TOWARDS MINING BRAND ASSOCIATIONS FROM USER-GENERATED CONTENT (UGC): EVIDENCE FROM LINGUISTIC CHARACTERISTICS

Gloria Volkmann  
*University of Cologne, volkmann@wim.uni-koeln.de*

Marc Egger  
*University of Cologne, egger@wim.uni-koeln.de*

Detlef Schoder  
*University of Cologne, schoder@wim.uni-koeln.de*

Follow this and additional works at: [https://aisel.aisnet.org/ecis2019_rp](https://aisel.aisnet.org/ecis2019_rp)

**Recommended Citation**

[https://aisel.aisnet.org/ecis2019_rp/82](https://aisel.aisnet.org/ecis2019_rp/82)

This material is brought to you by the ECIS 2019 Proceedings at AIS Electronic Library (AISeL). It has been accepted for inclusion in Research Papers by an authorized administrator of AIS Electronic Library (AISeL). For more information, please contact elibrary@aisnet.org.
TOWARDS MINING BRAND ASSOCIATIONS FROM USER-GENERATED CONTENT (UGC): EVIDENCE FROM LINGUISTIC CHARACTERISTICS

Research paper

Volkmann, Gloria, University of Cologne, Cologne, Germany, volkmann@wim.uni-koeln.de
Egger, Marc, University of Cologne, Cologne, Germany, egger@wim.uni-koeln.de
Schoder, Detlef, University of Cologne, Cologne, Germany, schoder@wim.uni-koeln.de

Consumers’ brand associations offer qualitative explanations on a brand’s success or failure and are typically elicited using survey-based instruments. Marketers are interested in time- and cost-efficient, automated brand association elicitation approaches. To enable an automated brand association elicitation, we show that brand associations can be formalized and described by patterns of linguistic part-of-speech sequences that differ from ordinary speech which is required for an automated extraction via text mining. Furthermore, we provide evidence that UGC is an adequate data-source for an automated brand association elicitation. We do that by comparing survey-based and UGC data-sources using linguistic part-of-speech sequence- and n-gram analysis as well as sequential pattern mining. We contribute to exiting research by establishing prerequisites for the construction of novel information systems that use text mining to extract brand associations automatically from UGC.

Keywords: brand associations, user generated content UGC, Part-of-Speech Patterns, Text Mining

1 Introduction

The content and structure of knowledge in consumers’ minds determines their behavior and has direct impact on brand image, position, and success in the marketplace (Keller 1993; Lee and Bradlow 2012). In marketing and cognitive science literature, consumer brand knowledge is typically represented and measured through associations that consumers hold in their memory and assign to the brand and thus determine a brand’s image (Gensler et al. 2015; Henderson et al. 2002; 1998; John et al. 2006; Keller 1993; Krishnan 1996). Thereby, associations can represent general product characteristics, related products, people, places, occasions, topics, product uses, benefits, attitudes, attributes and any other summary evaluations and memories from sources via direct (e.g. usage) or indirect experiences (e.g. word-of-mouth or advertisements) with the brand or its offerings (Henderson et al. 1998; Keller 1993; Krishnan 1996; Lawson 1998; Mitchell 1982). Furthermore, associations may vary according to consumers’ perceived favorability, strength, and uniqueness, which determines a positive or negative brand image and thus consumers’ differential responses to products or brand marketing, especially in high involvement decision settings (Gensler et al. 2015; Keller 1993). Thus, knowledge of consumer associations is important as it offers diagnostic information on a brand’s health, performance, or competitive position, and serves as a foundation for strategy development to position and adjust products, services, and brands, or to implement measures to affect consumer perception (Esch 2014; Keller 1993). Due to this practical relevance for explaining and controlling consumer behavior, the elicitation of consumer associations is of great interest for organizations. It has also received remarkable attention from marketing and market psychology research (Aaker 1992; Aaker and Keller 1990; Esch 2014; Keller 1993, Kaplan & Haenlein, 2010, 2011a, 2011b). Various reactive survey- and interview-based research instruments such as free association, projective techniques, focus groups, laddering, or brand concept maps have been proposed, developed, and utilized for this task (see Farsky (2007)). However, those approaches are faced with several challenges. First, reactive instruments (e.g. interviews, focus groups, surveys) are often criticized in terms of bias (e.g. interviewer bias, response bias like acquiescence bias or extreme responding), as
participant response might differ due to the (artificial) “reactive” research situation. Second, due to increasing market dynamics, the necessity emerges to also increase the frequency of conducting studies on consumer perspectives (i.e. associations) of Brands (Gensler et al. 2015). However, this leads to tremendous elicitation efforts with reactive approaches as participants have to be recruited and interviews or surveys have to be conducted. Third, consumer associations are typically elicited and evaluated manually which leads to cognitive efforts from the research team, which is time consuming and costly. Therefore organizations are increasingly interested in novel cost-efficient approaches (Urban and Hauser 2004).

