Quantifying polysemy in Cognitive Sociolinguistics

Justyna A. Robinson
University of Sussex

This chapter uses various statistical techniques to explore the extralinguistic grounding of individual conceptualisations of polysemous adjectives in English, such as awesome, gay, wicked. It considers the extent to which individual conceptualisations are non-random and can be related to the socio-demographic characteristics of the speaker. The experimental survey data collected from 72 speakers is analysed via hierarchical agglomerative clustering, decision tree analysis, and logistic regression analysis. The results reveal that not only individual adjectives, as indicated in Robinson (2010a), but whole groups of polysemous adjectives currently undergoing semantic change form usage patterns that can be explained by a very similar sociolinguistic distribution. This study demonstrates that employing a socio-cognitive perspective when researching polysemy is hugely advantageous.

Keywords: adjectives, decision tree analyses, hierarchical agglomerative clustering, logistic regression, semantic variation, semantic change

1. Polysemy

Polysemy, which is usually defined as one form that has several related yet distinct meanings, has been widely explored in various areas of linguistics (for an overview,

---

1. Polysemous meanings are related historically (as opposed to homonymous meanings which are not). From the methodological point of view there are different approaches as to whether a particular sense is a valid reading of a polysemous category. While certain studies include historically-related polysemous meanings in their dataset (e.g. Sweetser 1990), other studies consider only senses that are conceptually related in a given point in time (e.g. Beretta et al. 2005; Klein and Murphy 2002). Some discussion of diachronically-related senses that can be perceived as unrelated is available in Geeraerts (1997) and Blank (2003).

2. Challenges and potential solutions for determining the number and boundaries between polysemous readings are discussed in Dunbar (2001); Geeraerts (1993), Gries (2006), Hanks
see Allan and Robinson 2012; Cuyckens and Zawada 2001; Geeraerts and Cuyckens 2007; Lewandowska-Tomaszczyk 2007; Nerlich et al. 2003; Rakova et al. 2007; Ravin and Leacock 2000; Vanhove 2008). Although polysemy was traditionally associated with the lexicon, much of the research has shown that polysemy also emerges when syntactic, morphological, and phonological usage is considered (see e.g. Brugman 1981; Taylor 1995:142). It has become apparent that polysemy is not just a feature of certain words, but that it is a form of categorisation (see e.g. Taylor 1995:99; Lakoff 1987:12). Therefore, much of the research on polysemy has since focused on learning more about patterns of human categorisation. Some of the key observations of polysemous usage indicate that knowledge is categorised in terms of family resemblance models with less central meanings clustered around a prototype (see e.g. Geeraerts 1989, 1993; Janda 1990; Lakoff 1987:379).

Since the majority of these observations are drawn from the analysis of intralinguistic data only, there is little known about the extent to which the categorisation of specific linguistic events varies between speakers in the same community. For instance, can we assume that the prototypical centre of a polysemous category will be the same for each speaker in a given community? Much research on language variation indicates that linguistic usage does indeed differ between speakers in a community (Chambers et al. 2002; Coulmas 1997; Fought 2004; Labov 2001). These studies demonstrate that the way people speak may be predicted, for example, from their profession, the area in which they live, their social networks (Milroy 1980, 1987), practices they engage in (Eckert 2000) or identities they construct and adopt (Bucholtz 2011). Theoretically, Cognitive Linguistics agrees with the social grounding of linguistic variation by arguing for the experiential and perspectival nature of meaning (Geeraerts 1993:60). However, little research has been done to relate linguistic usage patterns directly to extralinguistic categories and to account for the socio-cultural grounding of categorisation. Notable exceptions are represented in the literature by Geeraerts et al. (1994), Geeraerts et al. (2010), Kristiansen and Dirven (2008), Pütz et al. (2012a), Pütz et al. (2012b), Pütz et al. (2014), Reif et al. (2013), Robinson (2012a), and Robinson (2012b).

2. Scope of the study

The current chapter contributes to the discussion of the extralinguistic grounding of individual conceptualisations of polysemous categories. It considers the extent to which individual conceptualisations are non-random and can be related to the socio-demographic characteristics of the speaker. The current chapter also demonstrates

---

how various statistical techniques (i.e. hierarchical agglomerative clustering, logistic regression, and decision tree analyses) may be employed in order to enhance the way investigations of polysemy are carried out.

The current chapter elaborates on my earlier study (Robinson 2010a) that demonstrated the benefits of implementing a sociolinguistic perspective in cognitive research on polysemy. In that study, I found out that the usage of innovative or conservative senses of the adjective awesome can be predicted from the socio-demographic characteristics of a speaker. For example, the use of awesome 'terrible' can be predicted from the speech of people of 60 years or older. Not only do these findings support cognitive linguistic understanding of polysemy but they also indicate that significant differences exist in the extent to which the different meanings of the same semantic category are salient for different speakers. Although these findings provide compelling evidence for the existence of socio-semantic usage patterns, the conclusions regarding systematic, socially-grounded usage can be only generalised as far as the adjective awesome is concerned. What remains to be verified is whether speakers' usage of other similar words (e.g. other adjectives currently undergoing change) is structured in a similar way.

The current chapter addresses this issue by investigating the usage of eight polysemous adjectives that are presently undergoing change. Firstly, I establish whether any meaningful patterns can be detected when the usage of several polysemous adjectives is considered simultaneously. Provided that meaningful usage patterns emerge, I then determine whether these are non-randomly related to the socio-demographic characteristics of the speakers who use these categories. In order to achieve the aims of the research, various exploratory and confirmatory statistical techniques are implemented. Thus, after introducing the data (Section 3), I summarise the aims of the hierarchical agglomerative cluster analysis (Section 4) and apply this exploratory method to analyse the dataset (Section 5). This analysis is supplemented with confirmatory analyses involving logistic regression (Section 6) and a decision tree analysis (Section 7) before the conclusions are drawn (Section 8).

