not limited to, 'puppy,' 'poodle,' and 'corgi.' A node with a selection of its hypernyms and hyponyms is illustrated in figure 5.

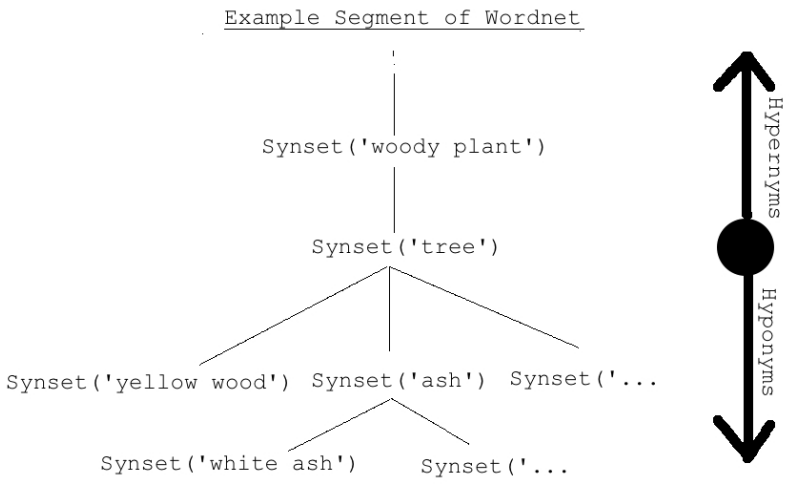


Figure 5: Example Segment of WordNet

The structure of WordNet closely resembles the hierarchies described by Atran and Berlin. This suggests that there might be some way to fit entries in a target language into the English taxonomic tree. Because of this, I came up with the idea of gloss assisted semantic tagging. By using the English gloss for each of a target language's words, I will be able to tag the words with English semantic fields. This will result in a catalogue of semantic categories. I can then walk through the semantic fields and examine the groups for morphological patterns.

One flaw with this approach is that it models the target language onto the English taxonomy, while the point of interest is the target's taxonomy. As is expressed in Penderson, Nimb and Braasch 2010, cultural backgrounds influence taxonomic structures radically. By mapping Matukar words onto English, the results may be unfaithful to Matukar's folk taxonomy. This would be a greater problem if the program were intended as a standalone analysis. The hope is that this initial mapping of the target language onto English's semantic tree might provide groupings of nouns that can later be analyzed for similar morphological qualities in the target language.

It should also be noted that it is only in the best-case scenario that this approach will remove all hand examination of the results. The main intent of this approach is to provide some semantic grouping to an untagged dictionary for the purpose of easing hand analysis afterwards. If the scope of this program is limited to all of the plants and animal words in the dictionary, this should accomplish a

categorization of all of the plants and animals in the target language into some more easily understandable format.

6 Implementation

In this section I will briefly describe the materials and methods I used to leverage the above tools on the Matukar talking dictionary. It should be noted that I was given an XML dump of the dictionary as my corpus. Both of these methods were implemented on Mac OS X using Python 2.7.1. The code for both of these implementations is available online.[4]

6.1 String Similarity

For string similarity, I initially implemented Levenshtein String Edit Distance, however; a problem quickly appeared. Levenshtein String Edit Distance does not reward similarity, while it does punish differences. For instance, the string edit distance between "white rose" and "yellow rose" is six, while the distance between "white rose" and "white house" is only two. Words that should have been grouped together were farther apart due to differing lengths, while words of similar lengths but differing meanings were being grouped together. This is counter to what one would want. The results returned by the simple Levenshtein String Edit Distance contained far more noise than they did signal. There are modifications that can be made for Levenshtein String Edit Distance so that, in addition to penalties for differences between words, it places a reward on similarities between words

[4] *The code will be made available at* http://www.sccs.swarthmore.edu/users/12/jgluck/Files/Linguistics/Matukar/src
It should be noted that the code will not work without NLTK having been installed.

however; this is more difficult to implement, and it would not necessarily remove the aforementioned problems. It was for this reason that I opted to look into other string similarity algorithms. I found a pre-existing algorithm, for this purpose in Pythons' difflib library. This function is called "difflib.get_close_matches()." The help file for difflib states that this function implements an advanced version of an algorithm called the "gestalt algorithm," by Ratcliffe and Obershelp[5], to produce similar strings that "look right to humans." This function works by finding the longest subsequence in common between two strings. It then runs the algorithm again on the sequences to the left and right of the previously matched sequences. This alternative sounded promising, and when it was integrated into the program it performed better than the basic string edit distance had, matching fewer sets of words erroneously. This program parsed the Matukar XML file into a dictionary of words, which was then analyzed. Groupings of similar words were generated for each noun in the dictionary. The runtime of this algorithm is relatively fast, taking on the order of a minute or two for the 3045 words examined.