However, over the past years two developments have taken place which may address those challenges and also expand the established reactive elicitation paradigm (survey- and interview-based research) to passive, non-obtrusive and non-reactive (i.e. automated) research. The first development can be seen in the advent of the web 2.0 that paved the way for social media applications that allow the creation and exchange of User Generated Content (UGC). (Kaplan & Haenlein, 2010). By now, a vast number of consumers voluntarily publish UGC in the Internet, and express their experiences, opinions, feelings and perceptions online on social networks, in fora, blogs or product review channels. Thus, UGC might serve as a novel, relevant, and publicly available data source for the elicitation of consumers’ brand associations. As UGC is preserved digitally it can be accessed automatically, which increases cost-efficiency due to saving costs for participant recruiting and conducting studies. Furthermore, as UGC is readily available, no participants have to be reactively asked for their associations, which shifts the elicitation paradigm to more passive observation and addresses the criticism reactive research instruments are typically faced with (like interviewer bias). The second development that may change the way brand associations are elicited emerges from developments in natural language processing (NLP), text and opinion mining. Over the past years, these research areas have evolved rapidly by proposing effective and efficient algorithms to analyze human language to find out meaningful words, phrases, and relationships among them. The application of those algorithmic approaches bears the potential to reduce, if not completely eliminate, the cognitive and manual effort to elicit consumer associations from unstructured text (e.g. from interview transcripts, free-text answers in surveys, or text-based UGC) and thus tremendously reduce elicitation costs. However, despite the enormous potential of these two developments for both exploiting a novel non-reactive elicitation paradigm and data source (UGC) and also reduce the elicitation efforts (natural language processing and text mining), important challenges emerge. First, when aiming to exploit UGC for research on consumers’ brand associations, it remains unclear whether this data source is adequate in relation to established reactive data sources (e.g. survey or interview data). Do consumers explicate brand associations in the same way in UGC as they would in reactive survey-based research? Therefore, when exploiting UGC for the elicitation of brand associations, it has to be evaluated whether UGC represents an adequate and comparable, if not equivalent, data source. Therefore, concurrent validity (APA 1954) to established (non-reactive) data sources has to be considered, which motivates our first research question RQ1: Can UGC serve as a non-reactive data source for the elicitation of brand associations (with respect to reactive data sources)?