---

3. This chapter is based on my doctoral research (Robinson 2010a). I would like to thank the University of Sheffield for funding this research project and Joan Beal, Kwa Dąbrowska, Philip Durkin, and Susan Fitzmaurice for generous advice and guidance on many aspects of this research. I would also like to thank Christopher S. Butler, Dagmar Divjak, and anonymous reviewers for comments on the earlier version of the current chapter. All other shortcomings are mine.
3. Data and method

Initial steps in the research follow those presented in Robinson (2010a). Eight polysemous adjectives currently undergoing change (Table 1) and five controlling adjectives have been chosen for the study. Available corpora (the British National Corpus (henceforth, BNC) and the Oxford English Corpus (henceforth, OEC)) and dictionaries (the Oxford English Dictionary (henceforth, OED)) indicate that the investigated adjectives have recently developed a distinctive meaning in British English. For a few of them, a potentially disappearing meaning has also been identified (see Table 1).

In order to determine the usage patterns of these adjectives in a speech community, I carried out interviews with 72 speakers from South Yorkshire, UK. The speaker sample was equally representative of both men and women, different age groups (11–94 years old), and socio-economic backgrounds. Each of the speakers was asked a series of questions aimed at eliciting the most salient usage of polysemous adjectives. These questions followed a schema of asking for a referent that could be best described with an adjective in question, as shown in the following example:

(1) Interviewer: Who or what is wicked?
   Participant: My mum. (referent)
   Interviewer: Why is your mum wicked?
   Participant: Because she lets me play my music as loudly as I want to.
   (justification for use)

Each participant provided, on average, three instances of use of the investigated adjectives, which yielded more than seventeen hundred cases for analysis (excluding counts of those for controlling adjectives).

The information obtained on both the referents and justification for the use of each adjective allowed me to put individual responses into groups of similar usage. Each of those usage groups was then given a sense label. This sense label was mainly

<table>
<thead>
<tr>
<th>Adjective</th>
<th>Incoming meaning</th>
<th>Potentially disappearing meaning</th>
</tr>
</thead>
<tbody>
<tr>
<td>awesome</td>
<td>great</td>
<td>terrible</td>
</tr>
<tr>
<td>chilled</td>
<td>good</td>
<td></td>
</tr>
<tr>
<td>cool</td>
<td>good/trendy</td>
<td></td>
</tr>
<tr>
<td>fit</td>
<td>attractive</td>
<td></td>
</tr>
<tr>
<td>gay</td>
<td>lame</td>
<td>happy</td>
</tr>
<tr>
<td>wicked</td>
<td>good</td>
<td></td>
</tr>
<tr>
<td>solid</td>
<td>hard, tough</td>
<td></td>
</tr>
<tr>
<td>skinny</td>
<td>'fat'</td>
<td>mean</td>
</tr>
</tbody>
</table>
Table 2. Example of the database structure

<table>
<thead>
<tr>
<th>Participant</th>
<th>wicked 'good'</th>
<th>wicked 'evil'</th>
<th>awesome 'great'</th>
<th>awesome 'terrible'</th>
</tr>
</thead>
<tbody>
<tr>
<td>Speaker A</td>
<td>2</td>
<td>1</td>
<td>3</td>
<td>0</td>
</tr>
<tr>
<td>Speaker B</td>
<td>3</td>
<td>0</td>
<td>2</td>
<td>0</td>
</tr>
<tr>
<td>Speaker C</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>Speaker D</td>
<td>0</td>
<td>3</td>
<td>1</td>
<td>3</td>
</tr>
</tbody>
</table>

derived on the basis of the match between the usage and the citation of senses used in the dictionaries. For instance, Example (1) above would be generalised with the meaning *wicked* 'good'. Another sense group that emerged for the adjective *wicked* was *wicked* 'evil'. Occasionally, speakers indicated that they were aware of a certain use of an adjective but they clearly distanced themselves from using this sense. In such cases, a category of a 'reported' sense was introduced. A category labelled 'N/A' was introduced in order to account for overlapping senses that could not be reliably assigned to any of the above groups or for other problematic answers. The raw frequency of use of each sense for each participant was recorded in the database.

The next step in the analysis was to verify if any common usage patterns emerge across all senses used by participants. Usage patterns can be determined by identifying clusters in which two or more senses are used similarly (frequently or infrequently) by a number of speakers. Let us examine Table 2 to illustrate this. Table 2 presents a mini database that is structured in a similar way to the one used in the current study.

One can observe that speakers A and B use *awesome* 'great' and *wicked* 'good' more frequently than other senses and that speakers C and D use *awesome* 'impressive' and *wicked* 'evil' more frequently than other senses of the adjectives. Thus, one may conclude that two distinct usage patterns emerge for two different groups of speakers: one for speakers A and B and another one for speakers C and D. In order to establish usage patterns in a large database, like the one used in the current study, I used the exploratory technique of hierarchical agglomerative cluster analysis (hereafter, HAC).

Once semantic usage patterns are established, I consider the question of why these senses cluster together. Two variants are considered. Since the adjectives used

---

4. This separation of 'reported' senses from 'non-reported' senses is performed for practical reasons of potentially showing where in a community constraints on usage emerge. This is not to suggest that these are two different senses.

5. For instance, when the adjective *gay* used as a female name *Gay*.

6. Before one starts to analyse the visual output of a cluster analysis, it must be stressed that every HAC will always cluster elements together even if the clustering makes no sense. In other words, HAC will impose a structure on any data even if such a structure does not exist (also see Divjak and Fieller, this volume).
in this study are undergoing change, one option is to assess whether senses that cluster together are similar in terms of the historical information we have about them. Do recently developed senses group separately from historically older senses? Another possibility is to consider whether speakers who share the same socio-demographic characteristics (in terms of age, gender, education, etc.) use the same combinations or clusters of senses. The findings of this analysis help to answer the question of whether there are any systematic semantic usage patterns and whether these are socially grounded. In order to find out more information about the users of the senses, I employed two confirmatory techniques: logistic regression and answer tree analysis.