6.2 Gloss Assisted Semantic Tagging

I implemented this method with the use of NLTK, the natural language toolkit, for Python. This method begins by finding all nouns in the dictionary which contain a word in their gloss that is part of a synset *A.* This synset *A* is, itself, a child of the synset containing "organism." The intention of this initial step was to collect all words to which Atran would refer as "living-kinds." The program then recourses

[5] A detailed explanation of this algorithm may be found at http://drdobbs.com/article/print?articleId=184407970&siteSectionName=

down the hierarchy of synsets, starting from Organism, creating lists of organisms that descend from the current synset. The program stops examining a branch when the current synset has no hyponyms. At each level, when the list of descendant organisms is compiled, if the list is non-empty, it is written to a file so that incremental results may be examined. This was implemented with an object oriented approach with a *NetWalker()* class handling the recursive process and a *NetOrganizer()* class handling the problem of listing descendants. The runtime of this program is rather long as there are many comparisons being made. For the Matukar dictionary it takes about an hour and a half to tag every word in the dictionary for every synset in WordNet that is a child of Organism.

7 Results of Natural Language Processing

I will discuss and analyze the outputs of these programs, and assess the usefulness of these methods in this section. Both methods had quirks; however, they both demonstrated potential for broader use for future automated analysis of target languages. Sample output for each of these methods may be found in the appendix. Additionally, full output of these programs is available online.[6]

7.1 String Similarity: Overview

The use of string similarity as a method of detecting similar orthographic patterns, which can subsequently be used to detect taxonomic groupings, returned some interesting results. There were some instances of success. To begin with, there

[6] *The output of these programs will be made available at* http://www.sccs.swarthmore.edu/users/12/jgluck/Files/Linguistics/Matukar/Results

appeared to be a fully formed taxonomic set of three birds. The set of similar words to *kukurek* the word for 'chicken' was as follows:

kukurek -chicken
kukurekparpar -hawk (chicken + sound of hawk?)
nubanen kukurek - goose (water + chicken)
kukurek katalun -chicken egg. (chicken + egg)

Figure 6: Strings similar to 'kukurek'

With the exception of *kukurek katalun*, 'chicken egg', each of these words represents a different type of bird. This is an exciting result providing evidence in favor of this method. By string similarity alone these words were separated from the entire dictionary. Unfortunately this is the only obvious example of primary/secondary lexeme interaction between classifications of species that I found. This does not mean that the method is unable to pick up on them; it just appears that they may not be present in the Matukar vocabulary, or (more probably) in the dictionary. This could be because the language does not include the primary/secondary lexeme feature, or it could be because the elicitation for the dictionary was incomplete.

Additional evidence for the utility of this method may be found in the plethora of terms associated with both coconuts and betel plants. In each case the basic words for 'coconut' or 'betel', *niu* and *mariu,* are appended with some other descriptor. (e.g.: *niu patawan*, meaning 'coconut milk'.) For each plant, these terms were grouped into that plant's similar strings. The relevant string edit distance groupings for these words are shown below in Figure 7.

niu ririn - fresh coconut meat remaining in coconut shell after scraping
niu dabin - coconut roots
niu patawan - coconut milk
niu raun - coconut leaf
==========
mariu bag - betel bunch
mariu - betel nut
mariu luwan - betel trunk
mariu digot - betel leaf attachment to tree
mariu sadaro - betel branch (broom)
mariu rau.un - betel leaf

Figure 7: String similarity grouping: niu- 'coconut' and mariu- 'betel'

These groupings do not represent individual species, however, and I have opted not to include them in my analysis of the folk taxonomy of Matukar.

There is also some evidence that, by using this method, the origins of analyzable primary lexemes, in a target language, may be more easily derived. For instance, one Matukar word for 'frog 'is *sidar.* The string similarity program returns that this word is similar to both the Matukar words for 'blood'*, dar,* and 'reef,' *sar.* It is possible that these words are conjoined in some way to create the primary lexeme *sidar.*

Overall there were some promising results for this method; however, due to the relative lack of biological terms in the dictionary, it is difficult to ascertain how effective it is. If there were more diversity in the species elicited for the dictionary, then it would be easier to gauge the effectiveness of this method.