Second, when using unstructured text as a data source, information on consumer brand associations is coded within verbal structures hidden in the textual content. Therefore, algorithms have to differentiate which verbal expression represents a brand association and which doesn’t. This task is challenging both for algorithms and researchers, as both a common understanding on the content-wise and structural characteristics of brand associations is necessary to reliably extract those from verbal speech. While automated information extraction methods need linguistic resources such as linguistic rules and patterns to operate (Béchet et al. 2012), researchers also have to establish a common sense of how consumers express their associations apart from abductive reasoning (which potentially leads to low inter-rater reliabilities). So, no matter how far we proceed in automatically mining and understanding natural language by algorithms or if we perform this task manually, we have to know the linguistic characteristics of consumers’ brand associations, as otherwise we will not know what to search for. Opinion mining researchers have structurally well-defined how product features and opinions about them are constructed, by e.g. using part-of -speech (PoS) - information (e.g. nouns, adjectives etc.) as a basic form of word-sense disambiguation (Liu 2012; Cambria 2013), which is valuable in product controlling or developing product recommendation systems. However, these formalizations might miss important brand associations. Consumers’ perceptions (i.e. associations) on a brand level, range from attributes of the product itself, attributes linked to the purchase and consumption of the product, such as price information, product appearance, usage and user imagery, but also perceived benefits and product experiences, including the feelings, thoughts and attitudes that consumers have towards a brand or other
entities associated with the brand. Hence consumers do use a wide range of language constructs to describe brand associations. So far, no previous research has investigated whether brand associations can also be characterized from a structural perspective to enable their detection and foster the construction of novel automated information systems (like e.g. product features have been characterized by PoS-patterns in opinion mining research). This seems surprising, as defining, formalizing, and operationalizing the structural characteristics of consumer brand associations would establish a common understanding of the anatomy of brand associations, approach aspects of construct validity for automated approaches, and also foster the development of information systems that make use of efficient, state of the art NLP and text mining techniques to extract consumer brand associations from unstructured text (such as UGC, survey or interview data). This leads to our second research question RQ2: How can verbal brand associations be formally described and operationalized with respect to the construction of information systems?

In answering these research questions, we target to two contributions. First, in providing evidence on the appropriateness of UGC for the elicitation of consumer brand associations (RQ1), a novel and relevant data source would be exploited for marketing and branding research. Several contributions have already provided quantitative evidence on the importance of UGC for brands’ market performance (Floyd et al. 2014), sales figures (Gensler et al. 2015; Henderson et al. 1998; 2002; John et al. 2006; Tang et al. 2014; Xiong and Bharadwaj 2014; Godes and Mayzlin 2004; Dellarocas et al. 2007; Berger et al. 2010; Ghose et al. 2012), as well as for deriving market structures (Netzer et al. 2012; Lee and Bradlow 2012), for eliciting consumers preferences for product attributes (Lee and Bradlow 2012; Archak et al. 2011), for predictions of consumer ratings (Büschen and Allenby 2016; Tirunillai and Tellis) and for inducing dimensions of consumer satisfaction with product quality (Tirunillai and Tellis 2014). However, the previously mentioned contributions that use UGC in form of unstructured text and apply text mining techniques to analyze UGC (e.g. Archak et al. 2011; Lee and Bradlow 2012; Tirunillai and Tellis 2014) focus mostly on product feature detection techniques. Validating UGC as an adequate data source for the elicitation of brand associations (that encompass a broader informational scope than product features) would strengthen analysis methods that enable qualitative explanations on causes of market performance (Esch 2014; Fairecloth et al. 2001; Keller 1993; Lilli 1983). Additionally, as UGC not only covers consumer associations from present and also from future customers (Archak et al. 2011; Floyd et al. 2014; Lee and Bradlow 2012), and furthermore consumers who use UGC for opinion forming, the further predictive potential on consumers’ future reactions towards brands is opened up. Second, if consumer brand associations can be comprehensively formalized and operationalized (RQ2), the construction of novel information systems that use natural language processing and text mining to elicit brand associations from unstructured text (emerging from UGC, interview transcripts, or surveys) becomes possible, which introduces automation into marketing research and fosters frequent, reliable, valid and cost-efficient research on consumers’ brand knowledge. The remainder of this article is structured as follows: After the introduction on the relevance of UGC for brand associations and the motivation to structurally describe and formalize them, we give an overview on the theoretical background and related research. We then describe our methodology and operationalize our research questions. These are then addressed in four comprehensive studies on the formalization and description of brand association characteristics within different reactive and non-reactive data sources, and by validating UGC as a meaningful data source for brand association elicitation. We close with a discussion of results, limitations, and future research.