4. Hierarchical agglomerative clustering

There are a number of exploratory techniques used by social scientists that can help to find groups in data, such as principal component analysis, factor analysis, and different types of cluster analysis. These techniques differ in respect to their aims. Principal component analysis and factor analysis are methods for reducing the dimensionality of data (summarising the information in a complete set of variables using fewer variables), whereas cluster analysis aims to organise observations in meaningful structures. I use the exploratory technique of HAC in the current study because I am not interested in data reduction (as the dataset contains a comparatively small number of dimensions) and I intend to investigate groups in the data which are non-randomly similar.

HAC belongs to a family of multivariate exploratory statistical methods (i.e. non-hypothesis-testing) for finding groups in data based on measured characteristics. HAC starts with each case in a separate cluster and then combines the clusters sequentially, reducing the number of clusters in each step until only one cluster is left. This hierarchical clustering process can be represented as a dendrogram, where each step in the clustering process is illustrated by a fork in the tree diagram. A detailed discussion of HAC as well as other types of cluster analysis is presented in Divjak and Fieller (this volume).

The dendrogram in Figure 1 is an example of how senses from Table 1 cluster\(^7\) when their usage evidence for awesome 'great' (meaning 1), awesome 'terrible' (meaning 2), wicked 'good' (meaning 3), and wicked 'evil' (meaning 4) is inputted into the cluster analysis.

The numbers visible on the junctions of the dendrogram represent the measurement of the distance at which clusters fuse together in the hierarchical cluster analysis. This example illustrates meanings 1 and 3 being combined at a fusion value of 2, whereas meanings 2 and 4 are combined at a fusion value of 1. The elements that are clustered earlier (represented by lower fusion values) are more closely related than

---

\(^7\) Distance measure used: phi-square; amalgamation strategy used: Ward.
elements that are clustered later (represented by higher fusion values). The cluster analysis groups senses that are used in a similar way by different people. This dendrogram indicates that different people in the research sample are mostly using the same combination of senses 1 and 3 (1+3) and 2 and 4 (2+4), rather than other combinations such as (1+4), (2+3), or (1+2+3).

5. Hierarchical agglomerative cluster analysis of collected data

Having briefly outlined what constitutes HAC, I move on to discuss details of the computational steps of performing HAC on the current dataset. The analysis was performed using software called ClustanGraphics 7.05 (hereafter, Clustan).

5.1 Selection of polysemous adjectives

Table 3 presents the eight adjectives that are included in the HAC together with relevant meanings (a total of thirty-five meanings). Mifflin and Cooper (1986, cited in Everitt et al. 2001:179) suggest that only variables that are expected to be forming clusters should be included in the analysis. Irrelevant or masking variables should be excluded, if possible. Therefore, controlling adjectives and meanings grouped in the category ‘N/A’ are excluded from the HAC.

The raw frequencies of the use of different senses are recorded in an SPSS database (following the format of the example of Table 2). Certain types of variables need transforming or standardising before the HAC is run (e.g. standardizing to z-scores). However, there is no need to standardise the variables in the current study as their scales do not differ. For more information on standardising variables for cluster analysis, see Divjak and Fieller (this volume).

5.2 Dissimilarity matrix

The next step involves generating a dissimilarity matrix which shows the distances between items. This procedure calculates either the similarities or dissimilarities (also

---

Figure 1. HAC of senses of presented in Table 1
Table 3. Adjectives and their meanings explored in the HAC

<table>
<thead>
<tr>
<th>Adjective</th>
<th>Meanings</th>
</tr>
</thead>
<tbody>
<tr>
<td>awesome</td>
<td>great, terrible, impressive</td>
</tr>
<tr>
<td>chilled</td>
<td>relaxed, cold, collected, relaxed</td>
</tr>
<tr>
<td>cool</td>
<td>good, trendy, calm, collected, cold, good, trendy</td>
</tr>
<tr>
<td>fit</td>
<td>attractive, athletic, healthy, suitable, reported</td>
</tr>
<tr>
<td>gay</td>
<td>lame, unmanly, homosexual, happy</td>
</tr>
<tr>
<td>skinny</td>
<td>latte, thin, showing, skin, mean</td>
</tr>
<tr>
<td>solid</td>
<td>hard, of one substance, hard, of one object, dependable</td>
</tr>
<tr>
<td>wicked</td>
<td>good, evil, reported, good</td>
</tr>
</tbody>
</table>

referred to as distances or proximities), either between pairs of variables or between pairs of cases. Between-variable dissimilarities are generated using SPSS 18 and they are copied into Clustan. Distances can be measured by using different metrics according to different data types (see Divjak and Fieller, this volume). For instance, proximities in count data in the current dataset are measured with phi-square.

5.3 Amalgamation strategy

Once distances between items are calculated, an amalgamation strategy is applied. This procedure involves using one of a few available algorithms (for a summary, see Divjak and Fieller, this volume) that define how separate elements are to be clustered. In the current dataset, the algorithm Increase in Sum of Squares (sometimes called Ward)8 is chosen. This method merges the two elements whose merging least increases their sum of squared deviations from their mean. The Ward algorithm is considered to be a robust method for data classification, which is sensitive to outliers (Gries

8. Other algorithms have also been considered (single linkage, complete linkage, and average linkage). The basic operation of such methods is similar as they fuse individuals or groups of individuals who are closest or most similar. Differences arise because various methods define the distance between individuals or groups of individuals in different ways (see Everitt et al. 2001: Chapter 3). Scholars agree that there is no single best amalgamation method (Everitt et al. 2001: Chapter 8; Mosi 2009; and Tan et al. 2006: 639–642).
2007). This method also appears to have affinity with other linguistic research, e.g., Beitel et al. (2001), Divjak and Gries (2006), Gries and Stefanowitsch (2006).

