Quantification of transcripts from depth interviews, open-ended responses and focus groups
Challenges, accomplishments, new applications and perspectives for market research

Marcus Schmidt  
*Copenhagen Business School*

Statistical software programs have enriched the analysis of text from depth interviews, open-ended responses and focus group transcripts. This paper addresses some of the most important problems involved in quantification of text, and suggests practical solutions. It presents new ways of employing multivariate analysis and data mining for the analysis of marketing-related textual information. Rule-based webs and multiple correspondence analysis may improve the researcher’s insight into a problem and reveal patterns of association inaccessible to traditional qualitative research methods like grounded theory. It is shown that even a relatively simple word count of a focus group uncovers gender-specific differences in the use of words. The paper argues that such dissimilarities can be used for a more efficient targeting of promotional campaigns.

The qualitative–quantitative schism in marketing research

To many qualitative researchers, using quantitative techniques for analysing qualitative data is inappropriate. At the same time, hard-core market research methodologists in most cases are reluctant to work with qualitative data like text excerpts from focus groups, in-depth interviews and responses from open-ended questions. According to the latter type of researchers, data analysis is based on quantitative information like raw numbers, data cells, tables and figures. Number crunchers prefer data that

Received (in revised form): 11 December 2009
are metric scaled, interval scaled or at least categorically scaled with a limited number of categories (i.e. 20) within a variable as an upper limit. 

One could argue, however, that qualitative research can benefit from creative quantitative treatment, and vice versa. Qualitative approaches can help provide understanding of the underlying dynamics and meaning-making associated with constructs, such as how these are enacted and how they evolve. Sometimes qualitative methods help the researcher in comprehending a phenomenon that quantitative methods cannot provide (Bartunek & Seo 2002). In this age of huge batteries of canned AIO statements and other such question collections, one can lose sight of the natural language of the consumer and the perceptions that underlie that language. Therefore, the analysis of open-ended qualitative responses still represents a fruitful area for (quantitative) market research (Green et al. 2000).

Up to now, efforts to quantify text of relevance to marketers have been limited to a small number of publications (see below). Despite this rather modest interest in the topic, we think that quantitative analysis of qualitative data constitutes a promising avenue of future research. To the best of our knowledge, some of the methods we suggest have so far not been used for text analysis in a marketing environment. When applied appropriately, statistical analysis of a marketing-related text in combination with data-mining tools could improve target market segmentation and company profits.

**The challenge of quantifying text excerpts**

The idea of applying quantitative techniques for a better understanding of text is not new. Advanced statistical methods have helped to decipher an ancient script like Linear B (Ventris & Chadwick 1953), to locate early Hittite towns via an MDS analysis of cuneiform tablets (Tobler & Wineburg 1971), to figure out whether Hamilton or Madison was the author of the Federalist papers (Mosteller & Wallace 1963), and to estimate the probability that a newly discovered poem was written by Shakespeare (Thisted & Efron 1987).

Continuous progress in computer hardware architecture and in software algorithms has strongly improved the speed and applicability of data analysis. Nevertheless, it is still a tedious and time-consuming task to quantify qualitative data (Catterall & Maclaran 1998). Text is an extremely complicated construction – much more complicated than numeric data: technically, each word comprises an observation or may be
regarded as a data point. Since a two-hour focus group discussion of eight to ten people contains approximately 5000 words, there will be 5000 data points. While many words appear only once, others appear frequently. Normally, a focus group contains about 1000 unique words. Unique words are characterised by identical spelling. So, even ‘advertisement’ and ‘advertisements’ are treated as two unique words. If we regard text as a statistical measurement task, this corresponds to 5000 observations of a single categorically scaled variable with 1000 different levels!

Present state-of-the-art computer programs for quantitative text analysis still have trouble ‘understanding’ textual information. One of several problems involves homographs – words with the same spelling but with different meanings. For instance, the ‘bank’ of a river is not the same as the ‘bank’ in which we deposit our money. A correct understanding of the kind of ‘bank’ that is meant can be decoded only by looking at the specific context. According to our practical experience, homographs are not that rare in focus groups. In our analysis of two focus groups (see below) there were several cases of homographs (i.e. *play* with Lego toys versus *play* with advertising typefaces; *ocean* of time versus *ocean* cruiser). Another problem concerns the use of synonyms. For instance, ‘children’, ‘kids’, ‘siblings’, ‘daughters’ are spelled differently although they are often used as synonyms with the purpose of making the conversation appear less monotonic. Sentence shifters like ‘not’ and ‘never’ also cause difficulties (Jörgensen 2005). Assume for a moment that homographs, synonyms and sentence shifters can be handled by the researcher through some manual work, and presume that plurals and other variants of a word can be conflated by a single term: before the final analysis can begin, the analyst will also need to discard 95% or more of the text body’s words because they are of no relevance to the problem under scrutiny. Adverbs, adjectives, articles, conjunctions, pronouns and prepositions (e.g. today, many, a, the, and, he, in) are rarely of interest to the researcher, although they appear very often. Likewise, most verbs and nouns will also be of little or no interest.

When working with numerical data we can rely on the three mathematical relations of reflexivity (*x = x*), symmetry (if *x = y* then *y = x*) and transitivity (if *x = y* and *y = z* then *x = z*) (see Tarski 1977, pp. 66–69). Unfortunately, these basic relations do not hold when working with text because of (1) homographs and (2) synonyms:

---

1 Note that 5000 words (and 1000 unique words) is a rough approximation. Focus groups may be considerably shorter or longer. However, we think that 5000 words is a good estimate of a typical focus group.
1. Bank (financial institution) ≠ bank (shore)
2. If daughter → child then child → daughter or child ≠ daughter (but son)
3. If daughter → child and child → son then daughter ≠ son.

Sentence shifters only make things more unpredictable since they need not directly precede the term they are affecting (like plus and minus in the case of numbers). They can be placed elsewhere in the sentence or even in another sentence, although they may reverse the meaning of the environing context.

There are other structural dissimilarities between numerical data and text, but we think that the problems mentioned above are the most important ones that the researcher needs to consider. If a text contains 1000 unique words, the researcher will need to focus on a list of somewhere between 25 and 50 keywords. Normally the list will be subjectively selected by the analyst, but ideally it ought to be chosen in agreement by a group of knowledgeable judges. If the qualitative study dealt with, say, the purchase of organic vegetables, then judges could consist of a consumer researcher, a retail business professional and a nutrition expert.

