

**The conceptualisation of FEMININITY on English  
Wikipedia**

*Max Bäckström*

LUND UNIVERSITY  
ENGL01, HT2012  
English, Linguistics  
Supervisor: Dylan Glynn

*Max Bäckström*

## **Contents**

<b>1. Hypothesis...</b>	<b>2</b>
<b>2. Method...</b>	<b>3</b>
<b>3. Analysis...</b>	<b>3</b>
<b>4. Results...</b>	<b>13</b>
<b>5. Summary...</b>	<b>22</b>
<b>6. References...</b>	<b>23</b>

## **The conceptualisation of FEMININITY on English Wikipedia**

*Max Bäckström*

### **1. Hypothesis**

English Wikipedia is today one of the most widely used websites on the internet; with over 4 million articles and an annual growth rate of 9% for 2011 it is the world's largest encyclopaedia with a content 50 times greater than that of the second largest encyclopaedia in the English language (<http://en.wikipedia.org/wiki/Wikipedia:Statistics>, 2012-12-20, 14:20).

However, what makes Wikipedia the most interesting and what radically differentiates it from similar resources and other types of media is that the complete contents of Wikipedia are exclusively written by its users with little or no control over content, meaning that it is a highly interesting and useful resource for examining the use of language in regards of cultural information, building on information put forth in previous research made by Lakoff (1987) and Wierzbicka (1985).

The representation and conceptualisation of FEMININITY has been chosen to be looked at in this study due to its likeliness of user influenced usage and thus likely to yield interesting results. Due to the fact that Wikipedia has contributing users from all over the world, with many different languages as their primary, the study will be done on what is known as 'international English' (Seidlhofer 2001, Brutt-Griffler 2002).

Different subjects vary quite significantly from one another in terms of formality, status and how they are perceived by the public, e.g. the subject of religion differs quite considerably from those dealing with popular culture. Taking this into consideration for the study, the research question will be on whether or not the conceptualisation and representation of FEMININITY differs between different articles on Wikipedia depending on the subject that they are covering. Seeing as Wikipedia is striving for objectivity and little or no influence of the article authors' subjective values should be present in an encyclopaedia, this study is very likely to produce interesting results.

## 2. Method

The method in this study uses samples of natural language data and treats them with various statistical methods in order to find usage patterns. The initial step of this study was to create a corpus out of the entire English Wikipedia and all of its articles; Wikipedia themselves supply dump files of their entire article database for free download, simplifying this process quite significantly. After having acquired the most current dump file at the time, containing all current articles as of 2012-10-01 and having looked into the possibility of making the dump file a searchable corpus, a python script with that specific function was found. However, after having let the script run for two full days, dividing each article into an individual text file, it was concluded that a lack of computer power meant that the process would be too extensive for a study on this level.

After further research into the subject of converting the dump file into a searchable corpus, an article by Stephen Marquard of Cape Town University on the subject was found. After consulting Stephen, a workable shell script was executed, successfully converting the dump file into a fully searchable text file. Following this, five lexemes of significance in regards of conceptualising FEMININITY were chosen; *ladylike*, *girly*, *girlish*, *feminine* and *womanly*. These lexemes were chosen on their basis of being likely to be used in order to represent FEMININITY in significantly different ways, 100 occurrences of each of these lexeme were then extracted from the dump file.

Following this, the data was analysed in regards of the features further explained in section 3; this process was significantly time consuming and much of the analysis was thought of as arbitrary whilst it was being done. Finally, the data was treated with usage-feature analysis (Geeraerts et al. 1994, Glynn & Fisher 2011) in order to be able to examine patterns in the language usage and produce the actual results; several different statistical methods were used, looking at different aspects of usage.

## 3. Analysis

The initial process in determining the variables for analysing the data were looking and examining the actual examples, trying to determine what features would be and would not be of interest in answering the hypothesis. Due to very little experience with this type of research, variables, which turned out to be of no interest when actually examining the results, were

added as well. Apart from purely linguistic features of the examples, a variable for topic was added; topic in this case meaning what subject the Wikipedia article it was taken from is covering.

As mentioned in section 2, five lexemes, *feminine*, *girly*, *girlish*, *womanly* and *ladylike*, were examined in the study and below is a detailed explanation of the variables used in the analysis, some of which have extracts from the examples for explanatory reasons.

### 3.1. Example

The actual text in which the lexeme is found, taken from the Wikipedia corpus which was created from the dump file of the entire English Wikipedia at 2012-10-01.

### 3.2. Lexeme

Which lexeme occurs in the example.

