**Preface**

**Compare it with the dendrogram of your clustering and discuss briefly.**

When we set the number of maximum clusters to 1, we have all the letters 'a', 'o', 'k' and 'e' in one cluster. As we increase the number of maximum clusters to 2, we have two clusters: a cluster containing 'a', 'o' and 'e', and another cluster containing just 'k'. As we continue to increase the number of maximum clusters to 3, 4, 5, and even 6, I noticed that there is no change in classification. This means that no matter how big the number of maximum clusters gets, there can't be more than 2 clusters. From this we can conclude that the letters are classified based on them being a 'vowel' or a 'consonant'. The letters 'a', 'o' and 'e' are all vowels, hence the distance between them is really small compared to the distance of each to the letter 'k', hence they can't be put in the same cluster with 'k', which is a consonant.

**Question 1**

**b) Which criterion do you think gives the most reasonable clusterings? Explain the differences between the criteria in your own words, and interpret their differences in clustering.**

The three criteria differ in the way they merge clusters and how the distance between each cluster.

Complete linkage hierarchical clustering measures the longest distance between two points in each cluster and merging the two clusters with the smallest maximum pairwise distance (clustering the two most dissimilar documents). Single linkage hierarchical clustering measures the shortest distance between two points in each cluster and merging the two clusters with the smallest minimum pairwise distance (clustering the two most similar documents). Average linkage hierarchical clustering, the distance between two clusters is defined as the average distance between each point in one cluster to every point in the other cluster.

Through inspection, I can see that Complete Linkage classifies the objects into their most appropriate clusters. You can tell that each member in a cluster shares a distinctive category, like how 'motorcycle', 'car', 'truck', 'boat', 'ship', 'helicopter', and 'rocket' are all vehicles, and how 'scissors', 'pen', 'pencil', 'screwdriver'. 'chisel', 'hammer', and 'knife' are all tools. There are some misclassifications like how 'telephone' is under the same cluster with other animals, but out of all the three criteria, I still think that the Complete Linkage criterion exhibits the most sensible and meaningful clusterings.

**Question 2**

**b) Which criterion gives the most reasonable clusterings according to the words helicopter and bottle?**

For the word 'helicopter', the most reasonable clustering is the Complete Linkage because it places 'helicopter' under the same cluster with the word 'rocket'. Both of these words have similar properties, such as how they represent 'vehicles' and they both have the behaviour 'flies'. Hence, it is only appropriate that these two words are clustered together. Furthermore, the cluster containing 'helicopter' and 'rocket' is also a member of a much larger cluster which includes the words 'motorcycle', 'car', and 'truck', which are also 'vehicles', but they travel on land (does not fly). In the Single Linkage clustering, 'helicopter' is clustered with 'rocket' as well, but then that cluster belongs to a bigger cluster that contains objects such as 'spoon', 'bowl', and 'pen', which falls under the 'tool' category. Average Linkage also clusters 'helicopter' with a set of tools.

The same goes for the word 'bottle'. The Complete Linkage clustering had clustered the word 'bottle' under the same cluster with the word 'cup'. Both of these words exhibit similar properties, which may include being kitchen tools that you drink liquids out of. Looking at the bigger cluster, the Complete Linkage clustering has also placed the cluster containing the words 'bottle' and 'cup' with the words 'telephone', 'bowl', and 'pen', which shares similar (but more general) properties and fall under the same category which is 'tool'.

**Question 3**

**a) How do the clusters obtained from the two vector representations (BNC and McRae) differ? Discuss which vector representations yield more intuitive and semantically coherent clusters, based on the clustering of the words helicopter and bottle.**

Overall, I think that the vector representations obtained from the McRae vector yields more intuitive and semantically coherent clusters than the BNC vector. The BNC vector representation tends to yield more individual clusters than subclusters, failing to group similar words together. The McRae vector representation has plenty of subclusters under the main clusters which groups words with similar features together. The McRae vector representation more successfully places words which are closely related featurewise in the same category, compared to the BNC vector representation.

The reason behind why the McRae vector cluster words better than the BNC vector is that the McRae vector works on the notion of clustering words that have common features, and these features are determined by humans themselves. So, clustering words with the McRae vector representation are closely related to the human intuition when it comes to categorisation of words. On the other hand, the BNC vector representation is based on the co-occurrence of a target word with a context word. So, when using the BNC vector to categorise things, it will be more accurate in categorising which words are more commonly used together than actually putting them in their appropriate semantic categories.

**b) What can you infer from the clustering with respect to the semantic category of chicken? Which representation (based on the BNC or McRae'sfeature norms) do you think yields a better clustering for it?**

In the BNC vector representation, the word 'chicken' is often placed within a cluster containing foods. In Single Linkage clustering, it is in a cluster with 'potato', and both of them are within another cluster containing the words 'mushroom', 'onion', and then in another cluster containing 'banana' and 'pear', which are all foods. In Complete Linkage clustering, the word 'chicken' is placed in the same cluster with 'potato', then with 'mushroom', 'onion', and 'lettuce', which are all under a bigger cluster containing foods such as 'pineapple', 'pear', 'banana', and 'cherry'. Finally, this trend of semantically categorising 'chicken' as food is also observed in the Average Linkage clustering, where 'chicken' is again placed under the same cluster with 'potato', then under a cluster with 'lettuce', 'mushroom' and 'onion', and then with 'pineapple'.

