Automated Coding of Qualitative Interviews with Latent Semantic Analysis

Ingo Feinerer
Fridolin Wild
Vienna University of Economics and Business Administration
{h0125130|fridolin.wild}@wu-wien.ac.at

Abstract: Coding and analysing qualitative interviews is one of several core techniques used in marketing research. Qualitative methods offer valuable information hardly gained by standard quantitative methods since open-ended questions and interviews provide deeper insight into customer demands. The main disadvantages of qualitative methods are their inherent subjectivity and their high costs. We tackle this problem by applying *latent semantic analysis* (LSA) in a fully automated way on transcripts of interviews and we propose two algorithms based on LSA. We evaluate the algorithms against two separate real-life cases taken from the automobile industry and from the Austrian mobile phone market. Thereby, we compare the machine results against marketing expert judgements and show that the algorithms proposed provide perfect reliability with appropriate validity in automated coding and textual analysis.

1 Introduction

The proportion of qualitative methods in market and opinion research has risen continuously in the last years. Common qualitative approaches in industry target methods with a high trade-off between quality and quantity, as effective qualitative research is costly. Effective qualitative methods which can be implemented and evaluated within *shorter time frames* tend to predominate: in Germany, for example, the share of depth interviews decreased from 66 % in 1995 to 28 % in 2005, whereas the share of expert interviews rose from 11 % to 53 % (Arb06).

At the same time, empirical qualitative research methods face the problem of *inherent subjectivity* when conducted manually. Typically, few researchers conduct, interpret, and analyse the latent structure in the empirical material collected, with an ever-increasing rate of errors the more interviews they have to process and with subjective differences grounding in their varying interpretations. Choosing representative and trained interviewers can circumvent this only to a limited amount.

Nevertheless qualitative research has several advantages. Closed-ended questions only shift the subjectivity towards the question creation by asking along a battery of predefined items thus preventing spontaneous associations. Qualitative research therefor is and has to be an integral part of empirical research, especially used for

hypothesis building.

In this contribution, we propose a technique based on latent semantic analysis (LSA) that tries to reduce the costs of analysis by providing an automation of analysis substeps and tries to reduce the subjectivity problem inherent to manual analysis.

To be more precise, we investigate, how LSA can be applied to analyse the transcripts of interviews in order to automate the coding of interview transcripts into higher level concepts to be investigated.

With an automated coding instrument at hand, transcripts can be investigated in more depth, as exploratory examination can be performed with significantly lower efforts compared to manual coding.

We therefore will present two algorithms with which we try to extract brand associations automatically from the interview texts which we then will evaluate against results generated by human coding experts.

The rest of this paper is structured as follows. First, we present the methodology applied in this contribution, thereby describing the data-sets used and the overall approach deployed for evaluating reliability and validity of the algorithms we propose. Second, we describe the two algorithms against the background of the latent semantic analysis theory when used to extract brand associations. Third, we report on the results of the automated analysis and discuss the findings in relation to their validity and reliability. In the last section, we conclude with a review on the applicability of the proposed method and identify research problems to be worked on in future investigations.

2 Methodology

In order to evaluate the applicability of the two algorithms we propose that base on LSA, i.e. the *headcount* and the *termcount* algorithm, we deploy to data-sets collected from qualitative interviews that were conducted manually by marketing experts.

One of these data-sets stems from a study conducted in Austria about brands of the Austrian mobile phone market. These interviews were conducted and analysed by hand through Marketmind, an Austrian-based marketing company. The other data-set was originally contributed by (SSOF98) and was conducted in the automobile sector investigating brand relationsships in the US and in Germany.

The mobile phone data-set consists of 969 interviews searching for associations to five mobile phone companies based in Austria. The interview questions aimed at activating brand associations, e.g.,

- *"Which image do you perceive if you consider brand X?" or*
- *"Please imagine brand Z. What do you associate?".*

Each interview consisted of up to 10 answers, mainly short sentences or phrases. Both the questions and answers are in German. The five mobile phone companies are *A1* consisting of 134 interviews, *T-Mobile* consisting of 125 interviews, *One* consisting of 120 interviews, *Tele.ring* consisting of 122 interviews and *Drei* consisting of 468 interviews.