### 3.3. Lexeme 2

A combination of lexeme *girlish* with *girly* and *feminine* with *womanly*, due to their proven similarities.

### 3.4. Topic

The topic of the Wikipedia article where the example was taken from, this first category contained a high amount of different topics. In order to find the article where the example had been taken from, the string of text was searched for on a search engine and the Wikipedia article was generally the first hit. However, due to the changing nature of Wikipedia, a few instances of difficulty finding the actual article occurred. Due to how important this variable was to the result, in cases where the article was not possible to find, the example was replaced with one that was actually in an existing article, in order to avoid NA's.

*Max Bäckström*

### 3.5. Topic 2

A coarser version of topic where a vast majority of the topics in the initial category were clumped together to create more workable data; the categories in this variable were Art, Society, Culture, Fashion, Female artist, Japanese culture, Female historical person, History, Literature, Misc, Movies & TV, Movies and TV (Animated), Music, Religion, Society, Sexuality, Science and Tobacco.

### 3.6. Topic 3

This category contains an even coarser version of the topics. The categories were Socio-Cultural, Person, History, Movies & TV, Misc and, interestingly enough a high amount of articles relating to Japanese culture were found, 88 occurrences, being the third most common after Movies & TV with 103 occurrences and the most common one was Socio-Cultural, with 171 instances found.

### 3.7. Collocation

The variable collocation contains the word which the lexemes often occurred with; a majority of the times it was the word which stood next to the given lexeme, however, this was not always the case and related words could be found in another clause or even sentence. For example, in “*combination of innocence and womanly awareness*”, “innocence” was chosen as the collocation.

### 3.8. Construction

A category containing the grammatical construction of the example where the lexeme occurs, if possible, the words prior to and after the lexeme were used to describe the construction. The most common result was “Determiner\_Adjective\_Noun, with 129 occurrences, followed by “Adjective\_Adjective\_Noun” at 82.

### 3.9. Construction 2

A coarser version of construction, where the most irrelevant of the two words chosen together with the lexeme, was removed. Due to the lexeme most of the times being used as an adjective, this resulted in a great number of examples with “Adjective\_Noun” as the grammatical construction.

### 3.10. Word class

The word class of the lexeme was recorded in this variable, due to the already mentioned nature of the word, a great majority of the words were adjectives; hence a distinction between attributive and predicative adjectives was made. Apart from adjectives, a few occurrences of noun and adverb usages were found. In a majority of the examples, the lexeme was used as an attributive adjective, with 375 instances, predicative adjective being the second most common at 123 occurrences and only one usage each of adverb and noun.

### 3.11. Word class and lexeme

A combination of the word class and the lexeme itself in order to examine whether or not the grammatical class of the word played a part in distinguishing its usage. The attributive use of *womanly* and *feminine* was by far most common and predicative usage of *womanly*, *feminine*, and *girlish* was very sparse.

### 3.12. Part of construction

The grammatical part of the sentence in which the lexeme was located; for example, as a subject in example 1

(1) *feminine lines and colours began to appear in the late 1930s*

and as a direct object in example 2

(2) *seeks to create a distinctive feminine persona.*

*Max Bäckström*

The lexeme was predominately found being used as a subject with 190 occurrences and second most commonly as a direct object being found 170 times.

### 3.13. Verb

The main verb of the clause where the lexeme occurs, which was noted in the infinitive. The most common verb was understandably *be*, found 117 times, and *have* as second most common occurring 29 times.

### 3.14. Emphatic

Whether or not the lexeme was emphasised was examined in this category; very few occurrences of emphasised usage were found in the examples, only 29 instances. There were various constructions which at times were unclear as to whether they were actually emphasised or not, most notably the word *more* caused these issues, where comparative was mistaken for emphasis, an instance of this can be seen in example 3,

(3) *Velma even goes to the point of trying to be more girlish to try and gain his affection.*

The word *more* in this case does not entail emphasis; it is used in contrasting an earlier state of less girlishness to the present one.

### 3.15. Axiology

This category contained the alternatives ‘positive’, ‘negative’ and ‘neutral’. It was used to examine whether or not the lexeme was used to portray things with different attitudes. Due to the subjective nature of this category, it was often troublesome to determine the proper axiology of the example, hence it was assumed that it was neutral unless markedly positive or negative. 58 occurrences of negative axiology were found in the study, the negative nature of the usage was often amplified by the words used together with the lexeme, as can be seen in example 4,



(4) *who is vain and girlish, mischievous, lighthearted, coquettish and gossipy.*

Where words *mischievous*, *vain* and *gossipy* have a negative tone to them and hence making the entire sentence of a negative nature. An example of positive usage related to a given lexeme can be seen in example 5 below,

(5) *As such it was common to hear praise of womanly virtues as though they were divine.*

As with the example of negative axiology, the positive axiology of example 5 is enhanced by the word *virtue*.