7.2 String Similarity: Room for Improvement

While the method of string similarity I used unearthed some interesting patterns, there was still much room for improvement. Some of the below issues are inherent to this tool, while others have the potential to be mitigated with more advanced techniques. To begin, this method categorizes groups of short words together. These short words, even when very similar in form, often seem to have little to do with each other. Such a grouping may be seen below in Figure 8:

yad - part of a canoe
ya - hole
yau - fire *[paia]*
yan - yellow
dad - buy
bad - pot

Figure 8: Improper Grouping of Short Strings

Aside from a potential relationship between *yau* - 'fire' and *yan* - 'yellow' the other words in this grouping seem unrelated. This occurs because the shortest words have the least opportunity to become distinct. Two three letter strings can only be, at most, three string edits apart (replace each letter in string A with the corresponding letter from string B.) This leads to misleading conclusions such as the strings 'cat' and 'sum', with string edit distances of three, being more similar than the strings 'friend' and 'friendship', with a string edit distance of four. The former are unrelated, while the latter have the same root. Ideally we would prefer to have 'friend' and 'friendship' marked as more similar than 'cat' and 'sum'. Potential solutions to this problem involve providing more complex rewards to strings with longer similar substrings. For instance, if we decremented the string edit distance for common sub strings then the distance between 'friend' and 'friendship' would be

negative two. Such a distance would provide strong evidence for the relatedness of two strings.

A second weakness in string similarity may be seen in the case of binomials with shared descriptors. These descriptors are usually common words. In the output of my program there are many groupings that appear similar to the following in Figure 9:

te dabok - big bilum
nina dabok - big knife
maror dabok - big chief
tamat dabok - big man

Figure 9: Improper Grouping by Binomial Descriptor

These strings were marked as similar due to their shared descriptor *dabok* - 'big.' This would be akin to grouping 'red rose,' 'red fox,' and 'red panda' in English. While these patterns might be interesting, they are outside of our desired results. These errors are an unavoidable byproduct of this method; however, they are usually put into their own groupings and do not impede hand analysis.

7.3 Gloss Assisted Semantic Tagging: Overview

The use of WordNet to analyze the glosses of the Matukar dictionary returned interesting results, both promising and problematic. It successfully placed many of the Matukar dictionary entries in their corresponding locations in the English semantic web. This was most often true in the case of plants and animals. I have included the output for the synset 'ant' below in Figure 10.

Synset('ant.n.01')
ror: type of ant (black)
dəm: type of ant (very small, eats sugar)
bakbak: type of ant (black and brown, really big ant...)
kakad: type of ant (big, red ant that goes up tree)
maniŋkal: type of ant (brown, middle sized)
wes: type of ant (black, little ant who bites)

Figure 10: Example Output of NetWalker

The above shows all of the dictionary entries tagged by NetWalker as ants, all of which were tagged correctly. The trigger for categorization into this synset and the synset in question were the same; both were 'ant.' This is not always the case. For instance, in the synset 'insect' we may see, amongst others, the Matukar words *gab rairai*, 'type of fly,' and *muimui,* 'louse larva.' These terms were both categorized into the synset 'insects' because NetCrawler identified a string in their gloss, 'fly' and 'larva' respectively, that was an inherited hyponym of 'insect.' In most cases this method of tagging was sufficient; however, it was not without its flaws.

One problem that appeared in my experiments with this method was that WordNet includes the synset containing 'person' as a hyponym of 'organism'. While people are certainly organisms, the hyponyms of 'person' in WordNet primarily denote societal roles. This is problematic because the program attempted to tag all of the nouns in the Matukar dictionary with societal roles such as 'painter' or 'law man'. Even these unintended person related tags were applied with some degree of success. For instance the Matukar words for both 'virgin male' and 'virgin female' were tagged under the synset 'innocent'. This example of a correct societal tagging represents the exception. The noise to signal ratio would have been greatly reduced had 'person' not been included as a hyponym of organism.