There is one more step to complete before we can start analysing the dendrogram. The order of the original data can also influence the amalgamation of values. In order to obtain the optimal ordering of the cases, I follow the 'serialize procedure'. This procedure yields the best order of variables that can be obtained from the current proximity matrix.

5.4 Dendrogram

An HAC of the meanings of the adjectives resulted in the dendrogram presented in Figure 2. The vertical layout of the dendrogram in Figure 2 means that we analyse it from left to right. The HAC of the current dataset clustered senses that are used in similar ways by the same people.

For example, the fact that wicked ‘good’ and awesome ‘great’ clustered together means that a number of people who used wicked ‘good’ in the interview were also likely to use awesome ‘great’. Skinny ‘thin’ and awesome ‘great’ are fused later and belong to different clusters. It does not mean that no participant exhibited the use of these two senses. Instead, it just means that people using skinny ‘thin’ were less likely to also use awesome ‘great’ in the same interview.

5.5 Best-cut

After generating a dendrogram, one needs to decide which subclusters are meaningful and therefore should be highlighted for analysis. There are many “rules of thumb” as far as the analysis of cluster levels is concerned, but I will briefly talk about two approaches to this process. First of all, one may delimit borders of individual clusters based on the structure of the dendrogram and perceived (dis)similarities between variables. This procedure involves considering whether various subclusters make sense from the point of the intuitive (dis)similarities between data and the scope of the investigation. Another possibility is to employ statistical measures to determine the best number and size of clusters in the dendrogram (also called best cut). Ideally, conclusions from introspection and statistical analysis should overlap (although this is not always the case). The last scenario is that none of the possible divisions of the dendrogram make sense in the context of a given research question (statistically and through introspection). This is always a possibility since “cluster analysis can create as well as reveal structure” (Breckenridge 2000: 261) (cf. Divjak and Fieller, this volume).

Initial inspection of the dendrogram in Figure 2 indicates that three large clusters could be delimited (see highlighted clusters in Figure 3), the top one being more independent from the remaining two. This three-cluster solution seems sensible in the
Figure 2. Dendrogram of clustered sense
Figure 3. Dendrogram with three and seven-cluster division
light of the current research project. Taking into consideration diachronic information on the usage of individual meanings (cf. Table 1), one can notice that each of the three clusters groups senses that are of different historical depth. Thus, the top cluster includes novel senses, the bottom cluster largely includes senses that are considered to be disappearing, and the middle cluster mostly represents diachronically 'middle' senses (neither recent innovations, nor necessarily disappearing senses). Moreover, the most recent and the oldest senses are grouped into two clusters positioned at the extreme ends of the dendrogram. These visual characteristics indicate that these two clusters are substantially different from each other in terms of usage/people who use them.

The statistical delineation of clusters was carried out by following the best-cut procedure. Best cut in the data was established by using a significance test (upper tail rule, cf. Mojena 1977) on the fusion values at every stage in which the clusters join together in the dendrogram. Best cut indicates the level at which the change in fusion values is significant for most groups. This is then displayed by highlighting partitions on the dendrogram (these clusters are delimited with squares on Figure 3). The partition corresponds to the largest number of clusters, which is significant at the level of 5%.

The best-cut procedure indicates that the seven-cluster solution turns out to be statistically significant. The seven-cluster solution breaks Cluster 2 down into two further clusters and Cluster 3 into four further clusters, whereas Cluster 1 remains unchanged. At first sight, it is less apparent why these sub-clusters would be separated. One potential explanation could involve historical information on the usage of some of these subclusters. Therefore, one could suggest that the cluster (skinny ‘mean’ and awesome ‘terrible’) contains disappearing senses.

At this point, a decision needs to be made as to the level (seven or three clusters) at which to carry out further analysis. This initial exploratory analysis shows that the three-cluster solution already seems to exhibit interesting sense groupings. Going deeper into subclusters (seven clusters) might yield more detailed, but not necessarily as relevant, information (from the point of view of the current research project) on groups of variables. Besides, one rule of thumb says that one should distinguish as few clusters as possible. In the current chapter, I present the analysis at the level of three clusters only.9

---

9. The analysis of clusters at the level that a researcher intuitively considers appropriate (three clusters) may lead to ignoring interesting nuances in use that can be revealed at a more detailed 'best-cut level of seven clusters'. Therefore, the analysis of the seven-cluster solution of the current data is presented in Robinson (2010b).
5.6 Validation of clusters

The validation of a given clustering involves a series of procedures that determine the robustness of a present solution for making predictions. Different studies suggest various ways of assessing the validity of a given clustering (see Duda et al. 2001: 557–559; Everitt et al. 2001: Chapter 8; Moisl and Jones 2005; and Tan et al. 2006: 532–555). In the current study, I examine both the internal stability and the external validity of the present cluster solution following suggestions presented by Clatworthy et al. (2005).

5.6.1 Confirmatory analysis: Internal stability

This first step in the validation procedure aims at answering the question of whether a given cluster solution can be replicated on new data. In order to validate the clustering in Figure 2, I run a tree validation procedure (also called bootstrap validation), which involves a series of random trials on randomised proximities. Each trial generates a different dendrogram for the given data and the series of trials provide a mean dendrogram and confidence intervals. The validation is achieved by comparing the initial clustering with the clustering in the randomised dendrogram. The validation procedure confirms that the obtained dendrogram and the best-cut division of the dendrogram can be replicated even on randomised data.

5.6.2 Confirmatory analysis: External validity (sociolinguistic analysis)

As Clatworthy et al. (2005: 333) point out, the internal stability of clusters is not sufficient evidence to determine the value of a cluster solution. External validity procedures are employed to determine whether a set of external predictors (i.e. variables which were not included in the clustering process) can be associated with the obtained cluster solution and whether the same cluster solution can be replicated from a new independent sample. This approach is considered to be one of the better ways to validate cluster solutions (Aldenderfer and Blashfield 1984: 66).