### 3.16. Theme

The theme of the example was a way of determining what was talked about using the given lexeme. It was not always completely clear what the primary theme of the example was, e.g. if the wearing of a *feminine* dress is being described, whether it was an action or a matter of appearance was not always easily determined. The most commonly occurring theme was “Way of being”, 231 instances, example 6 being a clear demonstration,

(6) *He agreed but the girl rejected Shivshankar for being too feminine*

And with “Appearance” being the second most common at 100 occurrences, shown in example 7 below,

(7) *possibly implying that he had a womanly face.*

Further features within this variable were action, example 8,

(8) *When a man marries and is about to offer himself to men in womanly fashion*

Thing, example 9,

(9) *his best friend, Aly's, girlish birthday presents*

Attribute, example 10,

*Max Bäckström*

(10) *That distinctively girlish vocal quality inclined Tattermuschová*

Moreover, features relationships, life and femininity were found, however, they had very few and rather uninteresting occurrences.

### 3.17. Referent 1 Type

The referents of the lexeme obviously play a big part in determining the usage of it. In Referent 1 Type, the actual person or thing referred to was determined as either specific human, generic human, concrete or abstract. Example 11 shows an occurrence of a concrete first referent, being the word *costumes*,

(11) *The womanly costumes were engraved*

Example 12 illustrates an abstract first referent, the word *wiles*,

(12) *used her womanly wiles and devious mind to manipulate those around her.*

Finally, example 13 shows the usage of generic human as the first referent, in this case *women*,

(13) *he did not admire very feminine women.*

The most common outcome of this category was abstract things with more than half of the examples, which is quite surprising results, a majority of these abstract referents were attributes of the second referents, explained below. Furthermore, the most sparsely used was generic human with merely eight examples.

### 3.18. Referent 1 Animacy

The referents in the Referent 1 Type were narrowed down to either animate or inanimate objects, adding up to 112 animate referents and 388 inanimate ones.

### 3.19. Referent 1 Animacy 2

The same thing was done in this category as in the previous one, however, with the difference that concrete and abstract things were still separated and all human referents were clumped together into one category, leading to a total of 112 human referents 324 abstract ones and 64 concrete.

### 3.20. Referent 2 Type

As with Referent 1 Type, the distinction between concrete, abstract, specific human and generic human was made. In contrast to Referent 1 Type, this variable had a specific human as its outcome in a majority of the examples; this is due to the fact that it was often a person performing an action or having an attribute described with the lexeme who was the second referent. An occurrence of a specific human as second referent can be seen in example 14 where *Rosie* is the second referent,

(14) *Rosie knows that she can use her womanly wiles to get what she wants.*

Seeing as a second referent wasn't always available, approximately one fifth of the examples were NA. In the cases where a second referent was not available and where this category was not applicable, there was only a single referent and this referent did not belong to or was the attribute of any other, i.e. second, referent.

### 3.21. Referent 2 Animacy

The same thing was done here as in Referent 1 Animacy, i.e. determining between animate and inanimate, this time in regards of the second referent. The results shows that there were 287 animate referents, 66 inanimate and a total of 147 NA's.

### 3.22. Referent 2 animacy 2

*Max Bäckström*

As in Referent 1 Animacy 2, whether the second referent was concrete, abstract or human was looked at. As in the previous category, there were 287 human referents with 50 abstract ones, 16 concrete and the same number of NA's, i.e. 147.

### 3.23. Referent Mix Type

Referent 1 Type and Referent 2 Type were combined in order to create this category. Due to the fact that the second referent was often the most significant one of the two, instances where Referent 2 Type had been NA, it was replaced with what had been the outcome of Referent 1 Type, hence eliminating the NA's and creating a category with more interesting and workable data. In total, 82 abstract, 36 concrete, 48 generic human and 334 specific human were found.

### 3.24. Referent Mix Animacy

The same procedure was done here as with Referent Mix Type, but with animacy instead of the actual referent, which means that when the animacy of Referent 2 Animacy was NA, it was replaced with that of Referent 1 Animacy, creating a more extensive category. After this had been done, there were 382 animate referents and 118 inanimate ones.