2 Theoretical Background and Related Research

The “association” construct has its theoretical foundation in the human associative memory model, which describes human semantic memory as a set of nodes and edges (Anderson and Bower 1980). Nodes store information and represent associations. According to the spreading activation theory (Collins and Loftus 1975), edges between association nodes describe activation probability patterns and speed. If a certain association node is activated (e.g. from an external stimulus), its activation spreads throughout the network according to the strength of connected edges, and finally describes which information comes into consumers’ minds. This perspective on semantic processing has been widely adopted in marketing research, which operationalizes the association network with regard to brands. Hence, brand association networks determine a brand’s image and represent the most accepted component of brand equity (Aaker 1992; Keller 1993; Biel 1992), which is the primary cause of why consumers react more or less favourably to brand-related stimuli. With this in mind, marketing researchers have provided content-oriented characterizations to better understand the type of brand associations. Brucks (1986), for example, describes eight categories of associations
which include product level aspects, brand level aspects, evaluations, and usage situations, while Krishnan (1996) includes brand attributes, benefits, and experiences. One of the most comprehensive and established conceptualizations of consumer brand associations in marketing literature has been proposed by Keller (1993, 2003). According to him, associations can be distinguished by the level of information that is subsumed in them and can be categorized in three major types of increasing information scope: attributes, benefits, and attitudes (Keller 1993). Attributes are descriptive features that, from a consumer point of view, characterize either a product or the purchase and consumption experience of a product. They are further distinguished in product-related attributes i.e. attributes of physical composition or service requirement and non-product-related attributes i.e. price, packaging, user and usage imagery). As descriptive features, attributes are the most objective type of associations (e.g. a cell phone case in plastic is in plastic for everyone), however the perceptions of these attributes may lead to different perceptions of benefits and attitudes. Benefits are the personal value consumers attach to product attributes and reflect what consumers think a product or brand can do for them. Attitudes on the other hand are defined as the consumers’ overall evaluations (i.e. like, dislike) of a brand and are not explicitly linked to or specified by an experience or benefit of any kind. For example if a consumer says “I like how the fabric of this sports-shirt feels on my skin” the statement refers to an association of the product-related attribute type e.g. the description as sports shirt and an experiential benefit, because the product brings the customer a pleasant wearing experience. In contrast a statement containing an association of the attitude type would be “I like this sports-shirt”. Existing conceptualizations of brand attributes provide a structure to understand brand associations from a content perspective that allows to qualitatively describe and differentiate several brand images. This is of particular practical relevance because the associations that construct the brand image can be drivers for marketing strategies such as e.g. brand “reinforcement” or “revitalization” strategies (Keller 1993, 2003; Roth 1994; Krishnan 1996; Faircloth et al 2001). For example a brand “reinforcement strategy” (Keller 2012, p. 451; Faircloth 2001) strengthens a specific mix of different types of long time consistent, strong, favorable, and unique brand associations to increase brand awareness and loyalty. A “revitalization strategy” (Keller 2012, p. 451; Faircloth 2001) on the other hand identifies a specific mix of different types of new brand associations or refreshes existing brand associations and thereby generates changes in competitive positioning which has not only an effect on targeted customer segments and competitors but also on the perceived quality of the product. However the described conceptualizations do not help towards understanding brand associations from a formal structural perspective, which is necessary in order to design algorithms to extract brand associations from unstructured text, as this requires the knowledge of structural characteristics such as linguistic rules and patterns (Béchet et al. 2012). However, the potential of formalizing the structural characteristics of certain entities for information extraction has been demonstrated in opinion mining and corpus linguistics research (Cambria et al. 2013; Liu 2012; Feldman 2013; Mäntylä et al. 2018). Therefore, opinion mining researchers are primarily concerned with formalizing implicit and explicit opinions towards products (e.g. “lasts long” vs. “long battery life”) (Liu 2010; 2012: ). Therefore Hu and Liu (Hu and Liu 2006; Liu et al. 2005) use class sequential rules to elicit linguistic characteristics of explicit product features, while linguistic characteristics of implicit product features were formalized from co-occurrence frequencies (Hai et al. 2011) or hybrid association rule mining (Wang et al. 2012). Besides opinion mining, researchers from corpus linguistics have also described and formalized linguistic characteristics for understanding poetry, letters, and fiction (Quiniou et al. 2012), differentiating childrens’ language at different ages (Tellier et al. 2014), detecting plagiarism (Alzahrani et al. 2012), or to extract information of judgment, sentiment and appositive qualifying phrases (Béchet et al. 2012). Those contributions demonstrate that the description of structural linguistic characteristics is a necessary prerequisite to understand, detect, and extract certain information entities from unstructured text. However, previous contributions differ from our research as they refer to different informational entities such as fiction, sentiment, or in case of feature-based sentiment analysis mostly product features. The latter are closely related to our research, as product features are also important components of brand associations that make up for a lot of the attribute and benefit associations (Keller 1993; Lawson 1998). However, product features do not capture the wider informational scope of brand associations which also cover attributes, and particularly benefits or general attitudes not only towards the brand but also other persons, places, or other concepts associated with brand that are important factors e.g. for the brand-leveraging process or target market definitions (Keller 1993). Therefore, to the best of our knowledge, our research is the first that investi-
gates brand associations from a perspective of structural linguistic characteristics and is located at the intersection of automated text analysis (i.e. NLP and opinion mining) anchored in foundations from marketing research and targeted towards the construction of novel information systems to elicit brand image.