In Sections 6 and 7, the external validity of the cluster solution is examined. I verify whether any information about the speakers (their age, gender, etc.) who provided the usage data for this study could explain the way in which the senses in Figure 2 are grouped. This practically means employing other statistical measures to validate the use of clusters (linguistic variables) against external variables (socio-demographic variables). In order to perform external validation, multivariate techniques (logistic regression modelling and decision tree analysis) are employed.

From a statistical point of view, verifying the external stability is a necessary element of cluster analysis. From the viewpoint of the current study, this procedure can be considered as a way of testing whether any meaningful semantic usage patterns emerge from analysing sociolinguistic variation. This naturally leads to gaining more insights as to whether conceptualisations of polysemous categories are non-randomly grounded in socio-cognitive contexts.
5.6.3 Summary of cluster-variables
Each of the three main clusters represents a usage pattern which is treated as a linguistic variable. In order to carry out statistical calculations, each of these conceptual patterns-variables needs to be presented as a binary variable to suit further statistical analyses. In order to do this, counts of senses belonging to a given cluster are added up. Following this, by the means of a visual bander10 (SPSS 18), a new variable is created that reflects the characteristics of a whole cluster. This is done through collapsing a large number of ordinal categories into a smaller set of categories, representing low and high usage of a given cluster of senses. Table 4 presents the summary of the dependent variables used, along with associated coding and category information.

Table 4. Linguistic variables (dependent variables) used to investigate their association with socio-demographic variables (independent variables)

<table>
<thead>
<tr>
<th>Dependent variables</th>
<th>Coded as</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>1, 0</td>
<td>High use, Low use</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>1, 0</td>
<td>High use, Low use</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>1, 0</td>
<td>High use, Low use</td>
</tr>
</tbody>
</table>

The following independent variables are considered in the analyses: age group, gender, education, National Statistics Socio-Economic Classification score for a participant's profession11 (hereafter, NSEC), and a postcode or a neighbourhood variable.

Table 5. Socio-demographic variables (independent variables) used to investigate their association with the use of different clusters of senses (dependent variables)

<table>
<thead>
<tr>
<th>Independent variables</th>
<th>Coded as</th>
<th>Categories</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age group</td>
<td>(1, 2, 3, 4)</td>
<td>Up to 18, Male, Female, 19–30, 31–60, Over 60</td>
</tr>
<tr>
<td>Gender</td>
<td>(1, 2)</td>
<td>Higher, Medium, Lower, University, Current student</td>
</tr>
<tr>
<td>NSEC</td>
<td>(1, 2, 3)</td>
<td>Schooling prior the age of 16, Secondary school, College,</td>
</tr>
<tr>
<td>Education</td>
<td>(1, 2, 3, 4, 5)</td>
<td>Lower property prices, Middle property prices, Higher property prices</td>
</tr>
<tr>
<td>Postcode</td>
<td>(1, 2, 3)</td>
<td></td>
</tr>
</tbody>
</table>

10. A visual bander is an SPSS tool to recode values of a variable into groups. Data frequently need to be manipulated before analyses are conducted. Data may need to be recoded, computations may need to be made, new variables may need to be created, or certain records may need to be selected (for more information, see Einspruch 2009).

which is based on property values in areas defined by the postcode of a participant's residence. For a summary of the coding of the independent variables, see Table 5.

6. Logistic regression

Having explored the data via HAC, confirmatory statistical analyses are carried out in order to validate emergent groups in the data. More specifically, I aim to assess the overall effect of socio-demographic categories on the use of particular clusters of meanings, hypothesising that the clustering solution can be explained by the categories of age, gender, NSEC, education and/or postcode value. In order to address the above-mentioned aim, a multifactor statistical model is employed, i.e., a model that considers several external factors simultaneously and measures their effect on the use of each cluster. In addition, the appropriate statistical approach needs to allow us to check for confounding variables. Socio-demographic factors may constitute such cases, i.e., education and occupation may be confounded, as people who are more educated are likely to have better jobs.

Logistic regression analysis fulfils these requirements. Logistic regression can be used to test hypotheses about the relationship of several independent variables to a dichotomous dependent variable (see Hosmer and Lemeshow 1989; Kleinbaum 1994; Speelman (this volume); Tabachnick and Fidell 2001 for introductions to logistic regression). Logistic regression is increasingly being used in linguistic studies (e.g., Benki 1998; Bresnan et al. 2007; Kallel 2007; Levshina et al. (this volume); Trommers et al. 2004).

The logistic regression model also allows for estimating odds ratios for each of the independent variables in the model. For instance, one may establish how many times a given meaning is more likely to be used by age group < up to 18> than by age group <19–30>. Logistic regression also provides information on variance (the percentage to which an independent variable is explained by the dependent ones) and is used to determine the importance of independent variables.

In the current study, logistic regression is performed using SPSS 18 to assess the overall effect of socio-demographic factors (independent variables) on the use of clusters. All responses (including missing values) for seventy-two participants are included in the analysis. All sociolinguistic factors are entered into the model, and then the factors are examined to verify whether they meet removal criteria using a forward stepwise method. The final model is established once no further variables are eligible for removal. The final model is then reported. The resultant fitted model informs us about significant changes in regression coefficients (expressed as $\beta$) between predictors.

In cases where a stable regressive model could not be established and a final solution could not be found (even by modifying model criteria, such as increasing the
number of iterations: i.e. the series of approximations used by the logistic regression), I obtain insights into investigated variation by using multivariate statistical modelling based on decision trees (for more details, see Section 7).