Additionally, while browsing the output, I noticed that the Matukar word for 'tilapia' had not found its way into the results tagged with 'fish'. This turned out to

be because WordNet categorizes some specific names of animals under the synset 'taxa' and not under 'organism.' I ran the program again, this time with “taxa” as the root, and it returned only one categorization. This was 'tilapia.' I am uncertain whether WordNet has any more words like this, but I am certain that beyond 'tilapia' the analysis of the Matukar dictionary was unaffected.

The most common error, and the only unassailable flaw of this method is improper categorization due to English sense ambiguity. An example of this is the improper categorization into the English synset 'gum tree' of the Matukar word *gahu,* which may be translated as 'my gums.' This is a relatively common error and suggests that the output of this method is most useful when checked by hand afterwards.

In the case of Matukar, the output of this program provided all of the same insights as did the string similarity program and more. One piece of information this method detected that string similarity missed was the taxonomic class of *is*, the Matukar word for 'mosquito.' When I examined the synset for 'mosquito' I noticed that this program had tagged *is*, *is kaduman,* and *is wawak* all as members of this synset. This, in addition to the earlier group (*kukurek)* appears to be a second taxonomic group in Matukar. The reason that string similarity had missed this group was that the element that they all shared in common, *is,* is only two characters long. String similarity did not give appropriate weight to the similar qualities between these entries, as their shared substring was short, and thus passed over them.

I believe that gloss assisted semantic tagging provides an interesting automated means of semantic tagging for any target glossed dictionary and seems to

produce an understandable hierarchy of organisms in that target language. This could be an invaluable tool to any ethno-biologist. It has a few kinks; however, many of these would be fixable with time, and all of them are recognizable on sight.

8 Analysis of Output

Matukar seems to have a structured Folk Taxonomy which exhibits binomials, as Berlin predicted is common. However, from the data provided in the online talking dictionary I can only find two cases of direct taxa hierarchy. Aside from *kukurek*, *is*, and their respective descendants the vast majority of the language appears to be at the level of Berlin's generic taxa. In Hunn's model, this would suggest that all of the words, save the descendants of the two taxa above, would be central taxa. At first this seems extremely unlikely; however, the purpose of the Matukar dictionary was not to elicit an exhaustive catalogue of their biological terms. Its purpose was to create an initial repository for the language in general. This suggests that the vast majority of living-kind terms elicited were those that were most important to the Matukar people. These relevant terms would be the generic, or central, taxa. A piece of evidence in favor of this explanation is that the vast majority of organism terms found in the dictionary are focused on coconuts, and swine.[7] These are both staples of the Matukar way of life and thus would be likely to generate several generics.

[7] *These terms were not individual species terms, they were terms for parts of a coconut tree, or for counting swine.*

One glaring oddity is the absence of life form words, which are hypothesized in both Berlin's and Atran's models. Examples of life form words in English are 'bird,' 'fish,' 'insect,' 'flower.' The only example of a life form word that I was able to find in the dictionary was found in the definition of "bark." This was *ai suluŋan* which literally translated to 'tree skin.' This suggests that the Matukar for tree is "*ai";* however, this term was not given its own entry in the dictionary.[8]

The results seen here suggest that, in an effort to uncover the folksonomy of Matukar, additional research into the ethno-biology of the Matukar people would be fruitful. From this initial elicitation few ethno-biological levels are discernable. It would be difficult to continue examination of the Matukar Folk Taxonomy without the ability to elicit additional biological terms, and investigate whether the Matukar people have sets of life forms.

9 Related Works in NLP

In this section I will endeavor to provide an explanation of some other research with the tools that I have mentioned above. I do not believe that any research has been put into explicitly detecting folk taxonomies with Natural Language Processing; however, all of the following relate in some way.

9.1 Related Work: String Similarity

String edit distance as a method of identification of orthographic patterns has been examined previously in the context of morphological splitting. In Baroni,

[8] *I have since hand checked the XML dump of the dictionary, and found that the word "ai" is included with the gloss "wood." This gloss did not trigger inclusion in Organism's hyponyms, because wood is an object, not an organism.*

Matiasek and Trost (2002) they attempt to uncover morphological patterns in English and German using string edit distance. The importance of longer similar strings is encoded by normalizing string similarity by the length of the longest string. This method rewards longer strings for being more similar; however, it punishes similar short strings. As with my work, their algorithm does not report the patterns it uncovers, but allows a researcher to analyze the output for meaningful results. In addition to groupings by orthographic similarity, they include the notion of mutual information (semantic relatedness.) This measure is taken by co-occurrence in original text. For instance, if 'dog' and 'leash' often occur together, then one may draw the conclusion that a 'leash' is a 'dog' thing (or vice versa). co-occurrence is a feature of complete corpuses. This notion of mutual information is only available to full corpora and, as such, is not applicable to the single dictionary corpus of Matukar. Additional research into pattern discovery by orthographic similarity has been performed by Schone and Jurafsky (2000) and by Wicentowski and Yarowsky (2000), amongst others. In most previous work, the patterns being unearthed are meant to be morphological markers in a target language. In general these patterns are small conceptual units, such as prefixes and suffixes, and not the larger patterns that are required for taxonomic discovery.

