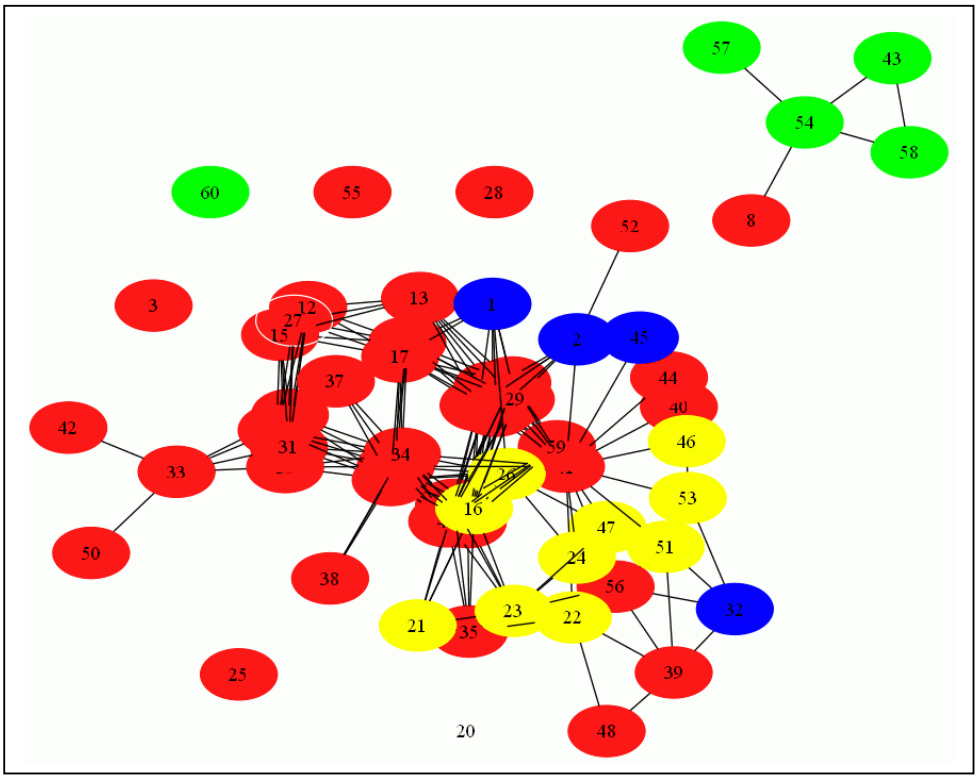
Feature Analysis of Containers

Bachelor's Thesis
Cognitive Artificial Intelligence
University of Utrecht
April 2011
7,5 ECTS



Author:
Maaike de Boer
3370852

Supervisor:
Dr. Joost Zwarts

Front page illustration: the illustration is one of the visualisations of the feature-analysis of containers.

Abstract

This thesis relies on a research by Malt et al. (1999). One of the experiments will be used to make a visualisation of the space of containers. This visualisation is a semantic map in the form of a graph. The method that will be used to make the semantic map is feature-analysis in combination with the hamming distance. The goal of this thesis is to investigate how connectivity in a feature-based graph is constraining for the categorisation of the objects.

Keywords: semantics, semantic map, graph, feature-analysis, hamming distance, connectivity

Contents

1. Introduction

The semantic systems of the world's languages vary greatly and because of this, two accounts of the relation between language and thought have been suggested (Khetarpal & Majid & Malt & Sloman & Regier, 2010). The first account is the Sapir-Whorf hypothesis, which says that cross-language differences exist because of corresponding differences in cognition. The speakers of different languages perceive the world differently and thus also think about the world differently. The Whorfian view predicts that speakers of languages with different semantic systems should perceive the world differently. Each group will perceive the world in line with their own language's semantic system.

The second account is the account of the universal conceptual space. This space is partitioned in different ways in different languages and therefore the semantic systems of the languages vary. This account predicts that speakers of different languages should perceive the world similarly, but that the partitioning of certain categories is different in a language invariant space.

In this thesis, I will focus on the second account.
I use a language invariant space of the semantic domain of containers to make feature-based graphs. These graphs will show that the partitioning or rather the clustering of the categories is different between languages, but also that there are similarities.

1.1 Containers

In this thesis, I will rely on research by Malt et al.(1999). One of the experiments in this article is a naming experiment. Malt asked speakers of the different languages English, Spanish and Chinese to name sixty different kind of common containers, like a jar, a bottle and a container. I will use the features of the stimuli and the most named name by the participants for a particular stimulus in their language for my experiment.

Malt also did a similarity-judgement experiment with the same stimuli as with the naming experiment. She used both experiments to make a map in a similarity space and investigated if there were similarity clusters in the distribution of names. I will use feature-analysis to make a map, more specifically a graph, of the stimuli and investigate if there is connectivity in the graph. This connectivity shows the resemblance between the different stimuli.

1.2 Relevance to the field of AI

"Artificial Intelligence is the science and engineering of making intelligent machines. … Intelligence is the computational part of the ability to achieve goals in the world." (J. McCarty, 2007). In order to make intelligent machines you have to create the ability to achieve goals in the world. You can create this ability if you make this implicit ability explicit. If you can make the implicit relation between language and thought explicit, you are a step closer to make an intelligent machine. In my thesis, I hope to help improve an account of this relation.