However, in the McRae vector representation, the word 'chicken' is often clustered with words that are semantically categorised as animals instead of food. In Single Linkage clustering, it is first clustered with the word 'peacock', then with 'eagle', then 'duck', 'owl', and then finally with 'penguin' and 'swan', which are all animals. 'Chicken' is also placed in a group of animals in Complete Linkage clustering. It is first clustered with 'peacock', then with 'eagle' and 'owl', and then with 'duck', 'penguin', and 'swan'. Finally, in the Average Linkage clustering, 'chicken' is first clustered with 'peacock', then with 'eagle', 'owl', and then with 'duck', 'penguin' and 'swan'. This further confirms the trend that in the McRae vector representation, the word 'chicken' is semantically categorised as an 'animal' and not a 'food'.

The fact that 'chicken' is categorised as a food in the BNC vector representation and an animal in the McRae vector representation shows us that the word 'chicken' frequently co-occurs with other words that fall under the 'food' category, meaning it is often contextually referred as a food than as an animal in British English. However, in the McRae vector representation, the word 'chicken' falls under the semantic category of 'animal' because it exhibits more properties that are distinctive towards animals, such as being a living thing and is able to move.

As to which representation yields a better clustering, I think it would be the McRae vector representation because it shows which semantic category 'chicken' falls into, whilst the BNC vector representation shows how often the word 'chicken' co-occurs with words in a particular category in daily usage.

**Question 4**

**Can you draw any conclusions regarding which semantic space representation (BNC vs McRae) and which clustering criterion works best? Discuss below.**

None of the semantic space representation perfectly matches the Gold Standard classification. There are several misclassifications that are unique to one semantic space representation but there are also some that are common in both.

The first common misclassification lies mainly in the category tool-artifact-artifact. Most of the BNC and McRae semantic space representation placed them in several different clusters. The reason behind this misclassification might be because the category 'tool' is very vague and not distinctive. To solve this issue, more categories can be added, and words within a cluster will have more similarity. For example, instead of putting 'knife' and 'telephone' under the cluster 'tool', you can put 'knife' along with 'spoon' in 'kitchen utensils' and 'telephone' in 'electronic devices' along with 'kettle'.

In addition to that, both the BNC and McRae semantic space representation tend to just combine the categories green-vegetable-natural with fruitTree-vegetable-natural under one big cluster which looks like it would be the category 'food'.

Furthermore, as I discussed in Question(3(b)), the BNC space representation classifies 'chicken' as something that belongs to the category 'food'. However, there is no such category called 'food' in the list of Gold Standard classes. You can't necessarily fit 'chicken' under 'green-vegetable-natural' nor can you fit it under 'fruitTree-vegetable-natural'. Therefore the misclassification of 'chicken' in any BNC space representation is as predicted.

The BNC space representation cluster the words based on the frequency of their co-occurrence with other context words in the corpus, and not their features. The Gold Standard classes differentiate the category 'animal' into two clusters: bird-animal-natural and groundAnimal-animal-natural. Since the BNC does not contain the features of each animal as well as the distinctive features of birds and ground animals, it does not know whether an animal falls under the 'bird' or the 'ground animal' category.

After a thorough inspection of the clusters, I think the semantic space and clustering criterion that works best is the McRae semantic space and the Average Link clustering criterion. Despite several misclassifications and overgeneralizations, I feel like the clusterings placed words in their relatively appropriate and meaningful semantic categories.

A probable reason behind why the McRae semantic space is much more suitable than the BNC semantic space is because the McRae semantic space clusters the items based on the similarity of their features. On the other hand, the BNC semantic space counts the frequency of the co-occurrence of the target word with the words the BNC corpus. Basically, the BNC does not tell you how similar two objects are, it tells you more about how often are the two words used together contextually.

Also, the Average Linkage clustering criterion provided the best classifications. This might be because the Complete Linkage clustering pays attention to clustering outliers and forcing them into what it thinks is the most appropriate category and that the Single Linkage clustering has a tendency to form individual clusters and link single words, forming a long chain of words that do not possess any meaningful similarity.

**Question 6**

**Which words were correctly classified, and which ones weren't?**

**Can you think of an explanation for the classifications and misclassifications?**

Correctly classified words = airplane, bathtub, broccoli, calf, dagger, fork, guitar, razor

Incorrectly classified words = belt, blueberry, hawk, house, missile, salmon, tank, worm

BNC classifies words in the notion of the frequency of a target word co-occurring with a context word. Hence, it places words that are often used together in the same category, but it does not necessarily mean that they belong to the same semantic category.

Let's take a look at the misclassified words and the possible reasons why they were misclassified.