The second dataset we will consider consists of 24 in depth interviews on brand perception of Mercedes in USA and Germany. Each interview is composed of 64 questions to trigger emotions and associations for the brand Mercedes, e.g.,

- *"If I buy a Mercedes I have a good feeling because ..."*
- *"Please characterise a typical Mercedes driver!"*

All answers are available in German. The answers to all 64 questions in one interview have lengths between about 3500 and about 11500 words, most interviews float around a length of 4000 words. The transcripts contain both questions and answers. We stripped the questions and extracted the subset of answers fitting to questions related to the brand Mercedes.

To evaluate the effectiveness of the algorithms we propose in this contribution, we investigate their reliability and validity.

Concerning reliability (irrelevant whether inter-rater-, test-retest- or parallel-forms-reliability) the LSA method has a very high reliability by construction. The LSA procedure is clearly defined in the literature, and is deterministically implemented. This means the LSA methods yields exactly the same results every time for the same set of parameters. This means our approach can help to deal with reliability problems as presented by (RLR$^+$05): they explain how brand attitude data which is typically collected through free choice attitude questions lacks reliability. Of course we cannot control what interviewees respond but we can guarantee that no subjectivity is introduced in the analysis of the transcripts.

The second aspect is validity. We perform external validity tests by checking how similar our automatically produced results are to results aggregated manually by human marketing experts with classical qualitative methods. Since the human scores are not perfectly reliable (you may find an interesting discussion on inter-coder reliability by (LSDB02)) it is clear that we cannot expect perfect external validity when comparing to human scores, as humans among themselves also do not perfectly correlate in their judgements. We expect the algorithms to provide near-human results with high correlations in order to provide evidence to be valid.

3 Algorithms

Applying Latent Semantic Analysis (LSA) for text-mining has a long-standing tradition since its invention in the late eighties of the last century. Dumais provides a comprehensive overview on LSA and its various application settings including

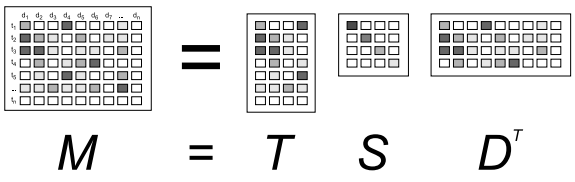

Figure 1: Singular Value Decomposition.

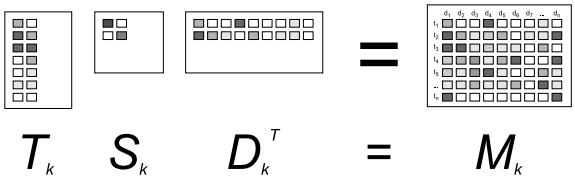

Figure 2: Truncated SVD.

applications for assessing semantic relations in human memory based on text collections (Dum03).

The basic idea behind LSA is that texts contain a semantic structure which, however, is obscured by variations in the wording, and that this structure of meaning can be (partially) unveiled by calculating conceptual indices derived through a truncated singular value decomposition. Comparing both texts and terms on the basis of the resulting lower-dimensional space thereby form the basic working principle of analytic applications using LSA (for a tutorial cf. (WS06)).

The various LSA application processes share a common core: typically first a document-term matrix M (a so-called 'textmatrix') is constructed from a given collection of texts denoting the number of occurrences of a specific term (the rows) for each document (the columns). Using a singular value decomposition (SVD), this matrix is decomposed into three partial matrices (the so-called 'latent-semantic space', see Figure 1).

Thereby T constitutes the left-singular vectors, D the right-singular vectors being both orthonormal and a diagonal matrix S which contains the singular values, so that $M = TSD^T$. By truncating these matrices to a relatively small number of singular values, the document-term matrix M can be restructured in such a way that it reflects only those k common dimensions that account for the greatest share of its underlying variance (cf. (VBRGK06)): $M_k = T_k S_k D_k^T$ (see Figure 2).