Cognitive Artificial Intelligence has four disciplines: psychology, philosophy, linguistics and informatics. This thesis combines the disciplines of CAI. The first discipline of Cognitive Artificial Intelligence is psychology. Language and thought are both represented in the mind, so this has a lot to do with cognitive science, a field of psychology. The Sapir-Whorf hypothesis is also a psychological hypothesis. Philosophy, the second domain, is not directly relevant for my thesis.
Linguistics, on the other hand, is the most important discipline in my thesis. This research is done, because of the question why there is difference between the semantic systems of the world's languages. Semantics is a part of linguistics. My thesis relies on a research of naming and pile-sorting objects, which is important for linguistics. The last domain is informatics. I will make a computer implementation to calculate the length of the edges in the visualisation. I will use this to make the graph in a computer program. Thus, for this thesis I used a lot of parts of the informatics domain.

1.3 Representation of the question

I want to investigate if there is a visualisation that can make a clear partition of the semantic domain of containers. Therefore, I want to investigate if there are different partitions of the categories in the different languages. This will be done by making a graph of the space and analyze the way the objects are placed in the graph. I will research the connectivity in the graph. Connectivity means that a graph is connected and if this is the case then all the nodes of the connected graph have at least one feature in common with another node.
I will also investigate the degree of connectivity. A high degree of connectivity means that in a subgraph there are a lot of edges between the nodes within the subgraph and less edges going to nodes outside the subgraphs. This means that some nodes have more in common with each other than with the other nodes.
I will use these notions to answer the following research question:
In which way is the connectivity of the investigated containers (Malt et al.,1999) constraining for the categorising of the objects in a feature-based graph and in which way do they differ and correspond to the different languages?

To achieve an answer to this question, I will use the following subquestions:

1. Which of the categories have the highest degree of connectivity and how can this be explained?
2. How is this result related to the results of the research by Malt et al.(1999)?

1.4 Structure

This thesis is structured as follows.
The following chapter contains information about the research by Malt. The study, results and conclusion of the research are explained. The result and conclusions are used as an inspiration for this thesis, but the way of analyzing the experiment is different.
Chapter three is about semantic maps. Both in Malts work and in this thesis, semantic maps will be used. The two approaches to make a semantic map are discussed and the space-driven approach, which is used to make the visualisation, will be further explained. The problems of this approach are also mentioned.
In chapter four, you can read about the hamming distance. Before you can make a semantic map in the form of a graph, you have to know the relation between the nodes. To express this relation the hamming distance is used. The hamming distance is used to calculate the distance between the nodes in the graph and express this by the length of the edge between the nodes.
In chapter five the notions of connectivity and a degree of connectivity are further explained.
In the first five chapters all the ingredients for my research are collected and explained. So in chapter six I will explain my research. First the computer implementation language Prolog and the computer program to make the graph, Graphviz, are considered. After that, the results are given. At the end of the thesis the answers to the questions, implications and suggesting for further research can be given. This will be in chapter seven.

2. Research by Malt

In *Knowing versus Naming: Similarity and the Linguistic Categorisation of Artifacts* (Malt et al.,1999) an experiment is described which argues that it is important to distinguish between categorisation as object recognition and categorisation as naming. Malt wants to rely on the study by Kronenfield et al.(1985) in which there are distinctive differences found in grouping of objects in linguistic categories by American, Japanese and Israeli participants, but there are relatively small differences found in perceived similarity among the objects.

In the next paragraph, there is information about the experiment Malt did. Paragraph 2.2 is about the result of this experiment and the analysis Malt did to get this results. In paragraph 2.3 the conclusions of this research are given.

2.1 Study

The participants of the experiments of Malt et al.(1999) were 76 native speakers of English, 50 Chinese speakers and 53 Spanish speakers. All the participants got a set of sixty common containers as stimuli. These stimuli were colour photographs of containers. You can see some of the stimuli in black-and-white on figure 1. The photos were taken against a neutral background with a constant camera distance to preserve relative size.

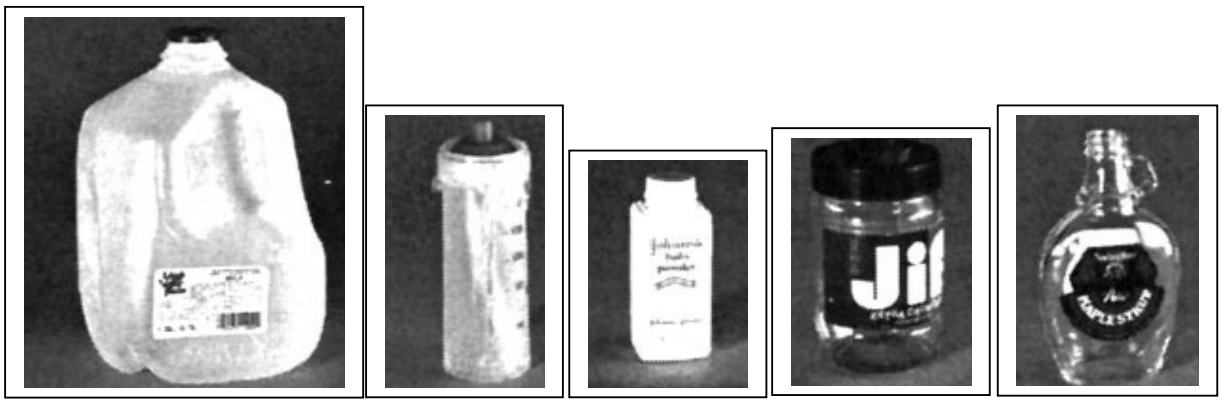


Figure 1: a few of the stimuli used in the experiment.

For the Chinese and Spanish speakers the contents of the objects were marked at the bottom of the picture.
The participants had to do two kind of experiments. The first experiment was to sort the objects into piles. The sorting was based on either physical, functional or overall similarity. The participants did two of the three sorting tasks. The second experiment was to give a name to each object.