Software for the analysis of text

Today the researcher can choose between several programs for the analysis of text. While all programs face the same structural problems involved in quantifying text, user interface, algorithms and output differ. Table 1 provides a list of some software programs. In the present paper we will be using CATPAC™ in combination with SPSS programs for multivariate analysis and data mining (Clementine™).

Since it was launched as commercial software about 20 years ago, CATPAC has been used by a number of researchers for different purposes.

Table 1  Selected software for analysis of text

<table>
<thead>
<tr>
<th>Software</th>
<th>Link</th>
</tr>
</thead>
<tbody>
<tr>
<td>Metamorph</td>
<td><a href="http://www.thunderstone.com/texis/site/pages/metamorph.html">www.thunderstone.com/texis/site/pages/metamorph.html</a></td>
</tr>
<tr>
<td>Ethnograph</td>
<td><a href="http://www.qualisresearch.com/">www.qualisresearch.com/</a></td>
</tr>
<tr>
<td>Kwalitan</td>
<td><a href="http://www.kwalitan.net/engels/">www.kwalitan.net/engels/</a></td>
</tr>
<tr>
<td>TEXTPACK</td>
<td><a href="http://www.gesis.org/en/software/textpack/">www.gesis.org/en/software/textpack/</a></td>
</tr>
<tr>
<td>CATPAC</td>
<td><a href="http://www.terraresearch.com">www.terraresearch.com</a></td>
</tr>
<tr>
<td>SAS Text Miner</td>
<td><a href="http://www.sas.com">www.sas.com</a></td>
</tr>
<tr>
<td>IBM SPSS Text Analytics</td>
<td><a href="http://www.spss.com">www.spss.com</a></td>
</tr>
</tbody>
</table>
The program’s algorithm is based on a neural network. It scans the text, counts all different words, establishes a frequency table of words and produces a matrix with weights reflecting the association between words (words that often appear close to each other get high weights, and vice versa). CATPAC has built-in features for conducting cluster analysis and MDS. For specifics, see Woelfel (1993) and Woelfel and Stoyanoff (undated).

Several researchers have used the program for analysing consumers’ responses to open-ended questions: Wassmann (1992) for studying people’s attitudes towards low-emission cars; Peacock and White (1994) for estimating the popularity of radio programmes; Sares (1998) for studying the socio-political viewpoints of graduates; Kikuchi and Ryan (2007) for assessing respondents’ perceptions of tourist attractions. Doerfel and Barnett (1999) employed CATPAC for conducting a semantic network analysis among all papers presented at a conference. In another academic study that the present author has recently been peer reviewing, researchers used CATPAC for comparing the texts of different accounting standards. For reviews of CATPAC, see Chakrapani (1995), Moore et al. (1995) and Belisle (1996).

Quantitative analysis of depth interviews with managers

In Tollin and Jones (2009), the authors report on their depth interviews with a small sample of managers. In the paper the authors provide ten quotes from managers. The quotes are used to illustrate their findings and to put them into perspective (the authors apply grounded theory). To simplify our discussion, we assume that the ten quotes originate from ten different managers: each manager has been asked to come up with a one-paragraph statement describing the role of innovation and marketing in his/her company.

Preparation for data analysis: standardisation of the text body and establishing a list of keywords

A transcript of the ten quotes is displayed in Figure 1a. Figure 1b very much resembles Figure 1a. However, there are some differences. Figure 1a is a Word document, whereas Figure 1b is an ASCII file. Furthermore, there are

---

2 The paper by Tollin and Jones was chosen as an Outstanding Paper Award Winner at the Literati Network Awards for Excellence 2010 by Emeralds Editorial Team.
some minor textual discrepancies. For instance, the sixth word in Figure 1a reads ‘demanded’, whereas in Figure 1b it reads ‘demand’. Figure 1a contains the word ‘demanded’ once [Manager ID1, line 1, word 6] and the word ‘demand’ once [3, 3, 7]. Since the word stem is the same, we select one of them – here ‘demand’ – to represent both cases. Therefore, the word ‘demanded’ in Figure 1a has been transformed to ‘demand’ in Figure 1b. Likewise, occurrences of the following words have been conflated (words in bold indicate selected version):

- brand/brands, product/products, develop/development, focus/focused, result/results, decide/decides, experience/experienced, function/functionally, scenario/scenarios, change/changes/changing, create/creating, document/documentation, technology/technological and organisation/organisation. The words customer, customers, consumer and consumers have all been conflated to consumers.

Apart from these exchanges, the texts of Figure 1a and Figure 1b are identical.³

Notice that, according to mathematical logic, conflation causes confusion because once several unique words have been conflated by a single word and the original record is lost, then the original version of the word cannot be re-established. To surmount this potential problem, it is advisable to employ a systematic approach that maintains a one-to-one correspondence between original and conflated version. Consequently, assuming that the text under scrutiny is much longer than that in Figure 1a, it is strongly recommended that the researcher keeps careful track of every single change between the original text and the conflated version. This can be done by way of a table (‘log file’) that in a row has an ID code of the word (page, line, word number in line) and then lists the original word(s) and the conflated version. While this may seem time consuming, it is indeed a rather quick procedure once the researcher has become used to it. If this systematic discipline is not upheld throughout the data-preparation procedure the researcher risks losing the overview of the process (what exactly was done earlier?) and (s)he may need to start again from the beginning (‘reboot the process’). A formalised and systematic procedure for preparing a text for statistical data analysis is presented in Schmidt (2001).

At first glance, a lot of the words in Figure 1a – like ‘shareholder’, ‘management’, ‘competition’, ‘innovation’ and ‘profitability’ – seem to be good prospects for a keyword list. Unfortunately, they all appear only once

³ Note that, in the following, we use the terms ‘ID’, ‘observation’, ‘respondent’ and ‘manager’ interchangeably. Likewise, we treat ‘variable’, ‘column’ and ‘keyword’ as synonyms.
[1] We experience that we become demanded if we are able to create value in the areas where we act. Because we are a support function, it is not us who decide so we must be able to document our results.