6.1 Logistic regression of Cluster 1

Logistic regression analysis is performed to verify the hypothesis that the likelihood of high use of Cluster 1 can be modelled from speakers’ age, gender, NSEC, education and/or postcode (neighbourhood). Logistic regression analysis on Cluster 1 yields an unstable solution, so one cannot make predictions regarding the use of this cluster. Nevertheless, interesting findings can be obtained from examining decision trees (see Section 7.1).

6.2 Logistic regression of Cluster 2

Logistic regression analysis is performed to verify the hypothesis that the likelihood of high use of Cluster 2 can be predicted from speakers’ age, gender, NSEC, education and/or postcode (neighbourhood) value.

The summary of the logistic regression analysis of the use of Cluster 2 is presented in Table 6. The final model reported includes variables that best account for the observed variation. Insignificant variables are excluded from the model. Table 6 shows the coefficients of regression Beta (hereafter, B), their standard errors, the Wald chi-square statistics, associated p-values, and odds ratios.\textsuperscript{12} The resultant fitted model indicates which independent variables are included in the final logistic model. It also informs us about significant changes in regression coefficients (B) between predictors. B determines the direction of the relationship between a given predictor and the dependent variable (the use of Cluster 2). If B is positive, the odds for the use of Cluster 2 are increased; when B is negative, then the odds are decreased; B equalling 0 leaves the odds unchanged. Explanations of the indicator variables can be found directly under each regression table.

\textit{Model summary.} According to the model, a high use of Cluster 2 can be modelled from speakers’ age (p = .005) and NSEC (p = .005).

\textit{Age.} The most significant differences of use exist between age groups \textit{<19–30>} and \textit{<up to 18>} (p = .001, B = –5.752), and also between age groups \textit{over 60} and \textit{<31–60>} (p = .009, B = 3.407). This means that ‘middle’ age groups speak more similarly to each other. The analysis of probability measures (hereafter, P) presented in

\textsuperscript{12} For further discussion of stepwise regression, regression coefficient, iterations and the output of the logistic regression analysis in SPSS, see Brace \textit{et al.} (2006) and Norušis (1999).
Table 6. Summary of the logistic regression analysis of Cluster 2

<table>
<thead>
<tr>
<th>Variables</th>
<th>B</th>
<th>S.E.</th>
<th>Wald</th>
<th>df</th>
<th>p</th>
<th>Odds ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>AgeGroup</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AgeGroup(1)a</td>
<td>−5.752</td>
<td>1.709</td>
<td>11.333</td>
<td>1</td>
<td>.001</td>
<td>.003</td>
</tr>
<tr>
<td>AgeGroup(2)b</td>
<td>−.313</td>
<td>.841</td>
<td>.138</td>
<td>1</td>
<td>.710</td>
<td>.732</td>
</tr>
<tr>
<td>AgeGroup(3)c</td>
<td>3.407</td>
<td>1.301</td>
<td>6.855</td>
<td>1</td>
<td>.009</td>
<td>30.179</td>
</tr>
<tr>
<td>NSEC</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NSEC(1)d</td>
<td>3.905</td>
<td>1.467</td>
<td>7.085</td>
<td>1</td>
<td>.008</td>
<td>49.672</td>
</tr>
<tr>
<td>NSEC(2)e</td>
<td>1.010</td>
<td>.929</td>
<td>1.180</td>
<td>1</td>
<td>.277</td>
<td>2.745</td>
</tr>
<tr>
<td>Constant</td>
<td>−.056</td>
<td>.380</td>
<td>.021</td>
<td>1</td>
<td>.883</td>
<td>.946</td>
</tr>
</tbody>
</table>

a: change between the age group <19–30> in relation to the age group <up to 18>
b: change between the age group <31–60> in relation to the age group <19–30>
c: change between the age group <over 60> in relation to the age group <31–60>
d: change between the NSEC group <Higher> in relation to the NSEC group <Medium>
e: change between the NSEC group <Medium> in relation to the NSEC group <Lower>

Table 7. Probability values used for logit of Cluster 2

<table>
<thead>
<tr>
<th>Age group</th>
<th>Age group probability</th>
<th>NSEC group</th>
<th>NSEC probability</th>
</tr>
</thead>
<tbody>
<tr>
<td>Up to 18</td>
<td>2.5%</td>
<td>NSEC 1</td>
<td>94.7%</td>
</tr>
<tr>
<td>19–30</td>
<td>88.9%</td>
<td>NSEC 2</td>
<td>26.5%</td>
</tr>
<tr>
<td>31–60</td>
<td>91.6%</td>
<td>NSEC 3</td>
<td>11.6%</td>
</tr>
<tr>
<td>Over 60</td>
<td>26.6%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 7 indicates that speakers of age groups <19–30> and <31–60> are most likely to be high users of Cluster 2 (P = 88.9% and 91.6% respectively).

NSEC. The significant ‘jump’ in B-coefficients exists between NSEC2 and NSEC1 (p = .008, B = 3.905). Speakers who occupy higher occupations (NSEC1) are most likely to exhibit higher use of Cluster 2 (P = 94.7%), in comparison to speakers of middle and lower occupations (P (NSEC2) = 26.5%, P (NSEC3) = 11.6%).

These findings are graphically presented in the form of logistic function estimate values (logit) in Figures 4 and 5. The bars with positive values (above 0) represent categories of independent variables (age group and NSEC, respectively), the occurrence of which correspond to a probability of the use of Cluster 2. In the figures, the taller the bar above 0, the higher the probability of high use of Cluster 2. The bars with negative values (below 0) represent categories of independent variables, the occurrence of which correspond to a probability of the use of Cluster 2. The lower the bar, the higher the probability of ‘not high’ use of Cluster 2.

In the logistic regression analysis, the predictive and explanatory power of the fitted model needs to be assessed. In order to validate predicted probabilities, the
c-statistic is used (see Peng et al. 2002: 6). The c-statistic analyses the proportion of observed to initially-predicted probabilities of occurrences of Cluster 2. In the case of Cluster 2, the fitted model (one that includes socio-demographic variables) achieves a success rate of 80.6%, which is an improvement over the intercept model (51.4%), i.e., a model that does not include any of the socio-demographic variables to account for the observed variation, but includes a constant term only and the model that does not take into consideration NSEC (73.6%).