2.2 Result

Malt et al. compared the correlation between the groups for the three kinds of sorting, through the results of the first experiment. All kind of sorting had a higher correlation with each other than the naming similarities. The functional sorting had the lowest agreement of all three kind of sorting. The agreement between groups was lower than within groups. Malt et al. also evaluated the results with the cultural consensus model of Romney, Weller and Batchelder (1986) and this evaluation indicated no significant group differences for overall sorting and smaller differences of functional and physical sorting than for naming.

Malt et al. did three kind of analysis on the second experiment.
In the first analysis, only the dominant name, that is the name that most participants said to be the name of the object, is used. They found out that the native speakers of English placed the objects into three categories, namely *bottle, jar* and *container*. All the three categories were approximately equal in size.
For the Chinese speakers, there arose one large category where almost all objects fell in. The category was *ping2* and it encompassed all the English jars, most of the English bottles and some of the English containers.
For the Spanish speakers there were 15 name categories. The most frequent category was *frasco*.

The second analysis uses a Pearson correlation to compare the linguistic categories of the three languages by comparing the similarity of each object's name distribution to every other object's name distribution. The name distribution does not only take the dominant name of each object into account, but all the names that were mentioned in the experiment. All the correlations resulted positive, but less than 1. This indicates that the three languages show differences and agreements. Another evaluation, with the cultural consensus model of Romney, Weller and Batchelder (1986), indicates that the groups from the different languages named the objects differently.

The third analysis is to make a map in a similarity space to investigate if there are similarity clusters in the distribution of names. She did that with a multidimensional scaling solution (Shepard,1974) of all three sorting experiment data using the KYST algorithm. The dominant linguistic category is used to label the objects in each solution. They mapped the objects to a similarity space and concluded that many members of each linguistic category cluster together, but some occur closer to members of other linguistic categories. Across all three sort types, Malt suggests that the linguistic categories are complex and do not map directly onto the similarity clusters.

The overall result is that differences in naming among the languages are only partially related to differences in perceived similarity and they show substantial independence.

2.3 Conclusions

The results supports the idea that it is important to distinguish between recognizing objects and naming, because people who speak different languages may have substantially different patterns of naming, while the similarities of objects among the groups are not different. The pattern of naming can therefore not arise only from the similarities people see among the objects. If that was the case, there should be similarity clusters in the mapping analysis. The pattern of naming can also be influenced by convention, pre-emption or chaining.

3. Semantic maps

In chapter two, there was a mapping analysis method to investigate if there are similarity clusters in the distribution of names. This chapter explains what a semantic map is, which approaches there are and what the problems of the approach I will use are.

A semantic map is a spatial representation of the different connections between linguistic meanings (Zwarts, 2010). It is important that there is a structure of meanings that is related to a set of forms each expressing one or more of those meanings.
A semantic map for a particular domain consists of two parts, namely a lexical matrix and a conceptual space (Croft, 2001; Haspelmath, 2003). A simple form of a lexical matrix is a table with words. For each word in a set of words it is showed which meaning from the conceptual space it can express. A conceptual space is a geometrically ordered set of meaning, typically a graph.

There are two mapping approaches. These are discussed in the next paragraph. In paragraph 3.2, there is a focus on one of the approaches. Paragraph 3.3 contains the problems with this approach.

3.1 Two Approaches

There are two mapping approaches (Zwarts, 2010).
The first approach is the matrix-driven approach. In this approach you build a conceptual space on the basis of a cross-linguistic lexical matrix in a data-driven fashion. In this way you build a map on basis of the words.
The second approach is the space-driven approach. The conceptual space is an existing conceptual space, which you confront with the cross-linguistic data in a meaning-driven fashion. For this approach you can use two ways. The first way is to make a map with similarity judgements, like Malt did. The second way is to make a map with feature analysis. This is the way I will use. The space-driven approach will be further explained, because this approach is used by Malt and by me.

3.2 The space-driven approach

In the space-driven approach, you take an existing conceptual space and investigate how the words are mapped onto it. The classic example of this approach is the study of colour terms (Berlin and Kay, 1969; Regier et al., 2007). An existing conceptual space, in this case an existing colour space, is used and the colour naming system of different languages is mapped against the existing space. They found out that the colour categories across languages are organized around a set of universal focal colours and that these colour terms near-optimally partition an underlying similarity space. So with the space-driven approach you can study cross-linguistic data with an existing space. Because of that, you can investigate the universal and language dependant factors.

3.3 Problems

But as with every approach, there are a few problems.
The first problem with this approach is that it is not always easy to find an a priori conceptual geometry. If you want to investigate an existing space, you have to find an unique structure of meanings for a domain. If you are not able to find this structure, you have to use the semantic map approach, because the cross-linguistic data can tell you the underlying conceptual structure.

Another problem with the space-driven approach is that the a priori conceptual geometry might bias the research and hinder us from finding patterns and dimensions in other languages. If you choose a particular geometry, you are stuck to this geometry and therefore you bias your research. But this is not always a weakness. It can bias your research in a good way and because of the ability to make a visualisation, you can analyze the data in a better way. Every approach has its advantages and disadvantages and I will use this approach because the advantages of this approach are more important than the disadvantages for my research. There seems to be no problem in finding an a priori conceptual geometry and the language independent space can help to improve the data analysis. But an important factor for making the visualisation in the space is to find a way to express the relation between the objects.

4. Hamming distance

The hamming distance will be used as a way to express the relation between the objects. The objects in the graph are nodes and the length of the edges between the nodes is the hamming distance between the features.