At first, I was confused as to why a 'blueberry' is classified as a 'bird' since it is obviously a fruit and it does not fly. However, I realized that blueberries do not grow on trees. But that still does not explain why it is classified as a bird. Reminding myself that the BNC vector representation classifies words based on co-occurrence frequencies, I typed in 'blueberry bird' in a search engine and found plenty of articles about how to keep birds from eating your blueberries. This frequent co-occurrence between the word 'blueberry' and 'bird' is the most probable reason why the model misclassified 'blueberry' as a 'bird'.

It is obvious that most of the animals are misclassified. Out of 4 animals (calf, hawk, salmon, and worm), only 'calf' was correctly classified. Unlike the McRae vector representation, BNC does not really classify words based on their features but based on their co-occurrence with other words instead, which results to illogically classifying such words in a category that lacks similarity featurewise but are often contextually used together in British English.

The words 'tank' and 'missile' are rarely used in daily British English and is more often found in military-related text/speeches. Hence, the word 'tank' does not occur with other vehicles and 'missile' does not occur with other tools regularly. This lack of co-occurrence causes the BNC vector representation to misclassify the words 'tank' and 'missile'.

For the word 'house', it is misclassified because I don't think people often use it contextually while referring to places to stay in. For example, a travel brochure might have the words 'hotel' and 'resort' when referring to accommodation, but rarely 'house'. Therefore the word 'house' rarely co-occur with words that fall under the 'accommodation' or 'lodging' category.

**Question 7**

**a) How does the cosine similarity differ from the Manhattan distance similarity? Give a pair of vectors that are close according to cosine distance but far according to Manhattan distance.**

Similarity in general measures how alike two data objects are. The distance measured represents the features of the objects. If the distance is small, the degree of similarity will be high. Otherwise, if the distance is large, the degree of similarity will be low.

Cosine similarity allows us to find the cosine of the angle between the two vectors. Cosine similarity measures the orientation of two vectors and not particularly its magnitude: two vectors with the same orientation will have a cosine similarity of 1, and two vectors that are right angled to each other (90 degrees) have a similarity of 0.

Manhattan distance similarity measures the distance between two points by calculating the sum of the absolute differences of their Cartesian coordinates. It is the total sum of the difference between the x coordinates and y coordinates. Hence it is the distance between two points measured along axes at right angles. The use of Manhattan Distance is only useful when you have a large dataset and you want to 'ignore' outlier data.

**b) Look at the differences in classification for the McRae and BNC data when using the two different distance measures.**

The Manhattan distance similarity classified almost all the words wrong in the BNC data, but correctly classifies some of the McRae data. Cosine similarity, although not perfect, is more suitable to be used to measure the distance for both vector representations.

**Try to explain the differences in classification.**

Cosine similarity is used to calculate the angle between two vectors. In this case, the magnitude of the vectors does not matter as much. A reasonable example to use cosine similarity aside from Manhattan Distance is when you want to count how often a certain word appears in a document. For example, the word 'Law' occurs more often in document A than document B, Cosine Similarity is able to conclude that document B is more related to the topic 'Law' than document B, despite the fact that document A might be longer than document B, because the magnitude of the vectors does not affect the conclusion drawn. The BNC dataset is not suitable to be used with Manhattan Distance, because Manhattan Distance can be influenced by the length of the vectors.

**Question 8**

Compare the predictions of the two semantic spaces. Which semantic space is better at inferring the categories of the new words? What limitations do you see with these representations? Can you think of ways to overcome them?

The semantic space that is better at inferring the categories of the new words is the ones that use the McRae vector representation. This is because the McRae dataset is based on features described by humans themselves, hence their method of classification is very close to real human intuition. Aside from that, I feel that the McRae dataset is designed to aid semantic categorisation based on features. On the other hand, the BNC dataset is more tailored to identify the frequency of the co-occurrence of a target word with a context word in the usage of British English. However, the frequency of the co-occurrence of words does not necessarily define if they fall into the same category, hence the BNC dataset is not really suitable for the purpose of semantic categorization.

Theoretically, the use of the BNC semantic space makes sense because words in the same categories tend to be used together, like how a grocery list would contain words such as banana, cherry, pear, pineapple, etc. However, words in the BNC can be used in different ways contextually. In addition, the BNC is unable to further classify a word because it does not have the feature of a certain word. For example, it is unable to determine whether an animal is a ground animal or a bird because it does not know the defining features of either category. Hence I believe the use of the BNC semantic space is rather inappropriate when we are trying to cluster words into their respective semantic categories.

The McRae semantic space is not perfect either. However, because it is more feature-oriented when it comes to categorising words, it is more appropriate to use in semantic categorisation. I think the reason behind most of McRae's misclassifications is because the number of words and categories used are very few. We can overcome this limitation by adding more words and categories so that we can clearly see the similarity between words in a cluster and the dissimilarity of words in different clusters. In addition to that, we can always add more distinctive features to make sure that words that fall under the same cluster are very closely related (share more than a certain number of definitive features). Together, this will make the categories much more distinguishable.