The explanatory power of the calculated model refers to how effectively it fits the actual data for estimating the outcome variable (Moss et al. 2003: 925). This could be assessed by a number of ‘goodness-of-fit’ measures. $-2 \log \text{Likelihood}$ (hereafter, $-2\text{LL}$) indicates the overall fit of the model. It reflects the significance of the unexplained variance in the model. Its lowering values indicate improvement of a model fit (increasing the likelihood of the observed results). Residual measurements (Cox and Snell, Nagelkerke tests) indicate how much variation the model actually explains. Sometimes these measures may yield different results (for further discussion, see...
Field 2005: 239–240). The Hosmer and Lemeshow test is another measure that is considered by some researchers to be a more accurate measure for assessing the goodness-of-fit of the model (Peng et al. 2002: 6). It tells you how closely the observed and predicted probabilities match and insignificant results in the Hosmer and Lemeshow test signify a model that fits the data well.

In the case of Cluster 2, -2LL (51.740) and an insignificant Hosmer–Lemeshow test indicate that the model fits the data well and is more adequate for explaining variation than models that do not consider socio-demographic factors. R-square measurements (Cox and Snell = .472, Nagelkerke = .630) indicate that the variation in the outcome variable is explained moderately well by the logistic regression model.

Logistic regression analysis evidences that the use of Cluster 2 can be satisfactorily modelled from the age and NSEC of speakers, although age group has a more significant overall effect on the use of the given variable than NSEC. Logistic regression analysis confirms our hypothesis and validates the external stability of Cluster 2.

6.3 Logistic regression analysis of Cluster 3

Logistic regression analysis is performed to verify the hypothesis that the likelihood of high use of Cluster 3 is modelled from speakers’ age, gender, NSEC, education and/or postcode (neighbourhood). Logistic regression on this cluster yields an unstable solution, so predictions regarding the use of this cluster cannot be made. Nevertheless, interesting findings can be observed from examining decision trees (see Section 7.3).

7. Decision tree analysis

In cases where the logistic regression model cannot be established, I use the results of another multivariate technique.

A decision tree analysis is a technique based on separating cases into segments that are as different from each other as possible. For instance, with a decision tree analysis one can easily detect segments and patterns such as ‘female bridge players with at least 5 years’ experience are likely to win a game’, or ‘students who miss more than 40 days of school a year are twice as likely to drop out’.

This procedure uses appropriate algorithms that predict the class (belonging) of a dependent variable from the values of predictor variables. The choice of algorithms largely depends on the type of data. The most appropriate algorithm chosen for our analysis is *Chi-square Automatic Interaction Detection* (hereafter, CHAID).\(^\text{13}\) This is

---

13. Other algorithms have also been considered: C4.5RT, QUEST (for a summary, see SPSS White Paper, Answer Tree Algorithm Summary, 2005), All p-values in the CHAID algorithm
a non-parametric stepwise regression procedure that produces splits until it gets a significant p-value for each split. The CHAID algorithm (available via Answer Tree 3.0) is used to examine factors predicting the use of senses in a cluster. It supplements logistic regression analysis, especially in cases where a logistic regression model cannot be determined. However, it does not mean that both techniques are different ways of answering the same questions. The CHAID algorithm identifies groups of speakers that use similar meanings in a similar way. Logistic regression estimates an overall effect of an independent variable (i.e. age, gender, or social class of speakers) on the use of a particular meaning cluster. Therefore, the two methods are two different ways of looking at the same data.

Decision trees have been widely used in database marketing research (Chaturvedi and Green 1995:245; Magidson 1994; and Rao and Steckel 1995) and in clinical science (Barrio et al. 2006:595; Boscaino et al. 2003:303; and Saltini et al. 2004:737) for performing classification or segmentation. However, the use of decision trees in linguistics is rare (but cf. Heylen 2005; Robinson 2012a; Schmid 2010).

7.1 Decision tree of Cluster 1

I run the decision tree analysis in order to verify the importance of socio-demographic factors (summarised in Table 5) in predicting high use of Cluster 1. More specifically, this analysis shows whether there are any significant socio-demographic groups (e.g. age) or subgroups (age by gender) that use the senses in Cluster 1. The output of the analysis is presented in the form of a decision tree. The decision tree presenting a multivariate analysis of Cluster 1 is presented in Figure 6.

The output presents several levels of significant splits (here two levels). Each split is based on the rule of the lowest p-value. In other words, if two splits are significant, the actual split in the decision tree follows the split according to the independent variable for which the p-value is the lowest.

In the case of a tie, the rule with the higher chi-square value is listed first. In the case of another tie, the rule with lower degrees of freedom (hereafter, df) is listed first. Both chi-square and df values are displayed in the tree for each split. In the decision tree in Figure 6, the square at the top (Node 0) represents the characteristics of a variable to be analysed. There are two categories in this variable: low uses of Cluster 1 and high uses of Cluster 1. There are 39 cases of the former and 33 cases of the latter, accounting for, respectively, 54.17% and 45.83% of the variable. These frequencies are visually presented in the form of bars at the bottom of each square (node). Low use of the cluster is represented by a darker shade of grey, whereas a high use of the cluster is represented by a lighter shade of grey.