[2] We have a shareholder oriented policy that says that the best protection we can have is a high share value. So we are very focused on this share value. And the consequence of this is that we must continually create value. And value for us is having robust key figures.

[3] The relevant business unit decides that they would like to develop a proposition to a new segment. That is they will identify the range of factors that are relevant for that segment. What does the new segment demand. Do we have to alter our some things in our contact structure, do we change our products because these are more relevant All these things are given to us. How we communicate this to people: should we write this with large text, or not. And how should one reflect over these things: such projects are our responsibility.

[4] I work really hard with finding a strategy to extend the life-cycle for one of our main products. My other task area concerns bringing marketing input to product development. Its a big cultural difference. Earlier, R&D handed over the products to us. Sell this. I want to change this. We have a strong line organization in our company. People are not used to working cross-functionally. Life cycle management is all about this. Its about all sharing the same information.

[5] Product development and distribution are not marketing tasks, but they contribute to creating a brand. The service concepts of our branch offices are not a marketing task. Our task is to secure power, control over the brand in all those situations where our brand is experienced.

[6] Our primary task is to grow brands. To develop our brands. You have to differentiate and you have to satisfy the trade. Industry is changing. The food chains are starting to sell our products and most products are the same. If you are to differentiate it is important to give the consumer some unique presentation and to stay one step ahead of competition.

[7] Our global strategy is all about generating growth. The focus has moved from profitability to growth. Another thing is that the market is changing rapidly. We must prepare ourselves by creating scenarios. Public relations is a new field of competence that we have been building up lately. Its actually a very recent development. The result of our scenario process, which predicted the emergence of a major new field: preventive health care.

[8] Our problem is that we are very constrained. We are not allowed to say that calcium strengthen your bones. Another aspect is that the consumers cant really see differences between alternatives. Thus, for us the core issue is to build up trust. This is not something we do alone. We are very much dependent on journalists and doctors. A large part of my job concerns getting some documentation on our products.

[9] Marketing has previously been very much sales oriented, but now it is becoming more marketing predominated. Traditionally we focused on the compounds, but today we are a lot more focused on the customers, not just the doctors. The consumers. We really believe in consumer input for product development. It is a very big focus for us and I think it is extremely exciting and we are making huge changes in integrating consumers in our marketing and product development.

[10] Our business foundation builds upon continual product development and innovation. Only by being at the forefront of technological development can we meet our customers needs for IT-implementation. We show the customer which way we see trends going. We give them some inspiration to try and transform them and say how they can use the technology in their organisation. And so together with them design some solution. We give them ideas to how they can develop and how they can develop their business from a technological point of view.

Figure 1a Ten quotes appearing in Tollin & Johnes (2009)
[1] We experience that we become demand if we are able to create value in the areas where we act. Because we are a support function, it is not us who decide so we must be able to documentation our results.

[2] We have a shareholder oriented policy that says that the best protection we can have is a high share value. So we are very focus on this share value. And the consequence of this is that we must continually create value. And value for us is having robust key figures.

[3] The relevant business unit decide that they would like to development a proposition to a new segment. That is they will identify the range of factors that are relevant for that segment. What does the new segment demand. Do we have to alter our some things in our contact structure, do we change our product because these are more relevant All these things are given to us. How we communicate this to people: should we write this with large text, or not. And how should one reflect over these things: such projects are our responsibility.

[4] I work really hard with finding a strategy to extend the life-cycle for one of our main product. My other task area concerns bringing marketing input to product development. Its a big cultural difference. Earlier, R&D handed over the product to us. Sell this. I want to change this. We have a strong line organization in our company. People are not used to working cross-function. Life cycle management is all about this. Its about all sharing the same information.

[5] Product development and distribution are not marketing task, but they contribute to create a brands. The service concepts of our branch offices are not a marketing task. Our task is to secure power, control over the brands in all those situations where our brands is experience.

[6] Our primary task is to grow brands. To development our brands. You have to differentiate and you have to satisfy the trade. Industry is change. The food chains are starting to sell our product and most product are the same. If you are to differentiate it is important to give the consumers some unique presentation and to stay one step ahead of competition.

[7] Our global strategy is all about generating growth. The focus has moved from profitability to growth. Another thing is that the market is change rapidly. We must prepare ourselves by create scenarios. Public relations is a new field of competence that we have been building up lately. Its actually a very recent development. The results of our scenario process, which predicted the emergence of a major new field: preventive health care.

[8] Our problem is that we are very constrained. We are not allowed to say that calcium strengthen your bones. Another aspect is that the consumers cant really see difference between alternatives. Thus, for us the core issue is to build up trust. This is not something we do alone. We are very much dependent on journalists and doctors. A large part of my job concerns getting some documentation on our product.

[9] Marketing has previously been very much sales oriented, but now it is becoming more marketing predominated. Traditionally we focus on the compounds, but today we are a lot more focus on the consumers, not just the doctors. The consumers. We really believe in consumers input for product development. It is a very big focus for us and I think it is extremely exciting and we are making huge change in integrating consumers in our marketing and product development.

[10] Our business foundation builds upon continual product development and innovation. Only by being at the forefront of technology development can we meet our consumers needs for IT implementation. We show the consumers which way we see trends going. We give them some inspiration to try and transform them and say how they can use the technology in their organization. And so together with them design some solution. We give them ideas to how they can development and how they can development their business from a technology point of view.

**Figure 1b** Figure 1a after a conflating procedure has been performed
Figure 1c How CATPAC reads (understands) Figure 1b
Quantification of transcripts from depth interviews, open-ended responses and focus groups

in the text. Other words, like ‘the’ and ‘to’, appear frequently, but they lack analytical relevance. For a word to qualify as a keyword it must (1) be of interest to the researcher and (2) occur with ‘some’ frequency. In the case of the ten quotes, we – subjectively – decided that a keyword must appear at least two or three times in the text, even if within the same quotation or paragraph. The text of Figure 1b consists of exactly 681 words. Of these, 281 are unique. Based on our introductory analysis, we have selected 31 keywords for further analysis (Figure 2). The remaining 250 unique words were regarded as unusable and were excluded from further analysis (i.e. ignored by the program). Figure 1c shows exactly how CATPAC reads Figure 1b (only those words appearing at least three times are shown). All keywords are denoted using different typographical indicators for matters of visual identification.