### 3.25. Referent 1 gender

The gender of the first referent was determined; no difficulties were encountered in doing this, as it was often quite clear. There were 99 occurrences of female, first referents, 17 male and a clear majority of NA's with 384.

### 3.26. Referent 2 gender

The same procedure as with Referent 1 Gender but with the second referent, a similar process meant that no difficulties were found here either. In this category, much fewer NA's were encountered, a total of 210 with 72 being male and 218 being female.

### 3.27. Referent Mix Gender

The NA's of Referent 2 Gender were replaced with the gender of Referent 1 to get a category with more substance, the combination yielded a result of 308 female referents, 87 male and 105 NA's; the NA's being cases where the gender of both referents was undeterminable.

### 3.28. Referent noun

Seeing as the lexemes were used as adjectives in almost all cases, this means that there was also a noun being referred to in them, e.g. *shows* being the referent noun in example 15,

(15) *while she watches girly shows on the big TV upstairs.*

There was a wide variety of nouns being referred to, in cases where it was a personal pronoun, specific human was put as the Referent Noun and if it was a name or similar, personal and geographical names were noted as Proper Noun.

### 3.29. Referent noun coarse

A coarser version of the Referent Noun with the different nouns divided into fewer categories to get a more manageable and workable data.

### 3.30. Referent noun coarse 2

When examining and plotting the results, it was noticed that even the coarse version of Referent Noun had too many categories; hence an even coarser version was created in order to be able to use the information properly. The categories in this variable were action, appearance, human and thing. The most common noun of reference was of human nature with 144 instances, however, both action and thing came very close to the same number at 140 each and appearance had 76 occurrences.

*Max Bäckström*

### 3.31. Negation

Only a few examples, 19, of all 500 were used with negation, due to this data sparseness, negation was not significant in plotting and examining the conceptualisation of FEMININITY.

### 3.32. Sexual

Whether or not the example had to do with sexuality was examined in this category; even though there were only 22 examples of sexuality involved, with example 16 being one of the examples where sexuality was highly involved,

(16) *deals with an aspect of the feminine experience, touching on matters such as sex, love, rape*

The instances where they were found were of interest when plotting the data at a later stage.

## **4. Results**

After the coding had been completed, the data was looked at with different statistical techniques and several different plots were produced to be able to interpret and look at the results. Each plot contains chosen variables believed to provide results capable of answering the research question as well as leading to new ones.

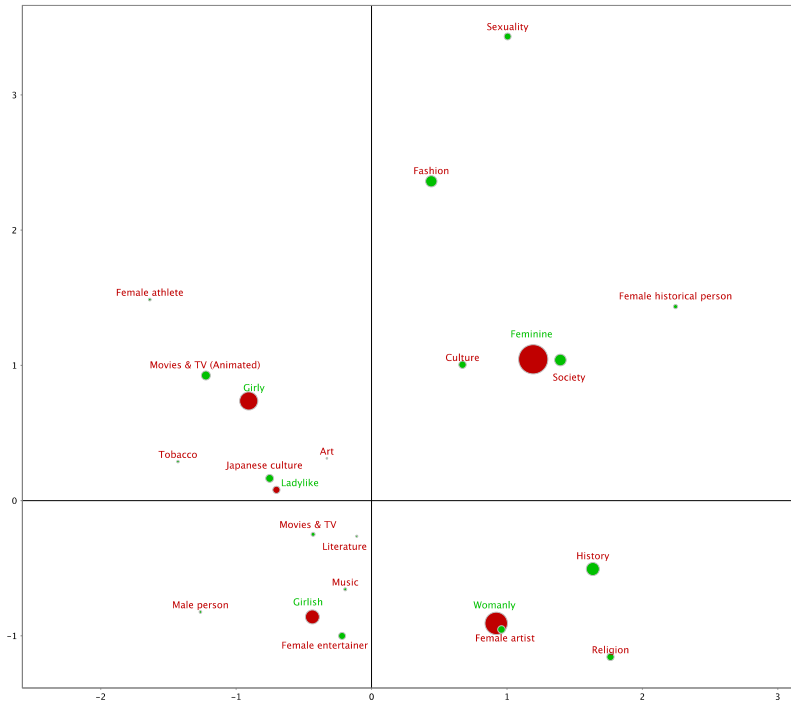


Figure 1. Multiple correspondence analysis of article topic and lexeme

Figure 1 demonstrates how the different lexemes of FEMININITY are used in relation to the various topics included in the study. As can be seen by the positioning of the lexemes, with four distinct clusters, this plot clearly shows that there are four conceptualisations of FEMININITY on Wikipedia. Firstly, the positioning of the lexeme *feminine* shows that it correlates primarily to articles covering the topics of fashion, sexuality, culture, society and female historical persons. Most notably the topics fashion and sexuality have a stronger correlation with *feminine* due to their clustering further from the centre of the plot and the other lexemes. However, it should be considered that the positioning of female historical person is most likely influenced by data sparseness. Moreover, the large size of the circle representing *feminine* indicates that this lexeme has a strong influence on structuring the plot and the result.