The hamming distance is named after Richard Hamming, who introduced it in the context of error-detecting and error-correcting codes (Hamming, 1950). It is defined as "the difference between two messages, each consisting of a finite string of character S with the same length, expressed by the number of characters that need to be changed to obtain one from the other" [1]. So it measures the minimum number of substitutions required to change one from the other. For example, the strings 'universal' and 'container' have a hamming distance of 9.
In my research, the strings are in the shape of [s,m,c,w,s]. The letters stand for the different features presented in the research by Malt. The features are size, material, shape, mouth and top. So, the first letter is the size, in this case *small*. If the letter of one of this features was not known, a 0 was added at that place of this feature in the string. The strings will be compared through the hamming distance and the length of the edges will be calculated.

There are also other ways to connect the objects in the graph, like with the Levenshtein distance. The Levenshtein distance, or edit distance, not only calculates the substitutions, but also the deletions and insertions. As seen above, the strings used in my research have a letter for each feature. It is not good to check if a letter of one feature is the same as a letter of another feature, because they refer to another feature of the object. This is why you don't want to use deletion or insertion as a way of calculating the distance. This is the reason why the hamming distance will be used and not the Levenshtein distance.

[1] Definition from : www.websters-online-dictionary.org > hamming distance > physics

5. Graphs and connectivity

After deciding what type of mapping and what kind of relation expressing you will be using, you have to make clear what kind of graphs you want to draw and what the terms connectivity and the degree of connectivity mean in order to answer the research question.

The graphs drawn in this research are undirected graphs with nodes and edges. The nodes represent the object with the name of the item and the edges are a representation of the connection between the nodes. If a graph is undirected, it means that the edges do not have a direction. This makes sense, because the length of the edges are represented by the hamming distance. If one object has a hamming distance of one with an other node, this other node also has a hamming distance of one with the first object. The length of the edge is the same and therefore the edge does not go from the first object to the second, but they have no direction.

In a graph there can be connectivity. In this case connectivity will be used in a meaning of a connected graph. A graph is connected if it is possible to establish a path from any node to any other node in the graph. So there is one graph and no node is excluded from this graph. To illustrate this principle, a connected graph is given in figure 2 and an unconnected graph is given in figure 3. Figure 3 has two connected subgraphs, but is not a connected graph.

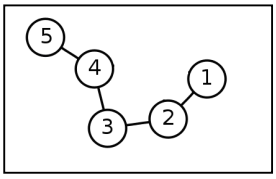


Figure 2: unconnected graph

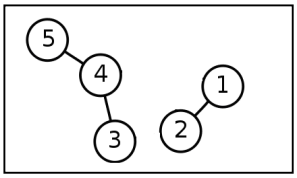


Figure 3: connected graph

There are degrees of connectivity. Some graphs have a higher degree of connectivity than other graphs. A high degree of connectivity means that there is a subgraph in a connected graph in which there are more edges between the nodes within the subgraph than there are edges going to nodes outside the subgraph. Figure 4 and 5 will help to understand the degree of connectivity. In figure 4 there is a higher degree of connectivity than in figure 5. In figure 4 you can see that the subgraph of numbers 2, 3, 4 and 5 have more edges inside the subgraph than going outside. In figure 5, the subgraph of numbers 2, 3, 4 and 5 does not have more edges inside the subgraph than going outside.

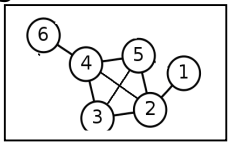


Figure 3: high degree of connectivity

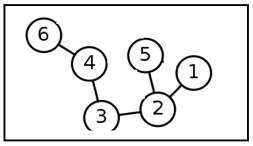


Figure 4: lower degree

6. My research

The ingredients for my research are the experiment of Malt, the space-driven approach, the hamming distance and the notion of connectivity. My goal is to make a visualisation of the semantic domain of containers, as investigated by Malt, that makes a clear partition of this domain. Feature-analysis in the space-driven approach is used and the hamming distance is used as a way to connect the objects in a graph. Prolog and Graphviz are used to implement and visualize the graph.

6.1 Prolog

Prolog [2] is used to implement the hamming distance.
Prolog, designed by Alan Colmerauer, is a declarative programming language. This means that the program logic is expressed in terms of relations, represented as facts and rules. A computation is initiated by running a query over this relations.

In this implementation, the strings that are compared are strings of features like [s,m,c,w,s]. Every place at the string stands for a different feature. If the letter of the feature was not known, a 0 stands at the place of the letter. The strings are part of an item. For example, item(1,[s,m,c,w,s]). If the you run the code in Prolog, for every item every other item is compared to this item and the hamming distance is calculated. If a 0 is read, it always matches with the letter of the feature of the other item. The hamming distance and the two items are transported to a file, in the form of 1 -- 2 [len = 1]. Len = 1 means that the hamming distance is 1 between item 1 and 2. If all the items are compared with each other, the file closes and the file will be used as the input for Graphviz.

6.2 Graphviz

Graphviz [3] stands for Graph Visualisation Software and it is initiated by AT&T Labs Research. Graphviz has tools to draw graphs in dot languages. This languages describe three main kind of objects: graphs, nodes and edges. The graph can be directed or undirected. In my implementation, I used the undirected graph, because the hamming distance works between two objects and not in one direction. The nodes are the items and the edges the line between the items. The length of these edges is the hamming distance.

The format that I will use is neato. Neato draws undirected graphs and uses a method equivalent to statistical multi-dimensional scaling for drawing the graph. It also uses shortest path distances to calculate the position of the node. The reason for using neato as format is, because neato gave me the best way of visualize the data.

[2] http://www.swi-prolog.org/
[3] http://www.graphviz.org/

6.3 Results

The graphs drawn with the Prolog implementation for calculating the hamming distance and the neato format in Graphviz are in the appendix. The appendix contains the graphs drawn with neato and only the hamming distance of one between two items. I choose this hamming distance, because it gave the most clarifying graph. If you choose the hamming distances of one and two or more, there are too much edges. With too much edges, it is hard to say which edges belong between which nodes and therefore you cannot analyze the graph in a good way. If the hamming distance between two nodes is one, all features but one are the same.