9.2 Related Work: Bilingual WordNet

Less work has been performed on cross language utilization of WordNet. One group of researchers Fernandez-Montraveta, Vazquez and Fellbaum (2008) present their method for creating a new semantic web for Spanish as a tool for research. They accomplish this task by, essentially, hand translating the Princeton WordNet

into Spanish. They do not mention the problems inherent in mapping one language's semantic taxonomy onto another; however, they make use of hand annotation in the target language which corrects incorrect entries. The goal of their work is to build a Spanish WordNet, with a focus on one-to-one translations whenever possible. They do not attempt to discern native folk taxonomic structures that are distinct from English, which is a failing as different cultures can encode data differently.

Another attempt at building a non-English semantic taxonomy may be seen in Penderson, Nimb and Braasch (2010). Building a Danish WordNet, they focus on only monolingual approaches. Using a pre-existing dictionary they attempt to build a semantic network that mirrors the structure of Princeton WordNet, while maintaining a unique cultural identity. They call this monolingual method of net building a "merge" approach, while they name the more common approach (as in Spanish WordNet) an "expand" approach. The reason for using monolingual tools, in their estimation, is that too many cultural aspects of a taxonomy may be lost in an "expand" mapping. In building this Danish WordNet, they focus specifically on plants, animals, and food. They do this because they feel that these are areas highly affected by folk taxonomic differences. Their analysis shows that culture affects the structure of folk taxonomies in a significant way, and that this should be taken into account when building a semantic net for a target language.

My work with WordNet falls somewhere in-between the "merge" and the "expand" approaches. I am attempting to map a target language onto the English WordNet, so in this way my work is an "expand" approach. However, the cues that I use to facilitate this mapping, are found in a gloss of the target language, in a pre-

existing dictionary so there are elements of the "merge" approach as well. In either approach using WordNet as a model for semantic mapping has been proven successful.

10 Concluding Thoughts

10.1 Future Work

This project has many potential extensions. The string similarity method that I use is more sophisticated than simple Levenshtein String Edit Distance; however, results could be improved further with the utilization of an even more sophisticated string similarity algorithm.

Additional Natural Language Processing tools could be mobilized for this problem. Morphological splitting is a method that, given a training set and a large quantity of words in a target language, attempts to split words into their morphological parts. Morphological splitting is an application used in ways similar to how I use string similarity. Morphological splitting; however, is tuned to search for small strings at the extremes of words. This method could potentially have detected the taxon *is* - 'mosquito' on which my string similarity failed. Initial applications of morphological splitting to this data set have proved somewhat problematic, as there is no large enough training corpus for Matukar; however, it could be a productive course to follow.

Currently I use two tools, string edit distance for orthographic similarity, and WordNet for semantic tagging. The most immediate extension to this project would

be to feed the results of the semantic groupings from WordNet into a string similarity algorithm. This might provide the researcher with a list of all of the most related words, in a combination of semantic category and orthographic relatedness. This combination approach is currently in the implementation stage.

One interesting aspect to the tools that I used can be utilized with the assumption that no large body of literary works exists for the target language. If the researcher had available a large corpus of natural text/speech in the target language, then additional tools would become available. One example of such a tool is traditional semantic tagging, which attempts to learn the sense of a word by examining co-occurrence (relatedness) in a large body of data.

Bioinformatics tools often provide a suite of web interfaces, and useful visualization tools to researchers. I feel that the methods used in my work with Matukar would scale well to web applications similar to bioinformatics tools such as Basic Linear Alignment Search Tool (BLAST)[9] or ClustalW[10]. These tools could be useful to field researchers who would like some basic automated analysis of a target language.