Cluster analysis of keywords

Figure 2 displays the result of a Wards’ cluster analysis of the 31 keywords (using CATPAC). The number preceding a keyword shows the frequency of occurrences.

Only cluster 8 and cluster 7 (basically consisting of two related sub-clusters) represent ‘substantial’ clusters, implying that they cover at least five keywords. One might ask why keywords are categorised into a specific cluster. If we look at Figure 2, we observe that the keywords ‘brands’ and ‘task’ are part of the same cluster. Why is this so? Let us take a look at Figures 1b and 1c simultaneously. As the reader can check, only IDs 5 and 6 use both keywords.

If we begin with ID4, line 4 word 2 (‘change’), and count this and the next six keywords, the list becomes (Figure 1c): 1. change; 2. product; 3. development; 4. marketing; 5. task; 6. create; 7. brands. The seven words (CATPAC default) comprise a so-called scanning window. Note that this is the first time that both ‘task’ and ‘brands’ appear in the same window (reading from the beginning, downwards from left to right). CATPAC’s algorithm reads the text one word at a time. Thus the next scanning

---

4 The latter precondition is a necessity since a mathematical algorithm can only unravel patterns and build relations between keywords as long as they exist more than once across the text body being analysed – just as two points are required for drawing a line and for measuring a distance. Notice that if the text to be analysed had been a 10–20-page focus group, then the minimum number of occurrences ought to be set somewhat higher, maybe five or six.

5 For some of the analyses that follow we will be employing only the 15 keywords that appear three times or more.

6 Note that, in the following, we use the terms ‘ID’, ‘observation’, ‘respondent’ and ‘manager’ interchangeably. Likewise, we treat ‘variable’, ‘column’ and ‘keyword’ as synonyms.
window starts with ‘product’ and stops after ‘marketing’ (and so on). This new scanning window also includes both ‘task’ and ‘brands’. It turns out that a total of 13 subsequent scanning windows incorporate both words (in seven cases five or more of the seven words in the scanning window are either ‘task’ or ‘brands’). Each pairwise occurrence of words in a scanning window increases the association weight linking the two words in the ‘covariance matrix’. Since ‘task’ and ‘brands’ frequently appear in the same scanning window within a relatively short array of keywords, the software’s algorithm puts them into the same cluster.\(^7\) While ‘task’ and ‘brands’ are strongly tied, they do not mix much with other keywords. For

\(^7\) The algorithm still remembers ‘task’ and ‘brands’ after they have left the scanning window, but for each sequence where a keyword is no longer present it loses 10% of its impact (the so-called decay rate is 0.9 – simulating the phenomenon of forgetting in neurological systems).
instance, they are not at all related to words like ‘business’ and ‘technology’ (no common occurrence in any scanning windows) and they are only weakly related to ‘consumers’ and ‘segment’.

An interpretation of a cluster analysis of words is different from and perhaps more difficult to comprehend than an interpretation of a cluster analysis based on metric data. Nevertheless it may help the analyst with unravelling patterns ‘hidden’ in the data. We do not think that it is necessary to treat each cluster as a separate entity. Provided that it facilitates an understanding of results, the analyst may subjectively ‘join’ clusters on a visual/interpretational basis. In the cluster analysis of Figure 2 it appears that two ‘macro-clusters’ make some sense.

1. Target Marketing issues: clusters 3, 6 and 7 (sell, marketing, product, consumers, segment, differentiate, focus, change, development).
2. Tasks/issues related to the company’s production and processes: clusters 2, 4, 5 and 8 (business, organisation, technology, growth, strategy, cycle, scenarios, results, documentation, decide, value, experience, demand).

Since cluster 1 is small and because it does not make sense to us, we abstain from interpreting it.

To sum up, the result of the cluster analysis suggests that managers differentiate between an external marketing focus and an internal management focus.

A rule model analysis of keywords

An alternative way of analysing the interrelation between keywords is to apply a rule model. But, before we employ this approach, we need to restructure the data matrix in such a way that it can be used as input for a rule model (we apply an algorithm included in SPSS’s data-mining application Clementine) (see Table 2 — this table is again based on Figures 1b–1c).

The column headings list the ID number of the ten managers, while the rows show the 15 keywords from Figure 2 that have three or more occurrences. In the first data column (ID1) we note that most cells contain the letter ‘F’ (False). Only in the rows ‘value’ and ‘create’ is there a ‘T’ (True).

Now let us look at Figure 1c. We notice that ID1 uses only two keywords: ‘value’ and ‘create’. So, concerning this manager, it is ‘True’ that
he uses these two keywords, whereas it is ‘False’ that he uses any of the other keywords. The rest of Table 2 is to be understood in exactly the same way. Table 2 satisfies the input structure for a rule-based web application. Figure 3 shows the rule-based web in circle layout (other layout formats available are network layout, directed layout and grid layout). Three of the 15 keywords from Table 2 – ‘Relevant’, ‘Segment’ and ‘Technology’ – do not appear in the web (Figure 3). They have been discarded because they have only one ‘T’ value in their respective rows in Table 2. At least two occurrences of ‘T’ within a row are needed for inclusion in a rule-based web. So, these keywords are excluded here, although all three appeared three times across the text body. The problem is that all occurrences of the respective keyword were by the same individual: ID3 used ‘relevant’ and ‘segment’ three times, while ID10 employed ‘technology’ three times (see Figure 1b). Therefore it is not possible to establish connections from these three keywords to keywords used by other individuals. One could say that

Table 2  Input matrix for a rule-based web

<table>
<thead>
<tr>
<th>Keyword/ID</th>
<th>ID1</th>
<th>ID2</th>
<th>ID3</th>
<th>ID4</th>
<th>ID5</th>
<th>ID6</th>
<th>ID7</th>
<th>ID8</th>
<th>ID9</th>
<th>ID10</th>
<th>Sum T</th>
</tr>
</thead>
<tbody>
<tr>
<td>development</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>7</td>
</tr>
<tr>
<td>product</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>7</td>
</tr>
<tr>
<td>consumers</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>marketing</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>3</td>
</tr>
<tr>
<td>brands</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>change</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>5</td>
<td></td>
</tr>
<tr>
<td>focus</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>task</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>3</td>
<td></td>
</tr>
<tr>
<td>value</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>create</td>
<td>T</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>new</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>business</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>relevant</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>segment</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>technology</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>F</td>
<td>T</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>Sum T</td>
<td>2</td>
<td>3</td>
<td>7</td>
<td>5</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>47</td>
</tr>
</tbody>
</table>

Note: The keyword ‘development’ is coded 7 times as T (true) in the first row, whereas the same word has a frequency of 11 in Figure 1c (see also Figure 2). If the same manager mentions a keyword more than once then it still is coded only as one incident of T. In Figure 2 every separate occurrence is counted. (The same holds for the other keywords.)
these keywords do not have *interpersonal capabilities* or that they are *not network enabled* – in the present context.