As you can see in the appendix, none of the graphs is connected, but in each graph is a connected subgraph. In this subgraph, there is a similarity between the three languages. You can see that the *jar*, *frasco* and *ping2* have all the same kind of clustering and this clustering has the same shape. Especially the Chinese and Spanish graphs stand out. In the Chinese graph the green objects have their own subgraph. The yellow objects in the large subgraph are close to each other. All the yellow objects can reach another yellow object through one other node. This means that these objects have a lot of features in common. The Spanish graph is less clarifying, but each yellow object is connected with all the other yellow objects. The most of the red objects are in the clustering of the subgraph. In the English graph, there is also a large subgraph, but the objects in that subgraph have less clustering of colour. This is very striking, because the experiment was considered by researchers that speak English. It is remarkable that the graphs of the Chinese and Spanish are in the two kind of graphs the most outstanding.

7. Conclusion

I wanted to investigate if there is a visualisation that can make a clear partition of the semantic domain of containers.
The question to be answered was:
In which way is the connectivity of the investigated containers (Malt et al.,1999) constraining for the categorising of the objects in a feature-based graph and in which way do they differ and correspond to the different languages?

To achieve an answer to this question, the subquestions were:

- Which of the categories have the highest degree of connectivity and how can this be explained?
- How is this result related to the results of the research by Malt et al.(1999)?

The answers of these questions are given in the next section. In section 7.2 the methodological implications are given, because the method I use is relatively new. I will give a few marginal comments of this method.
In the last section suggestions for further research are given.

7.1 Conclusion

The subquestions are answered first, because they build the answer to the researchquestion. The first subquestion is: Which categories have the highest degree of connectivity and how can this be explained?
The answer is language dependent, so I will discuss this question for each language.

In English, the category *jar* has the highest degree of connectivity. In the appendix, you can see that the green objects are connected and have a lot of edges inside the subgraph, which means that they have a smaller hamming distance with each other than with the objects of an other category. But there is one *jar* that is outside the subgraph, namely number 44. The percentage of people who labelled number 44 as *jar*, according to the experiment by Malt, is 78%. This is most of the people. Like Malt investigated, it can be influenced by convention, pre-emption or chaining. The other categories have a lower degree of connectivity, but in all cases there is connectivity in the categories. It seems that there is a lot of correspondence inside the categories.

In the Spanish graph you can see that the category *frasco* has the highest degree of connectivity. The categories *envase* and *bidon* have a low degree of connectivity. None objects of one of these categories has a hamming distance of one with another object of the same category. A reason for this can be that there are a lot of categories in Spanish and the difference between the categories is not strict. As the experiments by Malt showed, the dominant name of the objects is not convincing. In the Spanish categories the percentage of the people who named the same name to the object is often below fifty percent.

In the Chinese graph, the category *ping2* has the highest degree of connectivity. *Tong3* also has a high degree of connectivity, because the green objects in the graph have a separate subgraph in which these objects are connected. The category *guan4* also has a high degree of connectivity, but a lower degree of connectivity than *ping2* and *tong3*. There are less edges between two yellow objects than there are edges between red or green objects.

As you can see in all the languages, there is a category that has the highest degree of connectivity. This means that the nodes of the same colour are closer to each other than to nodes of other colours and there are more edges inside the subgraph of the category with the same colour than that are edges going outside. This means that the nodes in this category have a lower hamming distance to each other than they have to the nodes of different colours. Therefore these nodes have more features in common. This means that the objects of a category with a high degree of connectivity are named on base of the features. Because there are also categories with a low degree of connectivity, the features are not the only trigger to name an object.

The second subquestion is: How is this result related to the results of the research by Malt et al.(1999)?
Malt did the experiments to argue that it is important to distinguish between categorisation as object recognition and categorisation as naming. The results of the research by Malt was that differences in naming among the languages are only partially related to differences in perceived similarity and they show substantial independence. The data of the experiment support the idea that it is important to distinguish between recognizing objects and naming. There are substantially different patterns of naming among languages, while the similarities of objects among the groups are not different. The pattern of naming can therefore not arise only from the similarities people see among the objects. It can also be influenced by convention, pre-emption or chaining.

I did my research to investigate in which way the connectivity is constraining for the categorising of the objects. The result of my research is that there is no connectivity between the categories and that in each language there is a category with the highest degree of connectivity. This means that the objects of this category have more features in common and the participant have the same dominant name to these objects.
Also, there is a subgraph in the graphs for each language. The categories *jar*, *frasco* and *ping2* have all the same kind of clustering and this clustering has the same shape. This can be universal. Especially the Chinese and Spanish graphs stand out, which is striking.
So, there seems to be a similarity between the pattern of naming and the features the objects have. There is also a group of objects in every language that result in the same shape in a subgraph.

The research question is: In which way is the connectivity of the investigated containers (Malt et al.,1999) constraining for the categorising of the objects in a feature-based graph and in which way do they differ and correspond to the different languages?

As you can see in the appendix, connectivity is constraining for the categorisation of the objects. Some of the categories have a high degree of connectivity, which means that the objects have more features in common and therefore the naming can be dependent of the features. So, there seems to be a similarity between the pattern of naming and the features the objects have. Also, in each language there is a category in the graph that has the same kind of clustering and this clustering has the same shape as a category of another language. These categories are *jar*, *frasco* and *ping2*. On the order hand, categorisation is not only constrained by connectivity. In none of the graphs there is connectivity.

The way in which connectivity is constraining the categorisation differ in each language. The Chinese categories seems to be the most constrained by the connectivity. In three of the four categories there is a high degree of connectivity. With a hamming distance of one the objects that are named the same have a connection.