10.2 Conclusion

The study of humanity's categorization of its surrounding is fascinating. That we naturally store our experiences using models for easy recollection is a testament to the efficiency of the human mind. Progress in studies of this area can be easily augmented with several Natural Language Processing techniques. The two techniques examined in this discussion were helpful in making sense of the Matukar

[9] BLAST may be accessed at http://blast.ncbi.nlm.nih.gov/Blast.cgi
[10] ClutsalW may be accessed at http://www.ebi.ac.uk/Tools/msa/clustalw2/

Folk Taxonomy and pointing the way for further study. In testing these methods I received promising initial results which invite further exploration. These methods could serve as useful tools for researchers, and with sufficient revision could remove the need for human assistance in the analysis of folk taxonomies.

10.3 Acknowledgements

This work, which is partially described in an earlier report (Gluck 2011) was undertaken for K. David Harrison. I would like to thank Professor Harrison for giving me access to the Matukar language resources, as well as providing an important primer, and many sources for folk taxonomic science. I would like to thank Professor Wicentowski for giving me help wrangling some of the less cooperative NLP tools. I would like to thank Rebecca Knowles for being a fantastic first reader, and providing me with many future avenues of research. I thank Professor Fernald for all of his thesis writing advice (helping me form a more polished work), as well as for his support. Lastly, I would like to thank my friends and family, against whom I have bounced ideas for the past year.

11 Bibliography

Atran, Scott. "Classifying Nature Across Cultures." Edited by Edward E Smith and Daniel N Osherson. *An Invitation to Cognitive Science* III (1995).

Baroni, Marco, Johannes Matiasek, and Harald Trost. "Unsupervised discovery of morphologically related words based on orthographic and semantic similarity." *Proceedings of the ACL - 02 workshop on Morphological and phonological learning*, 2002.

Berlin, Brent, Dennis E Breedlove, and Peter H Raven. "General Principles of Classification and Nomenclature in Folk Biology." *American Anthropologist* 75 (1973): 214-242.

Fellbaum, Christiane. *Wordnet:An Electronic Lexical Database.* Edited by Christiane Fellbaum. Bradford Books, 1998.

Fernandez-Montraveta, A., G. Vazquez, and C. Fellbaum. "The Spanish Version of WordNet 3.0." *Text resources and Lexical Knowledge*, 2008.

Gluck, Jonathan. *NLP Assisted Analysis of Folk Taxonomy.* Swarthmore: Self, 2011.

Harrison, K. David, Gregory D.S. Anderson, and Danielle Mathieu-Reeves. "About The Dictionary." *Matukar Online Talking Dictionary.* 1 1, 2010. http://matukar.swarthmore.edu/about.php (accessed 5 4, 2011).

Hunn, Eugene. "The Utilitarian Factor in Folk Biological Classification." *American Anthropologist* 84 (1982): 830-847.

Miller, George A. *Princeton University.* 2011. http://wordnet.princeton.edu/ (accessed 5 4, 2011).

Penderson, Bolette S., Sanni Nimb, and Anna Braasch. "Mergin Specialist Taxonomies and Folk Taxonomies in Wordnets - A case Study of Plants, Animals and Foods in the Danish Wordnet." *Proceedings of LREC*, 2010.

Ratcliff, J., and D. Metzener. "Pattern Matching: the Gestalt Approach." *Dr. Dobb's Journal*, 1988: 46-51.

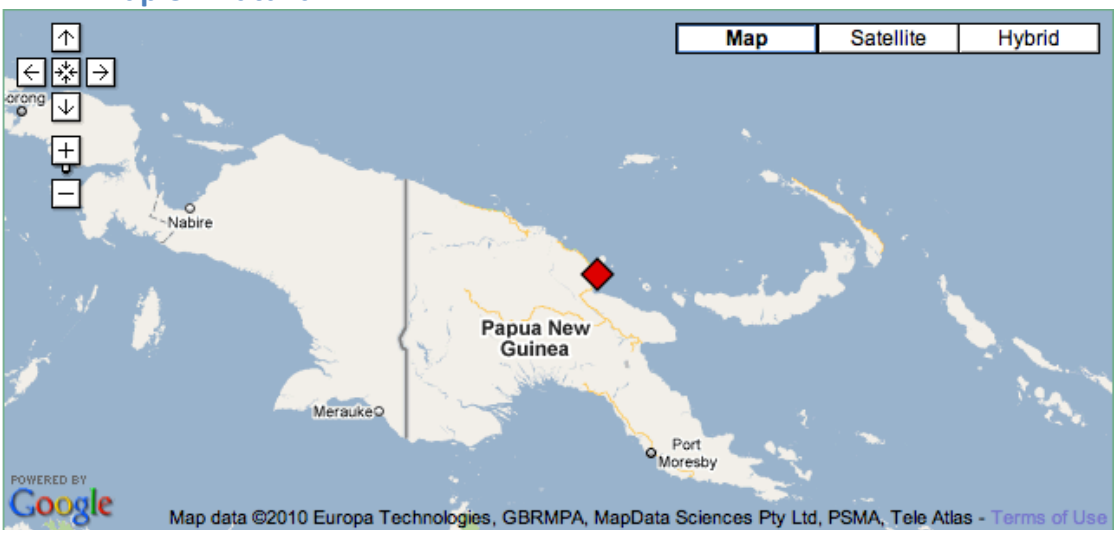
Schone, P., and D. Jurafsky. "Knowledge-free induction of morphology using latent semantic analysis." *Proceedings of the conference on Computational Natural Language Learning*, 2000.