---

*analysis were adjusted for multiple comparisons using the Bonferroni method (SPSS Answer Tree 3.0).*
Figure 6. Decision tree of Cluster 1

The first significant split takes into consideration the age group to which speakers belong. The statistics for the significance of this split are described just above the split. The statistics summary indicates that age group is the most significant predictor of using Cluster 1 ($p < .001, \chi^2 = 45.91, df = 3$). Speakers who use this cluster most frequently belong to the two youngest generations; age group <up to 18> exhibits high use in 94.44% of cases and age group <19–30> does so in 61.11% of cases. The results are also presented graphically: light grey bars at the bottom of Nodes 1 and 2 (squares in Figure 6 representing age groups <up to 18> and <19–30>, respectively) indicate the large frequency of the category representing high usage of Cluster 1. Nodes 3 and 4 (representing age groups <31–60> and <over 60>, respectively) indicate the high proportion of speakers who exhibit a low usage of the senses grouped in Cluster 1. This is represented graphically by the light grey bars at the bottom of nodes 3 and 4 (Figure 6).

The second significant split is based on the NSEC of speakers ($p < .0412, \chi^2 = 6.07, df = 1$). The multivariate analysis combined together NSEC2 and NSEC3 (medium

2nd proofs
and lower occupations) and separated them from NSEC1 (higher occupations). All speakers who are <31–60> years old and occupy higher professional positions indicate low usage of Cluster 1 (100% of low usage responses).

The risk estimate for this decision tree is 0.18, which indicates that if I use the decision rule based on the current decision tree I correctly classify 82% (100% minus 18%) of cases (the calculations of risk are not presented on the decision tree).

The multivariate analysis via Answer Tree 3.0 externally validates the use of Cluster 1, showing that the age of participants and, in the case of middle age speakers, their occupation, predicts high use of the senses grouped in Cluster 1.

7.2 Decision tree analysis of Cluster 2

The overall effect of socio-demographic factors in modelling high use of Cluster 2 is established using logistic regression analysis. Decision tree analysis is run in order to verify whether I could obtain any further insights into the use of Cluster 2 in relation to socio-demographic dimensions, especially in the context of determining significant subgroups of use.

Figure 7 illustrates the relative importance of socio-demographic factors in predicting the use of Cluster 2 meanings. The most important factor in predicting usage is the age of participants (p = .0002, \( \chi^2 = 20.8, \text{df} = 2 \)). Speakers of age <19–60> are grouped together as the highest users of Cluster 2. Moreover, multivariate analysis shows that in every age group, speakers living in the most affluent neighbourhoods (Postal code 3, i.e. above £142,795) or holding professional positions (NSEC1) most frequently exhibit high use of the meanings in the cluster (p < .05). Additionally, there is a distinction at the level of gender (p = .01) in the youngest age group of speakers living in the most affluent areas. In this group males are all ‘high users’ of Cluster 2 whereas females are all ‘low users’ of that cluster. Risk estimate indicates that 85% of variation can be correctly classified when applying the decision rule based on the current decision tree.

To conclude, decision tree analysis validates the findings of logistic regression and provides additional evidence for the external stability of Cluster 2.

7.3 Decision tree of Cluster 3

Decision tree analysis is run in order to assess the relative importance of socio-demographic factors in predicting high use of Cluster 3 (see Figure 8).

The multivariate statistical analysis shows that age group is the most significant predictor of using Cluster 3 (p < .001, \( \chi^2 = 44.43, \text{df} = 2 \)). Speakers who use this cluster most frequently belong to the two oldest generations. All of the speakers < over 60> are high users of the cluster. Speakers of age group 31–60> use this cluster in 77.78% of
Figure 7. Decision tree of Cluster 2.
cases (especially when they live in the highest and lowest postcodes ($p = .02, \chi^2 = 7.36, df = 1$). Risk estimate indicates that 87.5% of variation can be correctly classified when applying the decision rule based on the current decision tree.

Decision tree analysis confirms the hypothesis and provides evidence for the external stability of Cluster 3. Overall, the external validity of the cluster solution has been confirmed. The use of each of the three clusters can be predicted from the language of speakers who differ in socio-demographic terms.

8. Summary and discussion of results

Having carried out HAC on the usage data, it has become apparent that each of the main three clusters can be most satisfactorily predicted from the speech of different generations (see summary in Table 8). The use of Cluster 1 (innovative speech) is best predicted from the speech of the youngest speakers, the use of Cluster 3 (historically older senses) is best predicted from the speech of older speakers, and the use of
Cluster 2 (historically neither old nor recent senses) from the speech of middle age groups. Additionally, the results of the statistical analysis of Cluster 2 show that speakers in professional occupations are mostly 'high users' of the senses grouped here.

The IAC reveals that sociolinguistically meaningful semantic usage patterns emerge when usage evidence from several polysemous words is considered. It becomes apparent that the use of a selected group of senses can be most typical for a socio-demographically defined group of speakers. In other words, there are speakers for whom the same senses (e.g. fit 'attractive', gay 'lame', and wicked 'good') are the most salient readings of polysemous categories (fit, gay, and wicked). This finding suggests that not only individual words, such as awesome, but whole groups of polysemous adjectives currently undergoing semantic change form usage patterns that can be explained by a very similar sociolinguistic distribution. This study validates Robinson (2010a) by providing further evidence for the social grounding of polysemous conceptualisations and suggests that employing a socio-cognitive perspective in linguistic research is clearly advantageous. This study also showcases the benefits of engaging various statistical techniques to explore lexical meaning.\(^\text{14}\)

References


\(^{14}\) Although in this chapter I carry out statistical analyses by employing software packages such as ClustanGraphics 7.05, SPSS 18, and Answer Tree 3.0, the same analysis can also be performed with the help of R. The cluster analysis can also be performed with more recent versions of SPSS.

---

**Table 8. Summary of the exploratory and confirmatory analyses of the use of polysemous adjectives**

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Exploratory analysis</th>
<th>Confirmatory analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 1</td>
<td>More recent senses</td>
<td>Younger generations</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>Middle senses</td>
<td>Middle generations especially NSEC1</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>Oldest senses</td>
<td>Older generations</td>
</tr>
</tbody>
</table>


2nd proofs


---

2nd proofs


2nd proofs
Quantifying polysemy in Cognitive Sociolinguistics