The Spanish categories seem to be less constrained. The categories *envase* and *bidon* have a low degree of connectivity. None objects of one of these categories have a hamming distance of one with another object of the same category. A reason for this can be that there are a lot of categories in Spanish and the difference between the categories is not strict. It can also be that the Spanish language does choose the name on basis of other features than the investigated features.

Connectivity seems to be the lowest constraining in the English categories. There is connectivity in the category *jar*, but the other categories have a low degree of connectivity. There seems to be little constraining of categories in the feature-based graph.

7.2 Evaluation

The way the feature-analysis is combined with the hamming distance and Graphviz is very new. That is the reason why I will evaluate the method.

The method I used, is a good method to help analyze the data. Because the hamming distance equals the length of de edge in the graph, it is easy to see if two nodes are connected and how related the nodes are. In this way, you can see if the nodes have a lot of common features and therefore you can make your conclusions. With Graphviz you also get a clarifying graph. But nevertheless Graphviz has a good method of drawing the nodes, sometimes nodes that are not related are next to each other. These nodes do not have an edge, but if you do not look at the edges, it

seems that the nodes have a lot in common. So it is important to look at the edges and the length of the edges, before you make your conclusion.

As you can see in the graphs in the appendix, the nodes are not drawn on the same place in every graph. So you also have to look if there is a good language invariant space drawn. If the space is partitioned in different ways in different languages, it has to do with the universal conceptual space and not because Graphviz draws the space different.

A big disadvantage of this method is that you cannot see which feature is different. You know the hamming distance, but you do not know which feature or features created the hamming distance. You cannot investigate which feature is constraining the categories, but only if there are features constraining categories. So, if you use this method, you have to know which question you want to be answered.

Also, this method can only be used if there are clear a priori features. In the research of Malt it seems that all the features were good, but you have to be careful. The were only sixty common containers, which it not that much, so not every combination between features is made. Therefore it is necessary to have a good set of a priori features and a good research before you start with this method.

As seen in this thesis, this new method can help to create new insights in a research and come with amazing results. If you can handle the points above, it is a good way of analyzing experiments and create nice results.

7.3 Suggestions for further research

Because of this new method, a lot of suggestions for further research came up. This method is not on its peak now, because it is just introduced. A few suggestions for improving this method are to use an other program than Graphviz. This can help to take away a few of the difficult points of this method. You can choose for another graph visualisation program, but also for a multidimensional scaling program. Another improvement is to extend the feature-analysis. For example, you can use a way of minimum distance between nodes to calculate the edges (Zwarts, 2010). You can also use not all the features, or classify features.

Another suggestion would be to try to rotate the space of containers to investigate if each language uses a near-optimal partitions of the space, like Regier et al. (2007) did with the colour space. You can also try another way of analyzing this research by Malt.

Of course you can use feature-analysis on other research. It is a relatively new type of analysis, so there are a lot of researches where you can use feature-analysis.

References

Croft, W. (2001). Radical construction grammar. *Oxford: Oxford University Press*

Hamming, R.W. (1950). Error detecting and error correcting codes. *Bell System Technical Journal,* **29**, 147-160

Haspelmath, M. (2003). The geometry of grammatical meaning: Semantic maps and cross-linguistic comparison. *The new psychology of language(ed.Tomasello),* **2**, 211-243.

Khetarpal,N., Majid, A., Malt, B., Sloman, S., &Regier, T. (2010). Similarity judgements reflect both language and cross-language tendencies: Evidence from two semantic domains. *In S. Ohlsson and R. Catrambone (Eds.), Proceedings of the 32nd Annual Meeting of the Cognitive Science Society.*

Kronenfield, D.B., Armstrong, J.D., & Wilmoth, S. (1985). Exploring the internal structure of linguistic categories: An extensionist semantic view. In J.W.D. Dougherty (Ed.). *Directions in cognitive anthropology* (pp. 91-113).

Malt, B.C., Sloman, S.A. & Gennari, S., Shi, M., Wang, Y. (1999). Knowing versus Naming: Similarity and the Linguistic Categorisation of Artifacts. *Journal of Memory and Language* **40**, 230-262

McCarty, J. (2007). What is Artificial Intelligence?

Regier, T., Kay, P., & Kheterpal, N. (2007). Color naming reflects optimal partitions of color space. *PNAS,* **104**, 1436-1441

Romney, A.K., Weller, S.C. and Batchelder, W.H. (1986). Culture as consensus: A theory of culture and informant accuracy. *American Anthropologist,* **88**, 313-338.

Zwarts, J. (2010). Semantic Map Geometry: Two Approaches. *Linguistic Discovery,* **8.1**, 377-395

Appendix

The appendix contains the graphs of the English language, the Spanish language and the Chinese language. The graphs are made with the neato format of Graphviz and the length of the edge is calculated through the hamming distance of one. The nodes are the objects as investigated by Malt and the hamming distance is the distance between the features of the objects.

Figure 6 contains the graph for English. In this figure, the red objects are named *container*, the green objects are called *jar*, the blue objects are named *bottle* and the yellow objects are called *can.* The white objects are named otherwise.