The Central Intelligence Agency. "CIA-The World Fact Book." *Central Intelligence Agency.* 2009. https://www.cia.gov/library/publications/the-world-factbook/geos/pp.html (accessed 5 4, 2011).

Wicentowski, R., and D. Yarowsky. "Minimally Supervised Morphological Analysis by Multimodal Alignment." *Proceedings of ACL*, 2000.

12 Appendix

12.1 Map of Matukar



(Harrison, Anderson and Mathieu-Reeves 2010)

12.2 Example Output of String Similarity

tim: air
tim: wind
tidom: night
ti: no

nub yahai: waterfall
numau tahaik: five
nub narman: Water from yesterday
i yakai: he goes (but…)
ab yabi: S/he makes a house
nub wananan: hot water
nub koraman: puddle

kukurek: chicken
kukurekparpar: hawk
nubanen kukurek: goose
kukurek katalun: chicken egg

se paiin: paternal grandmother
sise paiin: old woman
sileŋ paiin: laughing woman
paiin: woman
kol paiin: female cousin
ham paiin: your wife
bagebage paiin: grandmother
ŋahau paiin: my wife
i wau paiin: my daughter-in-law
i wam paiin: your daughter-in-law

raurau uyan: Hello
garmaurau.un: my hair
abaŋ uyan: good day
garmauraun: my hair
mariu luwan: betel trunk
nal uyan: good day
fud uyan: good banana

12.3 Example Successful Output of NetWalker

```
============================================================
Synset('arthropod.n.01')
ror: type of ant (black)
kasaromrom: type of spider (lives in house)
dəm: type of ant (very small, eats sugar)
ləd: louse egg
is kaduman: mosquito larva
katabebe: spider
is wawak: mosquito (big)
bakbak: type of ant (black and brown, really big ant, goes up tree)
kabob: butterfly
altot: firefly
kalambu: mosquito net
kaiya: termites
alili: centipede
kakad: type of ant (big, red ant that goes up tree)
maniŋkal: type of ant (brown, middle sized)
is: mosquito
teratettet: type of insect
wes: type of ant (black, little ant who bites)
ut: louse
degadəg: cockroach
gab rairai: type of fly (big, blue)
muimui: louse larva
bukabuk: mosquito bite

============================================================
Synset('arachnid.n.01')
kasaromrom: type of spider (lives in house)
katabebe: spider

============================================================
Synset('spider.n.01')
kasaromrom: type of spider (lives in house)
katabebe: spider

============================================================
Synset('centipede.n.01')
alili: centipede
```

12.4 Example Improper Output of NetWalker

==========
Synset('producer.n.02')
mariu pidin: wood from betel nut tree
mariu digot: betel leaf attachment to tree
uləp: rope circle used for climbing trees (goes around feet)
ai suluŋan: bark (lit. tree skin)
nyat: hook for getting something from trees
tabe: brain, noodles, something inside of a rotten tree
pat: stone [(si)ton]

==========
Synset('film_maker.n.01')
pat: stone [(si)ton]

==========
Synset('architect.n.01')
kabakabman: eye white (possessed)
pat: stone [(si)ton]
kabakab: white

==========
Synset('maker.n.01')
laŋalaŋ tatuan: railing post
bag: post

==========
Synset('manufacturer.n.02')
laŋalaŋ tatuan: railing post
bag: post