Secondly, the lexeme *womanly* is positioned in the bottom right of the plot, clustering with the topics female artist, history, and religion. The topic female artist has placed directly on top of the circle representing the lexeme, meaning that there is a strong correlation between the lexeme

Max Bäckström

*womanly* and articles on the topic of female artists. Although not represented by a circle as big as that of *feminine*, the size of *womanly* indicates that this is also highly influential in structuring the plot.

Furthermore, *girlish* has placed at the bottom left of the plot, clustering with the topics movies & TV, female entertainer, literature, music and male person. The strongest correlation with the lexeme can be found in topics concerning male persons. The whole cluster, including the lexeme itself, is positioned fairly close to the centre of the plot, in combination with minor circles it can be concluded that *girlish* as well as the topics correlating with it, especially movies & TV, literature and music, do not play a very big part in structuring and determining the outcome of the plot.

Finally, lexemes *girly* and *ladylike* have clustered at the top left of the plot, correlating with topics movies & TV (animated), Japanese culture, art, tobacco and female athlete. *Girly* has a stronger influence on structuring the plot as it is represented by a bigger circle, *ladylike* is also located closer to the centre, hence it is not as significant to the clustering. Movies & TV (animated) has a strong connection to *girly*, whereas tobacco and art are inclining towards *ladylike*. However, since *ladylike* is close to the line, it means that it has a connection to all topics on the left side of the plot and not only the ones in the top left corner. In a similar manner, Japanese culture is also positioned close to the line, indicating correlation with all three lexemes *ladylike*, *girly* and *girlish*, having a stronger connection to *girly* than the other two.

In order to confirm the results seen in *figure 1*, the next step was to apply another statistical method to determine the actual number of distinct clusters. By combining the same variables, i.e. lexeme and article topic, *figure 2*, below, was produced.



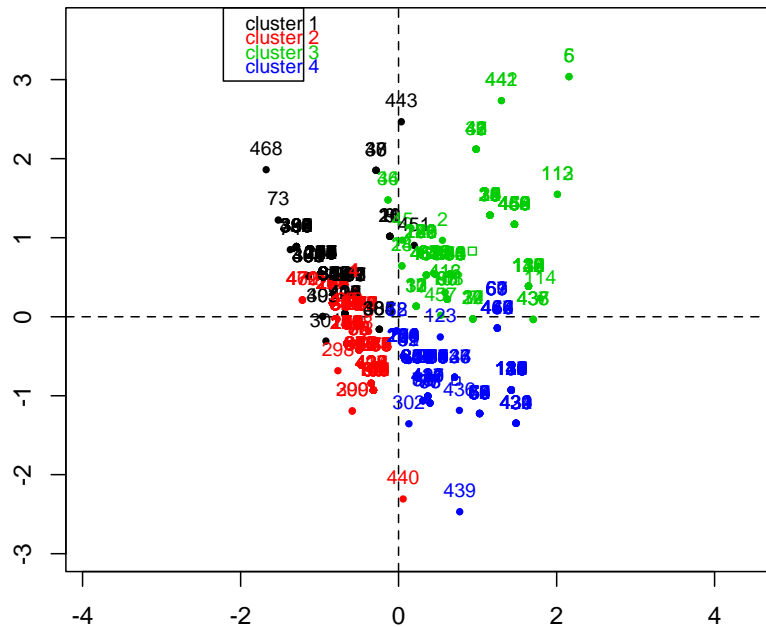


Figure 2. Factor map of article topic and lexeme

The factor map presented in Figure 2 was, as mentioned, produced with the lexemes and article topic as variables. As can be seen there are four different clusters in regards to the correlation between topic and lexeme, similar results to what can be seen in the multiple correspondence analysis in Figure 1. This distinctive four way clustering leads to the conclusion that in regards to article, topic and the lexeme used, there are four distinctive conceptualisations on FEMININITY on English Wikipedia. However, it should be noted, as was also seen in *figure 1*, cluster 1 and cluster 2, containing the lexemes *girly*, *girlish* and *ladylike* have tendencies of overlapping and mixing in contrast to cluster 3, *feminine*, and cluster 4, *womanly*, which are more distinct.