Notice that some lines in Figure 3 are thicker than others. The thicker the line, the closer the connection or association. The line connecting ‘create’ and ‘value’ is thin; this is because there are only two cases where both rows have a ‘T’ in the same column of Table 2 (ID1 and ID2). As the only keyword, ‘value’ links to only one other word in the web. We also notice that the line between ‘product’ and ‘development’ is thick. The reason for this is that, in the first two rows, there are six cases where both keywords have a ‘T’ in the same column. We notice a ‘T’ for ID3, ID4, ID5, ID6, ID9 and ID10. The explanation is not that the two words are always used as one term: both keywords appear 11 times but in only five cases are they directly linked.

Figure 3 shows 12 keywords, implying a maximum of 66 pairwise connections. However, only 25 connections are displayed in the web. This is simply because 0 and 1 cases of ‘T’ between two keywords are omitted. For instance, between ‘brands’ and ‘business’ there are zero cases of T in
the same column, and regarding ‘business’ and ‘change’ there is only a single simultaneous occurrence (ID3) (see Table 2).

Figure 3 reveals several interesting findings, as follows.

- The strongest link in the net is the ‘development’–‘product’–‘change’ triangle. We think this makes good sense.
- ‘Consumers’ links to these same three keywords and to no other word.
- ‘Marketing’ also links to the three words, but interestingly not to ‘Consumers’ (they only appear once together in the same column of Table 2 – ID9).
- The only word not linking to ‘development’ is ‘value’.
- ‘Business’ links (weakly) only to ‘product’ and ‘development’, and not to ‘marketing’ and ‘consumers’.
- ‘Brands’ neither links to ‘marketing’ nor to ‘consumers’
- ‘Create’ links (weakly) to ‘development’, ‘focus’ and ‘value’.
- ‘New’ links to ‘development’ and ‘change’.

One might wonder why ‘brands’ does not link well to other keywords, and why ‘consumers’ does not link to ‘marketing’. The reason may be that the ten managers interviewed by Tollin and Jones were CEOs of companies primarily selling to B2B markets (financial services, IT and pharmaceuticals), where branding is generally of less importance than in consumer markets.8

The focus of Tollin and Jones’ (2009) study is on how marketing tasks can become more involved in the innovation process. An analysis of the rule-based web indicates that the term ‘change’ plays a central role. Notice its ability to establish connections to other words. As the only word in the network, it produces a number of connections (seven), which is higher than the number of times it appears in the text (five) 7/5 = 1.4. ‘Task’ and ‘development’ are also good connectors, whereas ‘value’ (0.20) is not.

To sum up, it seems that, with regard to innovation (B2B), managers are primarily concerned with topics that relate to short-term ‘proactive’ activities like initiating change, developing the product and enforcing action, rather than deliberating about ‘long term’ or ‘visionary’ core competences like values, brand equity and consumers. Before drawing conclusions we must recall that the text used for analysis was limited to

---

8 Moreover, three of the eight occurrences of ‘consumers’ were ‘customer(s)’ prior to conflation. A Google search for ‘business to business’ combined with ‘customers’ gives about five times as many hits as when matched with ‘consumers’.
about one textbook page. Its primary purpose has been to illustrate the potential of applying quantitative methods to the analysis of text.

**Multiple correspondence analysis of keywords**

In some ways a correspondence analysis may be regarded as an alternative to a rule-based web. Basically, correspondence analysis is a graphical illustration of a cross tab. The analysis is derived from an input data matrix similar to that in Table 2. First, let us take a look at Table 3. There are two differences between Tables 2 and 3.

1. In Table 3 every ‘F’ of Table 2 has been transformed to ‘0’ and every ‘T’ to an integer (here varying between 1 and 4). The format of Table 2 is dichotomous, whereas Table 3 not only registers if a respondent uses a keyword or not, it also records how many times a respondent uses the keyword.

2. Columns/keywords with only a single occurrence of ‘T’ in Table 2 (‘relevant’, ‘segment’ and ‘technology’) have been omitted from Table 3.

<table>
<thead>
<tr>
<th>Keyword/ID</th>
<th>ID1</th>
<th>ID2</th>
<th>ID3</th>
<th>ID4</th>
<th>ID5</th>
<th>ID6</th>
<th>ID7</th>
<th>ID8</th>
<th>ID9</th>
<th>ID10</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>development</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>2</td>
<td>4</td>
</tr>
<tr>
<td>product</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>2</td>
<td>1</td>
</tr>
<tr>
<td>consumers</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>4</td>
<td>2</td>
</tr>
<tr>
<td>marketing</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>2</td>
<td>1</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>6</td>
</tr>
<tr>
<td>brands</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>3</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>change</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>focus</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>3</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>task</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>3</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>value</td>
<td>1</td>
<td>4</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>5</td>
</tr>
<tr>
<td>create</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>new</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>business</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>0</td>
<td>2</td>
<td>3</td>
</tr>
<tr>
<td>Total</td>
<td>2</td>
<td>6</td>
<td>6</td>
<td>7</td>
<td>11</td>
<td>8</td>
<td>6</td>
<td>2</td>
<td>15</td>
<td>9</td>
<td>72</td>
</tr>
</tbody>
</table>

Note: Shaded columns discarded prior to analysis.
Apart from these two modifications, the two tables are equal. However, before conducting correspondence analysis we need to ignore four keywords from Table 3 that have fewer than three non-zero numbers in their row (‘brands’, ‘value’, ‘new’ and ‘business’), simply because the multiple correspondence analysis program (SPSS Optimal Scaling) refuses to run the analysis as long as a keyword has fewer than three non-zero occurrences within a row. Thus the correspondence analysis will include only the eight non-shaded keywords of Table 3. After this last adjustment, the input data are in a suitable format for correspondence analysis. Figure 4 displays the object space, and in Figure 5 the subject space is provided (both in a two-dimensional space). The object space shows the degree of association between keywords (the closer the points, the bigger the association), whereas the subject space displays the relatedness of respondents regarding the use of keywords.
In Figure 4, coordinates marked with 1 represent absence of the keyword and are therefore not of any interest to us here. Only those points marked with 2 will be analysed. ‘Marketing’, ‘focus’, ‘development’, ‘create’, ‘change’ and ‘product’ are more or less associated with each other, whereas ‘task’ and ‘consumers’ appear to be less linked to these keywords.