In (WSSN05), the authors investigate recommendations on how to identify an ideal number of singular values k to be kept. To keep additional documents from influencing a previously calculated latent-semantic space and to reduce computational costs, new texts can be folded-in after the singular value decomposition (cf. (WS06)): first, a textmatrix V of the same format as M is constructed from the texts. Second, the textmatrix is mapped into the existing factor structure and subsequently reflected back to its textmatrix format: $V' = T_k S_k (V^T T_k S_k^{-1})$, see

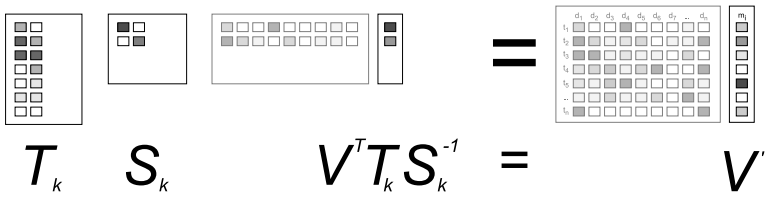

Figure 3: Folding-in of additional texts.

Figure 3.

Based on this general LSA framework, we tested several different methods of how to extract brand associations which mainly differed in how the input corpus is compiled from the available material and how similarities between brand names, constructs, and interview transcripts are measured. The following section will describe the two algorithms that emerged from our experiments in detail.

The first algorithm works on data-sets that contain only responses to one question. In such a data-set, the answers are extracted person by person, each statement in a different file. For the headcount method, the data-set is split into subsets separating the responses by brands and per person: each file contains only the response of one interviewee to one particular brand name (which can be filtered easily from the transcripts if the questions and answers are ordered according to the different brands). One data-set thus only contains the answers provided by the interviewees to one specific brand.

For the second approach, especially for larger texts, the interviews are split into questions and answers, each of them written into a separate text file, additionally separated by interviewees. As the interviews are significantly greater in length compared to the input seed-terms from the textual representatives of the coding scheme, the headcount algorithm cannot be easily applied to such a corpus: The correlation between any document and the pseudo-document of the seed-terms has to be expected to be rather low as the documents cover a greater domain and thus are not as sparse as the pseudo-document of the seed-terms. Therefore, we developed a second algorithm, which we call termcount, that deploys the full-transcripts of the larger interviews and uses the similarity between two pseudo-document vectors created by folding the brand name and the seed-terms of the corresponding construct into the latent-semantic space constructed over the interview transcript documents.

3.1 Algorithm 1: Headcount

A vector space V^n can be defined containing vectors of term frequencies within $\mathbb{R}$ (see Equation 1). The term frequencies f_n can be real numbers as a textvector can also be weighted using common termweighting measures such as inverse document

frequencies.

$$V^n = \{(f_1, f_2, \ldots, f_n) \mid f_1, f_2, \ldots, f_n \in \mathbb{R}\} \tag{1}$$

Within this vector space V^n a multiset M can be declared which contains the document collection, also called corpus (Equation 2). This multiset M can also be represented in matrix format, thus giving a textmatrix as described above.

$$M = \{d \mid d \in V^n\} \tag{2}$$

The multiset M has to satisfy the condition that for every two elements of the set, there is a term label function t which yields the same character-set (Equation 3).

$$\forall v, w \in M : t(v) = t(w) \tag{3}$$

The term label function relates every vector element of the vectors within M with its corresponding term label.

This multiset M is generated by a) constructing a textmatrix over the input corpus which contains the response of exactly one interviewee per document, b) erecting a latent-semantic space over this matrix, and c) converting this latent-semantic space back to a textmatrix representation format to allow for the more convenient application of operations (these steps a–c are described in more detail in Section 3 above). This output textmatrix contains now a representation of the training documents as termvectors again and is identical with the multiset M as defined above in Equation 2.

Stopwords (i.e., highly frequent terms like 'she', 'a' and 'was') are filtered from these texts, the minimum word length is set to two, thus excluding all terms with less than two characters. The documents are read into memory linearly. After they have been read into memory, however, they are converted to a bag of words textmatrix representation, where the order of terms is no longer of importance. The number of first k singular values to be selected for the truncation is calculated as described in (WSSN05).

A coding scheme CS is a set of constructs to be investigated, each construct containing seed-terms representing the construct (Equations 4, 5):

$$\mathrm{CS} = \{c \mid c \in V^n\} \tag{4}$$

such that

$$\forall c, q \in \mathrm{CS} \times M : t(c) = t(q) \tag{5}$$

This c is created by folding the terms from the text representation of the coding scheme into the latent-semantic space, thereby generating a (pseudo-)document vector of the same format as the set M.