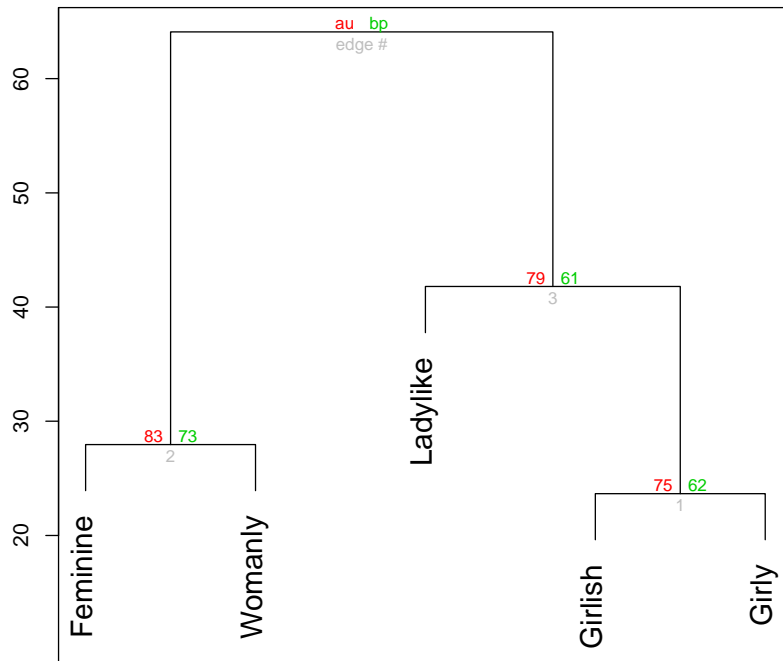


Figure 3. Hierarchical cluster analysis of lexeme with topic and referent gender

This HCA was done with the lexemes as the primary variable; in addition, the topic and the referent gender were used to produce the result. The analysis shows quite clearly that there are two main clusters, with two lexemes in each cluster, as well as one lexeme being positioned further from the others. The left cluster holds the lexemes *feminine* and *womanly* whereas the rightmost cluster is made up of lexemes *girlish* and *girly*. *Ladylike* has clustered on the same branch as *girlish* and *girly*, indicating that it has a stronger correlation with these two lexemes than with *feminine* and *womanly*, but due to its independent positioning, it can be concluded that it is indeed different from *girlish* and *girly*.

This way of clustering corresponds to that which was shown in *figure 1* and *figure 2* in which *girlish* and *girly* were overlapping one another and *ladylike* being positioned in between as a related, but less distinct, sense of FEMININITY. Moreover, by looking at the bootstrapped percentage numbers on top of the clusters, it can be seen that *feminine* and *womanly* are less likely to be chance due to its 83% and 73% chance, compared to that of *girlish* and *girly* with a 75% and 62% chance of not being chance; this is

supported in *figure 2* where the clusters representing *girlish* and *girly* has a much greater overlap than that of *feminine* and *womanly*, meaning that *feminine* and *womanly* are more distinguishable.

```

Deviance Residuals:
Min      1Q  Median      3Q      Max
-1.9761 -0.9936  0.5533  0.8956  2.0448

Coefficients:
Estimate Std. Error z value Pr(>|z|)
(Intercept)          0.7749    0.4223   1.835 0.066546 .
Topic.4History      -2.1138    0.6112  -3.458 0.000544 ***
Topic.4Japanese culture  1.0247    0.3929   2.608 0.009109 **
Topic.4Movies & TV    0.5512    0.3642   1.513 0.130193
Topic.4Socio-Cultural -0.6043    0.3210  -1.882 0.059791 .
AxiologyNeutral     -0.6197    0.3474  -1.784 0.074440 .
AxiologyPositive   -1.5344    0.4667  -3.288 0.001011 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Logistic Regression Model

lrm(formula = Lexeme.2 ~ Topic.4 + Axiology, data = data, x = T,
y = T)

Model Likelihood      Discrimination      Rank Discrim.
Ratio Test      Indexes      Indexes
Obs          376      LR chi2      67.14      R2
Feminine & Womanly 185      d.f.          6      g      0.218      C      0.729
Girly & Girlish    191      Pr(> chi2) <0.0001      gr      1.075      Dxy      0.457
max |deriv|      4e-06      gp      2.929      gamma      0.520
Brier      0.208      0.230      tau-a      0.229

Coef      S.E.      Wald Z Pr(>|Z|)
Intercept      0.7749  0.4223  1.83  0.0665
Topic.4=History -2.1138  0.6113 -3.46  0.0005
Topic.4=Japanese culture  1.0247  0.3929  2.61  0.0091
Topic.4=Movies & TV    0.5512  0.3642  1.51  0.1302
Topic.4=Socio-Cultural -0.6043  0.3210 -1.88  0.0598
Axiology=Neutral     -0.6197  0.3474 -1.78  0.0744
Axiology=Positive   -1.5344  0.4667 -3.29  0.0010