3 Methodology

The ultimate objectives of the research at hand lies in the provision of foundations (1) to exploit UGC as a novel data source for the elicitation of consumer brand associations, and (2) to foster the automated extraction of brand associations from unstructured natural language to construct novel information systems. From these objectives we derive two research questions: With our RQ1: Can UGC serve as a non-reactive data source for the elicitation of brand associations (with respect to reactive data sources)? We target the validation of UGC as an adequate data source for eliciting consumer brand associations with respect to established research data sources such as surveys and interviews. To construct automated information systems for marketing intelligence (such as reputation management, image analysis, or monitoring), text mining and natural language processing algorithms have to be employed. Thus the structural characteristics (Béchet et al. 2012) of consumer brand associations have to be described and formalized, which leads to our RQ2: How can verbal brand associations be formally described and operationalized with respect to the construction of information systems?

Figure 1 presents our research model, which shows that we approach both research questions from formalizing and describing structural attributes of brand associations by linguistic characteristics (RQ2). Therefore, we conduct two studies, to both elicit and compare linguistic characteristics of brand associations we collected from non-reactive (i.e. UGC, Study 1) and from reactive data sources (i.e. surveys and interviews, Study 2). With this, we aim to provide indications on concurrent validity (RQ1) (APA 1954) by discovering whether brand associations’ linguistic characteristics are consistent across data sources that fundamentally differ in their creation and elicitation paradigm. To rule out distortions, we conduct two additional control studies. In the first control study (Study 3) we conduct our own reactive free-elicitation study (Olson and Muderrisoglu 1979) to control for distortions that might emerge from Study 2 (associations from existing empirical studies). This is necessary, as in established reactive elicitation approaches (Study 2) associations are eventually post-processed by research teams, for example to rule out ambiguous or synonymous associations. However, these aggregations might lead to distortions in the associations’ structural properties as they might linguistically disintegrate the original “raw” associations. In our second control study (Study 4) we then compare linguistic characteristics of brand associations to those of ordinary speech by an NGram analysis and compare n-gram frequencies of ordinary written speech to the combined set of consumer associations from Study 1, 2 and 3. With this, we control whether linguistic characteristics of brand associations are simply dictated by the syntax of language. In this case, the problem of extracting brand associations from unstructured text by automated algorithms would not be decidable, our data source validation would be meaningless, and the construction of information systems to extract brand associations would be impossible.