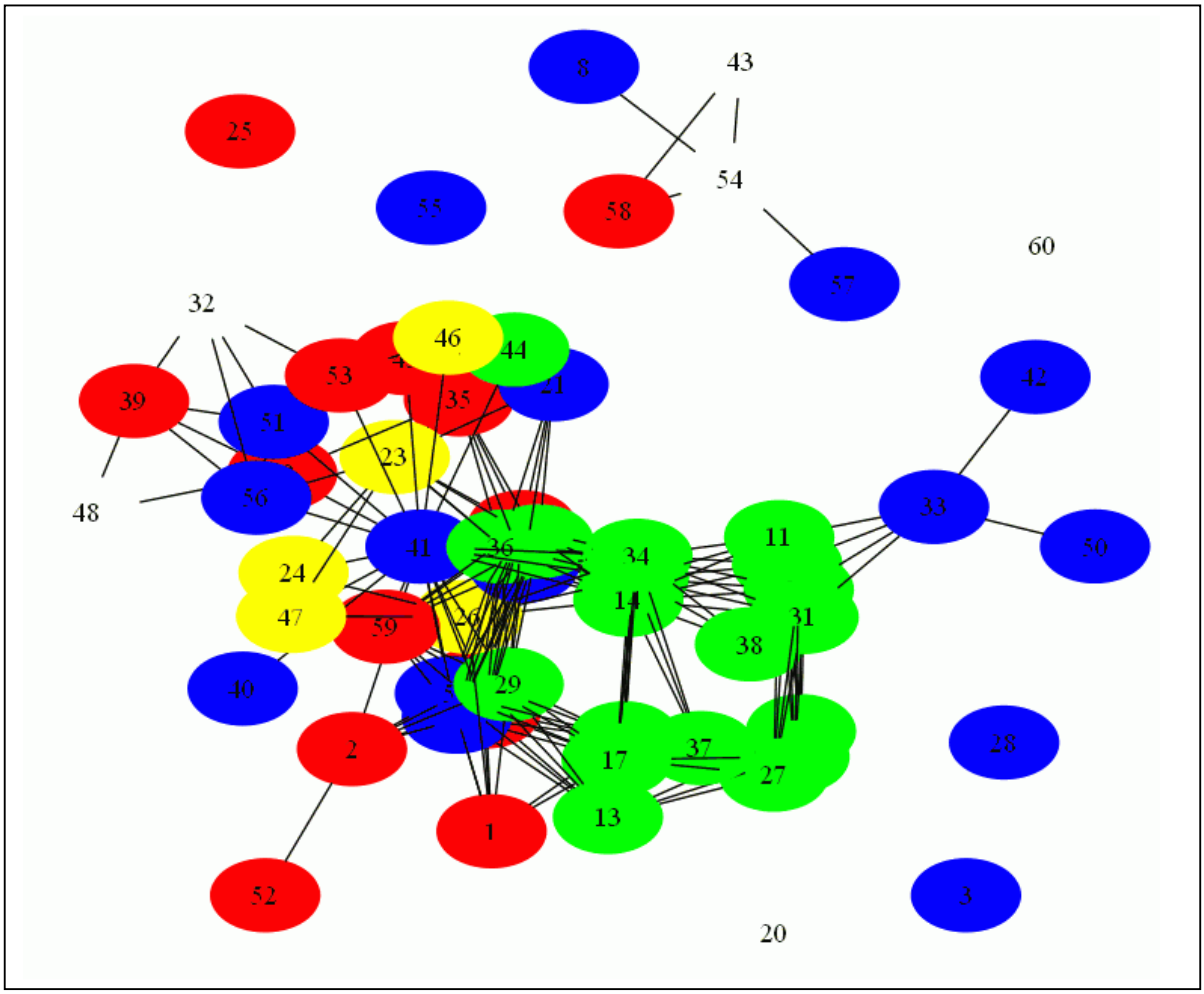


Figure 6: neato and distance for English

Figure 7 contains the graph for Spanish.
In this figure, the red objects are named *frasco*, the green objects are called *bidon*, the blue objects are named *envase* and the yellow objects are called *aerosol.*