```

*Figure 4.* Logistic regression with coarse lexemes as response variable with topic and axiology as variables

As was mentioned in regards to *figure 3* it can be seen that lexemes *feminine* and *womanly* correlate to each other and *girly* and *girlish* do the same, with *ladylike* not being fully included but inclining towards the same cluster. With this information a binary category of “feminine\_womanly” and “girly\_girlish” was created in order to perform a logistic regression analysis.

The topic History was initially used as the variable on which the coefficients were calculated, however, with “History” being one of the major and very distinct categories, the topic on which to calculate was changed to “Person”, yielding better results. The coefficients with a positive value at the top of the LRM shows that topics Japanese culture as well as “Movies & TV” point towards the use of *girly* or *girlish* whereas articles covering topics “History” and “Socio-cultural” point the other direction, towards the use of *feminine* or *womanly*. However, the

Max Bäckström

probability values of topics “Movies & TV” and “Socio-cultural” are notably high with “Movies & TV” having 0.1301 and “Socio-cultural” being at 0.0597, meaning that they are likely to be random and not consistent. Furthermore, their predictive values are rather close to zero with “Movies & TV” at 0.551 and “Socio-cultural” having a value of -0.604 meaning that they are not very predictable variables.

In contrast, “History” has a probability value of 0.0005 and a predictive value of -2.113; from this, it can be concluded that the results from this analysis have a 99.9995% chance of being consistent if doing it multiple times, moreover, the predictive value of -2.113 shows that the topic of “History” has high predictive value in determining the outcome of which lexeme chosen. In a similar manner, “Japanese culture” has a consistency number of 0.0091 and a predictive value of 1.024, meaning that it has a 99.9909% chance of not being random and a fairly strong relevance in predicting the outcome of the response variable. Finally, a positive axiology is also of significant value with a probability value of 0.0010, i.e. a 99.999% of being consistent, and a predictive number of -1.534 making it a reasonably strong variable predicting towards *feminine* and *womanly*.

In addition to this, the overall predictability is determined in the next part of the analysis. The model gets an  $R^2$  value of 0.218, meaning that it is not a fully predictable model but that it still provides a reasonable chance of predicting the outcome moreover, the C value of 0.729 means that the model will predict correctly in approximately 73% of the cases.

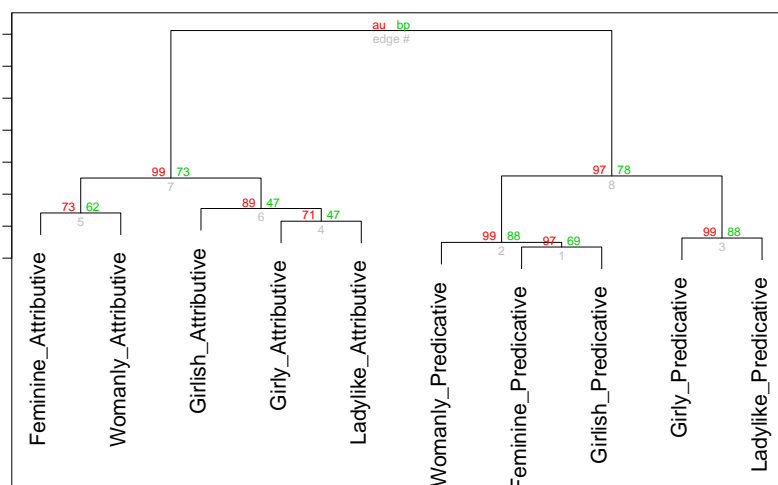


Figure 5. Hierarchical cluster analysis of word class & lexeme, referent noun and referent type

To establish whether or not the word class of the lexeme played a part in determining its usage a category was created by combining the lexeme with the word class. This category was combined with the categories relevant to the referent, i.e. referent noun and referent type and *figure 5* was produced.

This HCA reveals two very distinct clusters of the word class and lexeme combination; interestingly enough, it is solely based on grammar. High percentage values shows that the model is accurate in determining that the choice of lexeme in regards of the referent is based purely on grammatical semantics and that there is no significant lexical influence in this process. Additionally, it can be seen that there are tendencies of *feminine* and *womanly* clustering together and *girlish* and *girly* being connected, similar to the results noted in *figure 3*.