The reader may wonder how Figure 4 fits with Figure 3. In Figure 3, ‘task’ links to five and ‘consumers’ to three other keywords. Note, though, that ‘product’ links to ‘consumers’ in Figure 3, and simultaneously the two words are placed relatively close to each other in Figure 4. Moreover, ‘task’ links to ‘development’ in Figure 3, and in Figure 4 ‘development’ next to ‘focus’ is placed closest to ‘task’. Another relevant question is to what degree Figure 4 fits with Figure 2. Recall that Figure 2, the cluster analysis, is based on all available information (dimensions), whereas Figure 3, the rule-based web, uses only dichotomous F/T information. The weighted correspondence analysis has visual properties but uses only information generated from the first two principal components. Moreover, the number of keywords differs: Figure 2 (31); Figure 3 (12); Figure 4 (8). Finally, cluster analysis is run undifferentiated across the whole text, whereas the

Figure 5 Correspondence analysis – subject space
rule-based web and the multiple correspondence analysis separates the text and allows for analysis on the individual response/ID level as well as for comparisons across responses. The latter feature makes a rule-based web and correspondence analysis attractive for the analysis of open-ended responses and depth interviews. The two methods can also help in analysing words from focus groups that have been split across gender, age groups, etc.

However, directly comparing the three output plots of Figures 2–4 and looking for similarities would correspond to comparing pears and apples. Each output has its own properties.

In Figure 5, we notice some similarities between respondents. Managers 3, 4, 5 and 9 are placed close to each other. Why? Refer to Table 3: all four observations have a non-zero value regarding keywords ‘development’ and ‘product’, and three out of the four also have a non-zero value regarding ‘marketing’ and ‘change’. Likewise, managers ID1 and ID2 appear to be similar. This is because they both share a non zero entrée for ‘create’ and for ‘value’, while they both have zeros for the remaining valid keywords (except for ‘focus’ regarding ID 2).

While our empirical study is small and exploratory, we think that the analysis of the subject space has the potential to categorise managers into clusters with different approaches to marketing strategy. However, such an analysis would necessitate (1) longer interviews, with (2) a bigger sample of managers (maybe 30–50), each of up to, say, one page. While the analysis of such a 25–40-page text becomes a somewhat complex task, we know from our earlier work with complete transcripts of focus groups (see below) that it is manageable in practice.

**Post hoc analysis of sentence shifters**

Did sentence shifters play a role in the analyses carried out above? It does not seem that the shifter issue comprised a serious problem in the text under scrutiny. There is only one sentence shifter, ‘not’, a word that appears eight times in the text. Only with regard to ID5 does ‘not’ work as a sentence shifter of the keyword ‘marketing’, which, erroneously, is positively associated with ‘product development’ (the second ‘not’ has no effect). However, since the affected keywords appear many times we decided to do nothing.9

---

9 It is our experience that sentence shifters play a marginal role. In the present text one could solve the problem by introducing a new keyword: ‘notmarketing’.
Example of a quantitative analysis of focus groups

As part of a commercial study, a quantitative analysis was performed on a complete transcript of two focus group discussions. The overall theme of both discussions was the respondents’ attitude towards the annual tourist catalogue published by the Danish national tourist organisation (VisitDenmark). The 100-page colour catalogue was distributed to all of the country’s 2.2 million households, and therefore it was a very expensive marketing campaign. Understandably, the organisation’s management was interested in assessing the effectiveness of the catalogue.

The selection of respondents was based on the following screening criteria:

- Group 1 – parent with youngest child less than 7 years old
- Group 2 – parent with youngest child between 7 and 13 years.

An equal number of males and females (five of each sex) was selected for each group. The two focus groups contained 6000 and 8000 words, respectively, and both texts contained approximately 1200 unique words. Table 4 shows a list with 10 (out of 42) – conflated – keywords used in the final analysis. The table shows how many times each keyword was used across both groups. Moreover, the results are broken down by (1) age of youngest child, (2) gender, and (3) both background characteristics simultaneously. The analysis utilises CATPAC word frequency count.

Now let us look at Table 4. What does it tell us?

1. Families with small children in general, and females in particular, use the terms ‘children’ and ‘amusement park’ much more often than do, say, males with children that have started school.
2. While females frequently use the keywords ‘pictures’ and ‘confusing’, they are not used by a single male. Generally, families with school-age children use the terms ‘holiday’\(^\text{10}\) and ‘advertisement’, ‘read’ and ‘confusing’ more often than do families with one or several small children.
3. Males use terms like ‘catalogue’ and ‘camping’ a lot more than females.

A chi-square test for homogeneity reveals that the differences (1–3) are significant at the 95% level, some of them even at the 99% level.\(^\text{11}\)

\(^{10}\) We use the term holiday but might have used vacation instead (the Danish word is ‘ferie’).

\(^{11}\) We note that a chi-square test cannot be carried out in cells with zero occurrences.
What do these general findings indicate? We try to provide an explanation, as follows.

- **Ad 1:** in particular, females that have small children are ‘obsessed’ with the comfort and well-being of their ‘small wonders’. Whenever their perceptual system encounters holiday-related issues, they start thinking about what serves the child best and what is necessary for its well-being: playgrounds, swimming pool, a zoo with animals, and so on.