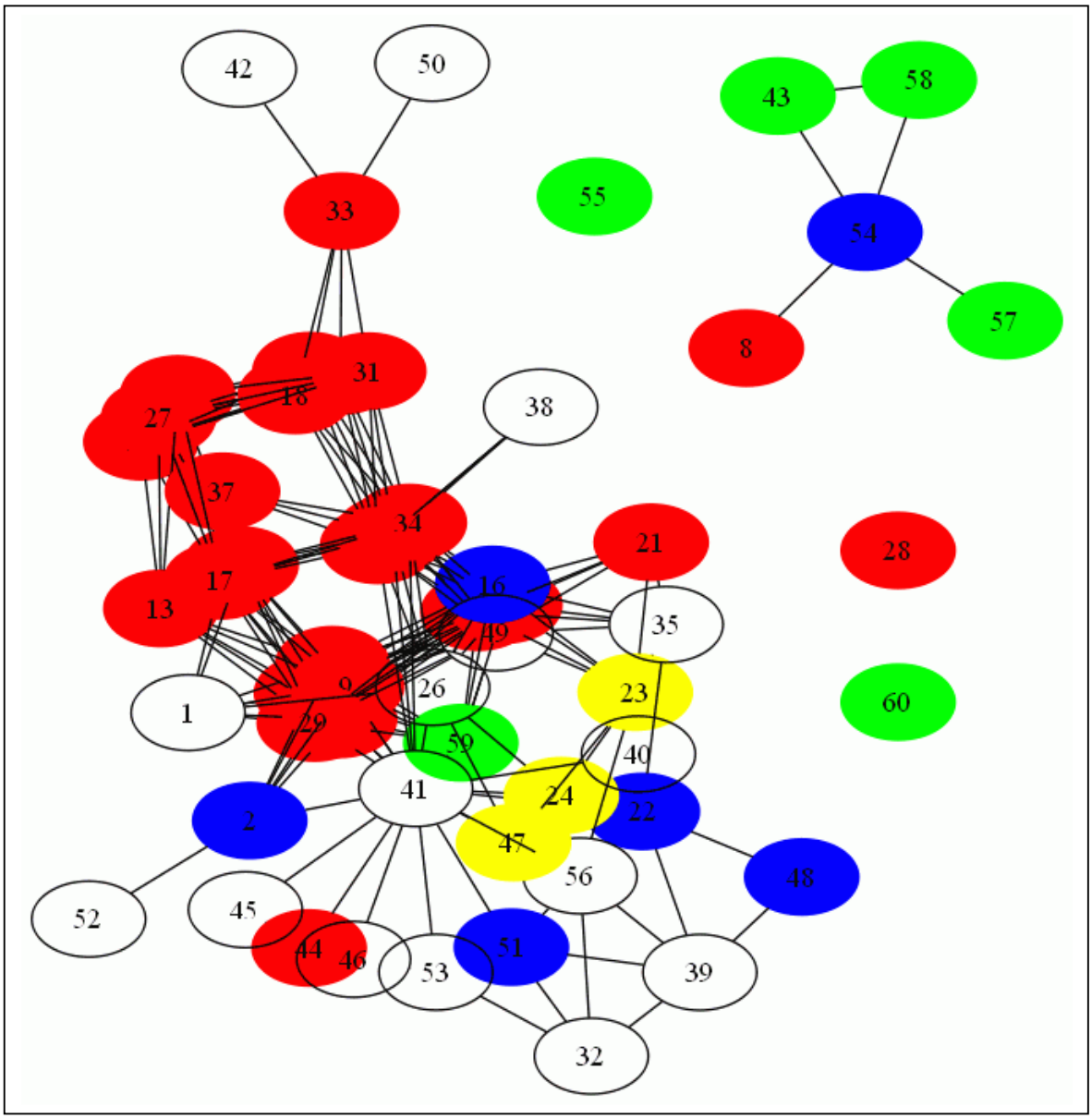


Figure 7: neato and distance for Spanish

Figure 8 contains the graph for Chinese.
In this figure, the red objects are named *ping2*, the green objects are called *tong3*, the blue objects are named *he2* and the yellow objects are called *guan4.*

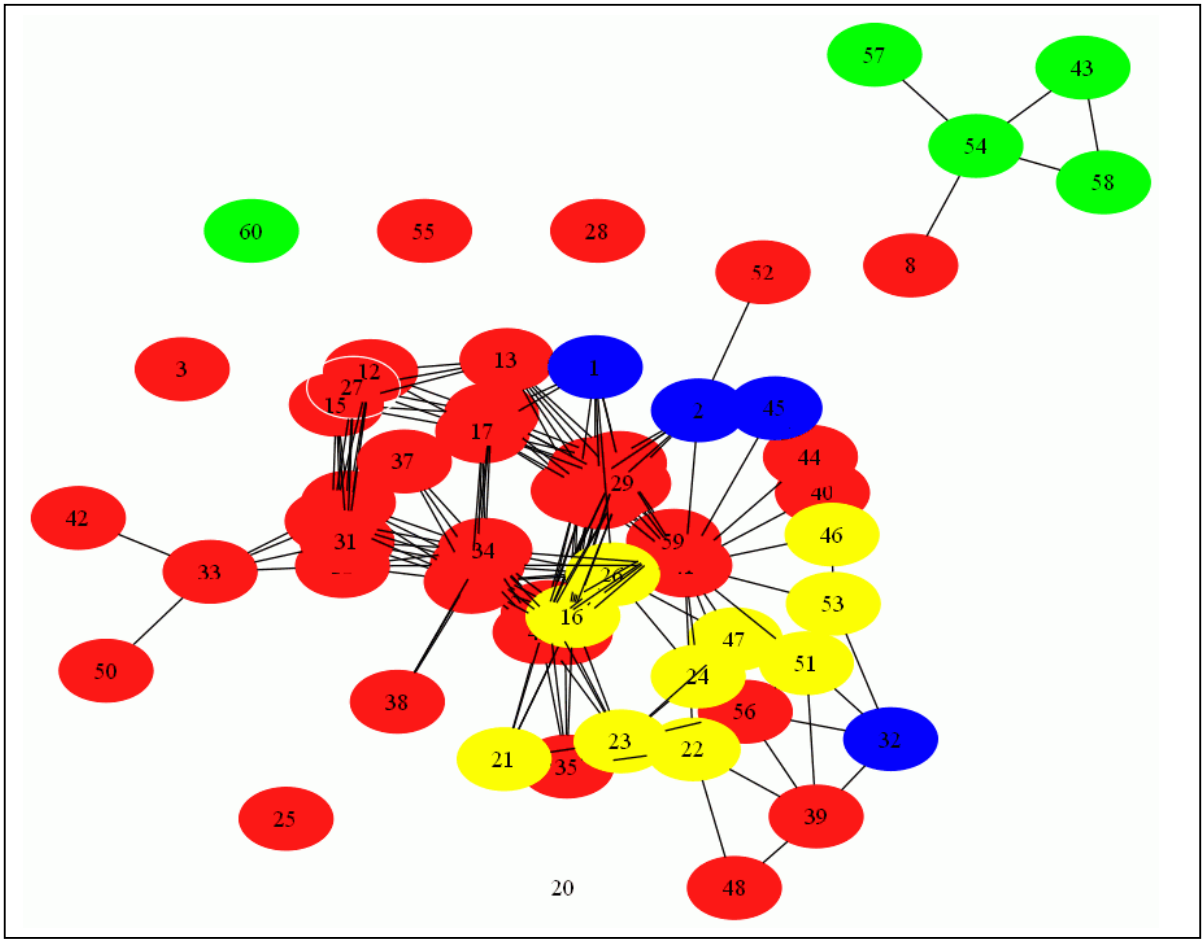


Figure 8: neato and distance for Chinese