## 5. Summary

In summary, the study yielded results which were highly interesting and helped in answering the research question posed. As was mentioned, very little research on the socio-cultural landscape coming to be on Wikipedia has been done, hence this study is merely scratching the surface of what is to be found, but nevertheless, the results produced were satisfying.

The question posed as to whether or not FEMININITY is conceptualised and represented in different ways in regards of what article it was used in was found to be confirmed; as can be seen in *figure 1* and *figure 2*. From these results, it can be concluded that there are four basic conceptualisations of FEMININITY on Wikipedia when looking at the lexemes in combination with what article they were used in; this is most clearly seen in *figure 2* where four distinct clusters were produced. Most interestingly, lexemes *girly* and *ladylike* correlated and formed a cluster, leaving the similar *girlish* to form a cluster on its own; however, it should be considered that there is quite substantial overlapping the two clusters just mentioned, meaning that they are not absolutely distinct. Moreover, *feminine* and *womanly* form two distinct clusters on their own. Nevertheless, it should be noted that catching the entire concept of FEMININITY by examining a group of lexemes is near impossible, the results produced are clear and substantial, but it should be taken into consideration that this is on a very basic level.

Moreover, *figure 3* shows that when adding the gender of the referent to the data structuring and doing a hierarchical cluster analysis, *girlish* and *girly* do indeed cluster together, just as *feminine* and *womanly*, with *ladylike* as an independent lexeme, loosely correlating to *girlish* and *girly*, similar results to what was found in *figure 1* and *figure 2*. The logistical regression analysis performed on these to clusters proved to be quite predictive, entailing that these two groups of lexemes are distinct from each other and that the analysis offers enough information to distinguish between the two.

Furthermore, it was proven that the grammatical semantics had more significance in determining the lexeme than the actual lexical semantics, as can be seen primarily in *figure 5* and *figure 6*. These plots show that it is not the lexical semantics influencing the choice of lexeme but rather the grammatical, leading to the conclusion that something as subtle as grammatical difference is more important in determining what word to be used in representing FEMININITY. Consequently, this is further proof of

what was mentioned earlier in this chapter, that the examination of these lexemes is not extensive enough to catch the entire concept of FEMININITY.

In conclusion, the study provided satisfying results and proved the hypothesis true; FEMININITY is indeed represented and conceptualised in different ways in different types of articles, meaning that people do indeed represent women in different ways when talking about different things. At a glance, this is understandable; the lexemes differentiate quite significantly and in choosing to use a certain one, people signal their different understandings of FEMININITY, *girly* and *girlish* having a somewhat frivolous undertone and *womanly*, *feminine* and *ladylike* being of a more serious and reserved nature. Furthermore, it was proven that the referent itself does not play a part in the determination of the lexeme and concept, in this case, the variables after which the lexeme is chosen proved to be purely grammatical, separating the attributive from the predicative usages of the adjectives.

However, it should be noted that depicting an entire concept by examining 500 occurrences of a lexeme related to it has its setbacks; factors such as data sparseness and arbitrary analysis may influence the results. Nevertheless, this method is a very effective way of examining natural language usage and seeing as language is an excellent way to socio-cultural information it is a very powerful tool. Finally, there is most certainly potential for an enormous amount of research to be done on Wikipedia and due to the content of it being produced by the audience, with very little influence and control, an impartial and fair content is an indication of moral values and good judgement amongst the millions of people who produce it.

## 6. References

- Brutt-Griffler, Janina  
2002 *World English*. Clevedon, England: Multilingual Matters Press.
- Geeraerts, Dirk, Stefan Grondelaers, and Peter Bakema  
1994 *Structure of Lexical Variation. Meaning, naming, and context*. Berlin: Mouton.
- Glynn, Dylan and Justyna Robinson  
2011 *Polysemy and Synonymy. Corpus methods and applications in Cognitive Semantics*. Amsterdam: John Benjamins
- Lakoff, George  
1987 *Women, Fire, and Dangerous Things. What categories reveal about the mind*. Chicago, University of Chicago Press



*FEMININITY on English Wikipedia*

Seidlhofer, Barbara

2001 Closing a conceptual gap: the case for a description of English as a lingua franca. *International Journal of Applied Linguistics* 11: 133-158.

Wierzbicka, Anna

1985 *Lexicography and Conceptual Analysis*. Ann Arbor: Karoma.