The headcount algorithm hc compares every element c of a coding scheme prepared in advance by the analysts with every existing document d in M and maps this comparison to a natural number (Equation 6).

$$\mathrm{hc}\colon M \times c \mapsto \mathbb{N}_0 \tag{6}$$

Again c and d satisfy the condition that they have the same format, as requested in Equation 7.

$$\forall d \in M : t(d) = t(c) \tag{7}$$

This natural number indicates the percent share of how many persons (compared to the overall number of interviewees = documents) have mentioned terms strongly associated to the seed-terms contained in the particular coding scheme element.

$$\text{hc} = \frac{\left|\{d \mid (\text{cor}\,(c,d) > l) \wedge (d \in M)\}\right| \cdot 100}{\left|M\right|} \tag{8}$$

Therefore, simply the cardinality of the set created by filtering those elements from M that correlate with the pseudo-document containing the seed-terms for the coding scheme element has to be divided by the cardinality of M (see Equation 8). For more convenient interpretation by providing natural numbers as output, this fraction was normalised to the number of documents (=number of interviewees) and converted to percentages. The output of hc thus represents the percentage of the share of persons. The threshold used as limit l in our experiments was found out to yield best results when set to $l = .5$, which is also in the range of good correlation of the correlation measure applied. In our case, we used Pearson's product moment correlation coefficient, other measures, however, are valid as well and produce similar results.

To evaluate the validity of the headcount algorithm, we compare the machine extracted association strengths with the previously established human classifications (external validation). See Section 4 for the results.

We call this method 'headcount', as it basically analyses the transcripts on a person-by-person basis.

3.2 Algorithm 2: Termcount

The second algorithm differs in so far as the association strength is measured directly via similarities in the textmatrix. A pseudo-document containing the brand name is compared against a pseudo-document containing the constituting seed-terms of a coding construct. Seed terms thereby are usually very prominent terms able to stipulate associations of the concept under investigation. For example, a coding construct 'security' could be composed of the term set ('secure', 'safe', 'stability'). We call this second method 'termcount', as it directly measures association strengths between the brand name and a seed-vocabulary.

The document vectors c and b represent pseudo-document vectors folded into the existing space, b containing a frequency of 1 at the position of the term of the brand name (e.g., 'Mercedes') and c containing frequencies of 1 at the positions of the terms representing the variable under investigation (e.g., $(\text{`}secure\text{'} = 1) \wedge (\text{`}save\text{'} = 1) \wedge \text{`}stability\text{'}$, see Equation 9).

$$c \in V^n, b \in V^n \tag{9}$$

Again c and b satisfy the condition that they have the same format as all existing documents d, as requested in Equation 10.

$$\forall b, c, d \in M : t(b) = t(c) = t(d) \tag{10}$$

The termcount algorithm uses the folded-in pseudo-documents for the construct constituting seed-terms and the brand name to map every tuple of them to a real number, indicating the association strength (see Equation 11).

$$\mathrm{tc}\colon c \times b \mapsto \mathbb{R} \tag{11}$$

The association strength is defined to be the similarity of the two vectors, measured with Pearson's correlation coefficient (Equation 12). Other similarity measures apply as well and yield similar results.

$$\mathrm{tc} = \mathrm{cor}\,(c, b) \tag{12}$$

We used the 'lsa' package for the (statistical) programming language and computing environment R in it's current version 0.58 (Wil06) for the basic LSA operations. The textual preprocessing was conducted with the 'tm' R package (Fei06) in version 0.1. The two presented algorithms have been separately implemented and were used in the following experiments. The implementations are GPL licenced and can be requested from the authors.

4 Evaluation of the Algorithms: Analysis and Discussion

4.1 Discussion of Mobile Phone Market Results

Table 1 shows the correlations between the scores of our automated latent semantic analysis and the scores by human marketing experts (Mar06) for the Austrian mobile phone market dataset. In detail the table presents several parameter combinations:

Brand: denotes a mobile phone company,

$\mathcal{D}$**:** a fraction of the sum of the selected singular values to the sum of all singular values. I.e., $\mathcal{D}$ is a shorthand for the dimension share,

$\mathcal{T}$**:** gives the threshold which correlations between two terms in the LSA space must succeed to be considered relevant,

ρ**:** the value of Spearman's rho statistic, and

p**-value:** its corresponding p-value.