- **Ad 2:** according to some research, females are more attracted by pictures than are males (Basow 1980; Johnson et al. 1987). Note also the growing use of the terms ‘catalogue’, holiday’, ‘read’ and ‘confusing’ when moving from females with small children to females with school-age children. Why does this happen? Most probably because mothers begin reading tourist catalogues aloud in front of their literate child, or

---

**Table 4** Use-frequency of selected keywords by focus group participants, broken down by demographic background characteristics

<table>
<thead>
<tr>
<th>Keyword</th>
<th>Total</th>
<th>Age of child (years)</th>
<th>Gender</th>
<th>Males</th>
<th>Females</th>
<th>Age of youngest child</th>
<th>Males</th>
<th>Females</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>&lt;7</td>
<td>7–13</td>
<td>&lt;7</td>
<td>7–13</td>
<td>&lt;7</td>
<td>7–13</td>
<td>&lt;7</td>
</tr>
<tr>
<td>Children</td>
<td>51</td>
<td>37</td>
<td>14</td>
<td>16</td>
<td>0</td>
<td>21</td>
<td>14</td>
<td></td>
</tr>
<tr>
<td>Holiday</td>
<td>55</td>
<td>19</td>
<td>36</td>
<td>30</td>
<td>25</td>
<td>13</td>
<td>17</td>
<td>6</td>
</tr>
<tr>
<td>Advertisement</td>
<td>56</td>
<td>16</td>
<td>40</td>
<td>29</td>
<td>27</td>
<td>7</td>
<td>22</td>
<td>9</td>
</tr>
<tr>
<td>Pictures</td>
<td>12</td>
<td>7</td>
<td>5</td>
<td>0</td>
<td>12</td>
<td>0</td>
<td>7</td>
<td>5</td>
</tr>
<tr>
<td>Catalogue</td>
<td>55</td>
<td>26</td>
<td>29</td>
<td>38</td>
<td>17</td>
<td>20</td>
<td>18</td>
<td>6</td>
</tr>
<tr>
<td>Camping</td>
<td>15</td>
<td>10</td>
<td>5</td>
<td>11</td>
<td>4</td>
<td>7</td>
<td>4</td>
<td>3</td>
</tr>
<tr>
<td>Read</td>
<td>25</td>
<td>8</td>
<td>17</td>
<td>13</td>
<td>12</td>
<td>6</td>
<td>7</td>
<td>2</td>
</tr>
<tr>
<td>Confusing</td>
<td>14</td>
<td>3</td>
<td>11</td>
<td>0</td>
<td>14</td>
<td>0</td>
<td>0</td>
<td>3</td>
</tr>
<tr>
<td>Interest</td>
<td>18</td>
<td>11</td>
<td>7</td>
<td>6</td>
<td>12</td>
<td>2</td>
<td>4</td>
<td>9</td>
</tr>
<tr>
<td>Amusement-park</td>
<td>26</td>
<td>16</td>
<td>10</td>
<td>6</td>
<td>20</td>
<td>5</td>
<td>1</td>
<td>11</td>
</tr>
<tr>
<td>Total</td>
<td>327</td>
<td>153</td>
<td>174</td>
<td>149</td>
<td>178</td>
<td>76</td>
<td>73</td>
<td>77</td>
</tr>
</tbody>
</table>

Note the keywords here include different versions of the word. For instance, ‘children’ covers child, child’s, children, children’s, kid, kids, kid’s, kids’, son, sons, son’s, sons’, daughter, daughters, daughter’s and daughters’ – but the majority of occurrences were ‘children’ and ‘child’.

---

12 Johnson et al. (1987) conducted open-ended interviews with a sample of respondents. The responses were then content analysed. According to the researchers, females tend to process information in the context of social relations (son, mother, family, dad, kid, etc.), whereas males tend to look at situations as distant from social relations. As was the case in our study, differences across gender were statistically significant.
even encourage the child itself to read from the catalogue in front of the mother. She uses the catalogue as a means to generate a discussion with the child, and for involving or engaging it directly in the holiday decision-making process (selection of destination, etc.). Therefore, it is of utmost importance for the positive experience of mother and child that the catalogue is interesting, that it has nice pictures and that it is not confusing. It is reasonable to believe that this pedagogical trend is more prevalent among mothers with children that have started school than among those with younger children (one does not discuss such issues with babies). Families with school-age children seem to perceive the term ‘holiday’ as leisure and not just as ‘How do I arrange a fun time for my children?’

- **Ad 3:** males are more prone to non-figurative information, and perhaps more interested in camping as such.

Table 5 provides a list of words that have been used (at least once, but often several times) by female respondents but not by males, and vice versa. It also summarises a few gender-specific phenomena. We note that females use words that reflect their role as a mother in charge of the household’s practical matters: food, clothing, attractions for the kids, and so on. Also, Table 5 reveals a gender-specific use of words: whereas females employ words dealing with mood, temptation, empathy and structure, males use

<table>
<thead>
<tr>
<th>Table 5</th>
<th>A selection of words used exclusively by one of the sexes</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Females</strong></td>
<td><strong>Males</strong></td>
</tr>
<tr>
<td>Packed lunches, picnic, meals, cook, menu, bedclothes, toilet paper</td>
<td>Railway station, bus stop, airline, airport, charter, gasoline station, kiosk, freeway, road, pull-in, map, bridge, glove compartment, cigarettes, beer, relax</td>
</tr>
<tr>
<td>Horses, shark, sea lion, horses, icebergs, zoo, museum, roller-coaster, playground</td>
<td>Nature, environment, reusable</td>
</tr>
<tr>
<td>Beautiful, happy, in love, cute, nice, negative, spoil, unhappy</td>
<td>Coupon, competition, win, gift</td>
</tr>
<tr>
<td>Frustrating, irritating, difficult, confusing, odd, system, systematic, classify, group, categories, simple, orderly, well planned</td>
<td></td>
</tr>
</tbody>
</table>

- Females used 16 different references to family relations (daughter, cousin, parents, father-in-law, grandmother, etc.) compared to only four used by males
- Females referred to 7 different names of colours (compared to 2 for males)
- Males mentioned 7 different foreign countries (compared to 0 for females)
words focusing on infrastructure (freeway, airport, etc.) and ‘egocentric’ issues (beer, cigarettes).

For more than a century there has been an ongoing debate among anthropologists: do differences in perception cause linguistic categories or does the causal direction work the other way round – that is, from language and culture to perception (Segall et al. 1966, p. 43). It appears that our empirical study renders some support to the presumption that it is the perceptual system and the mental processes of males and females that influence and determine their use of words.