Brand	$\mathcal{D}$	$\mathcal{T}$	ρ	p-value
A1	0.5	0.6	0.52	0.000
One	0.5	0.6	0.51	0.000
T-Mobile	0.5	0.6	0.40	0.004
Drei	0.5	0.6	0.37	0.008
Tele Ring	0.5	0.6	0.36	0.012
A1	0.5	0.5	0.61	0.000
One	0.5	0.5	0.59	0.000
T-Mobile	0.5	0.5	0.35	0.014
Drei	0.5	0.5	0.46	0.001
Tele Ring	0.5	0.5	0.45	0.001
A1	0.3	0.5	0.55	0.000
One	0.3	0.5	0.52	0.000
T-Mobile	0.3	0.5	0.31	0.030
Drei	0.3	0.5	0.44	0.001
Tele Ring	0.3	0.5	0.37	0.009

Table 1: Correlations between LSA and human scores in the Austrian mobile phone market.

The results of Table 1 indicate that LSA with 'headcount' (cp. section 3.1) is a viable method for analysing open-ended questions automatically. In fact virtually all parameter ranges and combinations (of dimension share $\mathcal{D}$ and threshold $\mathcal{T}$) unveil correlations between the LSA headcount method and the human reference group above 0.3 on a highly significant level.

It is especially notable that the LSA headcount method shows the best results for default parameter settings, i.e., both dimension share and threshold with 0.5. With this settings we obtain a maximal correlation of the LSA scores with human scores of 0.62 at a significance level of $4.087 * 10^{-6}$ for the brand A1. We receive similar results for One with correlation 0.6 at significance $9.188 * 10^{-6}$. Both results' importance is underlined by the fact that typically the correlation between two human groups' scores (i.e., the intercoder reliability) also floats around at 0.7: "The criterion of .70 is often used for exploratory research." (LSDB02, page 593). For the two brands Drei and Tele Ring we calculate correlations at almost 0.5, where T-Mobile's result stays back at 0.35. This lower value can be explained with the small set of open-ended questions for this brand ($n = 125$).

4.2 Discussion of Mercedes Results

Figure 4 shows free associations for the brand Mercedes with human scores and in comparison thereto the machine assigned scores created with the proposed term-count algorithm. The values correlate with Spearman's rho statistic with $\rho = .51$

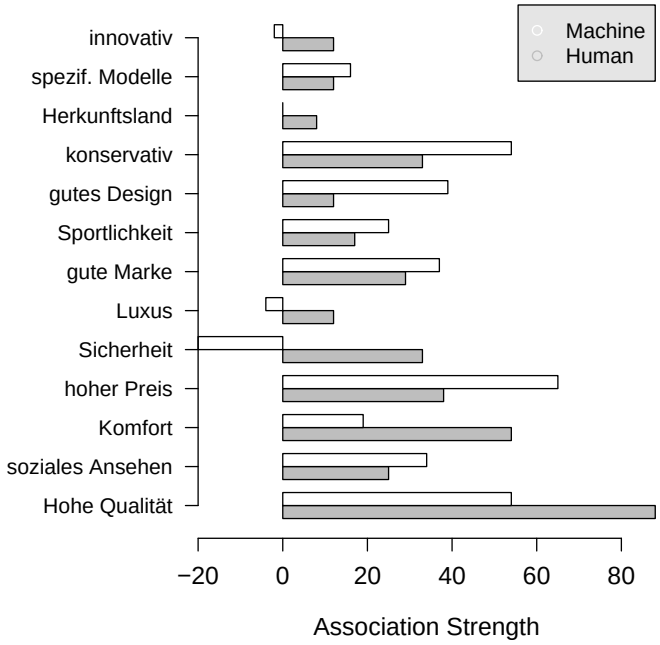

Figure 4: Free Associations for Mercedes scored with termcount in comparison to the humans coding.

with a significance of p-value= .07.

Differently than the human scores which represent the share of people who used terms aggregated in the corresponding coding construct in their responses, termcount is the percentage of the Pearson correlation. Therefore, it can also take negative values. The outlier at 'Sicherheit' (engl. 'security') can be explained as follows: obviously the seed-terms used were not sufficient enough, compared to what the human coders had in mind when they were coding the interviews.

5 Conclusion

We have presented an approach based on latent semantic analysis for investigating and coding qualitative texts which tackles the main problems of qualitative research methods: high costs and inherent subjectivity. Since our method works automatically it can contribute to cost reduction as it speeds up the coding process significantly. Subjectivity is fundamentally reduced as the algorithms are well de-

fined and provide perfect reproducibility of results. In other words the LSA method is highly reliable.