We think that females’ and males’ use of ‘gender specific’ words provides an interesting insight into the cognitive processes of the two sexes. By focusing on such differences, marketers should be able to target advertising messages and thus make promotional campaigns more effective.

**Discussion, marketing implications and future research**

The excerpt from interviews with managers was limited to about one textbook page. Moreover, we assumed that the ten quotes originated from ten different, or ‘independent’, managers. The primary purpose of the present paper – devised as an exploratory study – was to demonstrate some of the potential advantages of carrying out a quantitative analysis of text. We think that such an analysis – when carried out carefully – can be used as valuable inspiration for the development of hypotheses for subsequent quantitative modelling to be tested in a follow-up field survey. For instance, a battery of appropriate items motivated by the exploratory analysis could be analysed either by way of multivariate techniques like a discriminant analysis, or by building and testing a structural equation model.

As noted, there has been impressive progress with regard to the quantification of text during past decades. However, so far there is no reason to believe that text quantification efforts will make traditional qualitative approaches like grounded theory and laddering obsolete. Rather, quantification of text may be used by the researcher as a supplementary analytical tool for obtaining an improved insight into qualitative data. Normally, an experienced focus group analyst will be able to unravel the most important findings. However, he or she may overlook important patterns ‘hidden’ in the data.

For instance, the findings presented in Tables 4 and 5 indicate that promotional messages aimed at female parents ought to differ from

---

13 It should be noted, though, that language and cultural issues were not the primary focus of the present study.
messages targeted at male parents. Advertisements targeting females – i.e. in a female weekly – should stress ‘classical motherhood deeds’ where children are the focal point. The advertisement should be well structured, could use pictures of animals, and should inform about attractions for children. Promotional messages targeting males – say, in an automobile magazine – seem to obtain more attention when focusing on practical issues like infrastructure. The conclusions that we were able to draw from Tables 4 and 5 were not at all considered in the comprehensive report provided by the research agency’s expert in focus group analysis (a veteran business psychologist). Instead, he focused on how effective the catalogue was with respect to displaying the attractions of the different parts of the country.

A critical issue regarding a quantitative analysis of text concerns time. Higgins et al. (1996) report that it took them a minute to code each half line of text when working with the NUD.IST program. Assuming that the analysis concerns a two- to three-hour focus group, then a complete transcript could fill 20 pages, each consisting of 50 lines and corresponding to 1000 lines in total. Taking the pace set by Higgins et al. as the norm, the coding would take 2000 minutes, or 33 hours – almost a whole working week for a trained person – just to do the coding or data preparation! If we estimate the cost of an hour at €100–200, then the added cost is likely to exceed what clients are willing to pay for an extra analysis of a focus group. However, the analyses that are presented in this paper do not necessitate nearly as much time as that for data preparation.

In the case of the manager interviews, it took us six runs (‘iterations’), or less than an hour, before the software (CATPAC) had scanned the whole text and provided an alphabetically sorted list with all unique words. Establishing a list of keywords for analysis, and conflating different versions of the keywords by using the word processor’s search/exchange facility, also consumed a limited amount of time. Presuming that the analyst is familiar with the software (CATPAC, SPSS, etc.), the statistical analyses carried out above are not time consuming either.

The text with the managers’ responses was relatively short (under 700 words); a transcript from a focus group will typically be 15–30 times longer. Fortunately, our work with the topic over the years has indicated potential for substantial experience effects: many practical arrangements can be employed to speed up the process. To sum up, we believe that

---

14 Our version of the program could not scan all of the text in Figure 1a in one run. Newer versions allow the scanning of much more text during a single run.

---
a useful quantitative analysis of a focus group can be completed within 1-2 days by a trained analyst. Significant time could be saved by not transcribing/analysing the whole text, but only the critical/relevant parts of it. A focus group about supermarket shopping behaviour could consist of, say, ten different sub-topics (availability of the store, the quality of the store’s sales flyers, the supply of fresh fruits, etc.). In such a case the researcher together with the client could concentrate on a limited set of sub-topics, maybe three out of ten, thereby making the analysis faster and more focused. The result would be a considerable saving in time and cost without important loss of insight.

The analysis of the focus group of the tourist catalogue was a basic research project and therefore it took a lot of time to complete. But the experience from the empirical analysis resulted in a set of guidelines for systemising and speeding up the process. Moreover, the costs of producing and distributing the tourist catalogue to 2.2 million households were substantial. Looked at in this light, taking a couple of days for the additional analysis of the focus groups might very well justify the extra cost.

This paper shows how the researcher can use word-frequency analysis along with multivariate methods and data mining to analyse text. Future research with regard to the quantification of text could focus on issues like the effectiveness of branding campaigns, of advertising slogans, and of sales promotion material. Analytical methods – not considered in the present paper – that might be applied could involve Bayesian heuristics, CHAID and Latent Class Analysis.

References


Quantification of transcripts from depth interviews, open-ended responses and focus groups


About the author

Marcus Schmidt is an Associate Professor in the Department of Marketing at Copenhagen Business School. Besides teaching and researching he has for many years been a consultant to companies and, in particular, the marketing research agency GfK. He is a frequent speaker at seminars and at PhD courses. Marcus Schmidt is a co-editor of Marketing Research – An International Approach (2007), Product Development and Target Market
Segmentation (2010) and Marketing Casebook 2nd edition (2009). All books were co-authored by S. Hollensen.

His research has appeared in International Journal of Advertising, International Journal of Opinion Research, Journal of Services Research, Qualitative Market Research, Tourism Analysis, Futures and European Advances in Consumer Research, among others. He is a contributor to the Wiley Encyclopedia of Marketing (forthcoming). Marcus Schmidt is a three-time winner of the Best Paper Award in Marketing Research at the AMA’s Summer Educators’ Conference. His research primarily deals with the application of multivariate analyses in marketing and with new ways of collecting and analysing data. His book – in Danish – on inventing direct democracy using new electronic technologies (1993) triggered a wide debate in Denmark.

Address correspondence to: Marcus Schmidt, Copenhagen Business School, Marketing, Solbjerg Plads 3 Frederiksberg, 2000, Denmark
Email: ms.marktg@cbs.dk