Besides reliability validity is of great interest because we expect the method to deliver near human results. The main consideration is that a high matching between human and machine judgements is sufficient but not neccessary (since humans are error-prone). Instead we can guarantee reliability on a very high level with reasonable validity. Concerning validity it is clear that the coding strategy has to be adapted to the considered methods: in the human coding, labels should be the most concise, in the automated coding the most significant.

In this respect we conclude from the results that the two algorithms suggest that LSA per se can be deployed for conducting stages of the qualitative research process and that they are viable and powerful analysis methods for increasing the quality: LSA offers highest reliability combined with appropriate validity.

We propose that the headcount method is applied to larger corpora containing smaller, statements-sized interviews as in the case for the corpus provided by Marketmind. The termcount method should be applied with smaller corpora containing lengthy, in-depth interviews, as it provides a solution to work with the latent-semantic structure mined from these documents without directly referencing to the underlying documents needed to construct this latent-semantic space.

In our future work, we intend to investigate more closely, how correlations in the modified vector space can be better interpreted to directly reflect human semantic association strengths.

We see at least two possible starting points here: at first relate association pairs to each other to evaluate associations against other brands, against its corresponding antonym, or in comparison to other highly associated terms. The second option we see, is to separate domain-specific frequency distributions from the generic vocabulary use by identifying significant differences and to work with these significant differences.

6 Acknowledgements

We would like to thank Sonja Ehrenberger and Wolfgang Rejzlik (both Marketmind Austria) for providing us with up-to-date industry market data and for giving us feedback on the methodological background of this paper. Furthermore, we would like to thank Kurt Hornik and David Mayer from the Vienna University of Economics and Business Administration and Andreas Strebinger from the York University for their invaluable feedback in various discussions.

References

[Arb06] Arbeitskreis Deutscher Markt- und Sozialforschungsinstitute e.V. Proportion of qualitative interviews by ADM members. December 2006. [Online; accessed 21-December-2006].

[Dum03] Susan Dumais. Latent Semantic Analysis. *Annual Review of Information Science and Technology*, 38:189–230, 2003.

[Fei06] Ingo Feinerer. *tm: Text Mining Package*, 2006. R package version 0.1.

[LSDB02] Matthew Lombard, Jennifer Snyder-Duch, and Cheryl Campanella Bracken. Content Analysis in Mass Communication: Assessment and Reporting of Intercoder Reliability. *Human Communication Research*, 28(4):587–604, October 2002.

[Mar06] Marketmind. Qualitative interviews in the Austrian mobile phone market. Technical report, Marketmind Austria, 2006.

[RLR+05] Cam Rungie, Gilles Laurent, Francesca Dall'Olmo Riley, Donald Morrison, and Tirthankar Roy. Measuring and modeling the (limited) reliability of free choice attitude questions. *International Journal of Research in Marketing*, 22:309–318, 2005.

[SSOF98] Günter Schweiger, Andreas Strebinger, Thomas Otter, and Gereon Friederes. Qualitative Untersuchung der Marke Mercedes-Benz in vier Märkten unter Berücksichtigung von PKW und Nutzfahrzeugen. Technical report, Vienna University of Economics and Business Administration, Advertising and Market Research Institute, February 1998.

[VBRGK06] Jan Van Bruggen, Ellen Rusman, Bas Giesbers, and Rob Koper. Latent Semantic Analysis of Small-Scale Corpora for Positioning in Learning Networks. Technical report, Open University Netherlands, November 2006.

[Wil06] Fridolin Wild. *lsa: Latent Semantic Analysis*, 2006. R package version 0.58.

[WS06] F. Wild and C. Stahl. Investigating Unstructured Texts with Latent Semantic Analysis. In Lenz and Decker, editors, *Advances in Data Analysis*, Berlin, 2006. Springer.

[WSSN05] F. Wild, C. Stahl, G. Stermsek, and G. Neumann. Parameters Driving Effectiveness of Automated Essay Scoring with LSA. In Myles Danson, editor, *Proceedings of the 9th International Computer Assisted Assessment Conference (CAA)*, pages 485–494, Loughborough, UK, July 2005. Professional Development.