In the following we describe our methodology to formalize linguistic characteristics of verbal associations collected in Studies 1, 2, and 3. The studies’ data collection approaches are described their respective sections.
Given a set of verbal brand associations in text form, first tokenization is applied. Tokenization is a standard technique in natural language processing to cut text into a set of tokens according to specific rules (Feldman and Sanger 2006). To receive word tokens for each association we use the tokenization rule of whitespace characters. Thus, a single association like “nice appearance” would be tokenized into the word tokens “nice” and “appearance”. Second, linguistic categories (part-of-speech tags like noun, verb, etc.) are assigned to each of association’s words. For this task a so-called PoS-tagger is applied where we rely on the Stanford log-linear PoS-tagger (Toutanova et al. 2003) which is often used in text mining research. For the English language, the standard of how PoS-Tags on words are described is the so-called Penn-Treebank tagset (Marcus et al. 1993) which defines detailed categories of PoS-Tags (e.g. whether an adjective is a superlative or comparative, or a noun being a proper noun in singular or plural). As we are interested in linguistic characteristics from a more abstract perspective, we are interested in the top linguistic categories (e.g. a word being a noun, adjective, or verb but not the tense, or whether the word describes comparatives or superlatives). Thus, we map each PoS-Tag to a more abstract PoS-Tag using the universal PoS tagset proposed by Petrov et al. (Petrov et al. 2012). Here, for example, the different PoS-Tags for nouns (e.g. NN, NNP, NNPS, NNS, etc.) would be mapped to the more general PoS-Tag <NOUN>. Thirdly, after PoS-tagging, we perform PoS-sequencing by creating a secondary representation of each association. This representation solely consists of a linear sequence of PoS-Tags, which creates a structural linguistic representation of the association. For example, the association “design is nice” consisting of the words “design”, “is” and “nice” would be mapped to the secondary representation of the sequence <Noun, Verb, Adjective>. In the following we denote those sequences as PoS-sequences. From these representations of brand associations, we perform quantitative and pattern mining analyses to describe and compare the associations’ linguistics characteristics. Thereby we use three different variables. First, the number of associations a consumer typically retrieves when being confronted with a brand stimulus is of interest. The variable is of importance, as consumers retrieve information in chunks (Mitchell 1982) and therefore report generally more than one association in each memory probe. So, the number of associations retrieved helps to give a hint as to how many associations might be mined by algorithms from single consumer utterances and helps to evaluate whether the number of associations extracted from non-reactive data sources (i.e. UGC) is comparable to the number of associations typically retrieved in reactive research instruments. As the second variable, we investigate the length of an association by the number of words the association is composed of. The length provides initial information on brand associations’ structural properties. While the average length shows whether consumers usually report rather short or long information chunks, the standard deviation of an association’s length helps to evaluate how diverse consumers express their associations. Furthermore, from the transformation of each verbal association to a PoS-sequence, a linear sequence of linguistic categories in form of PoS-tags (e.g. noun, verb etc.) is received and allows the investigation of structural linguistic characteristics. The PoS-sequence thus allows us to measure and describe patterns of linguistic categories consumers typically use when verbally expressing their associations and serves as a necessary formalization fostering the construction of algorithmic approaches to uncover associations within unstructured text. Finally, we use a technique from data mining known as sequential pattern mining to uncover hidden linguistic patterns of PoS-tags found in the associations’ PoS-sequences. The task of sequential pattern mining was initially motivated by (Agrawal and Srikant 1995), who proposed the Apriori algorithm to identify frequent itemsets in transactional data such as market basket analyses to, e.g., uncover purchase patterns of which products are typically purchased together. As frequent itemset mining does not consider the order of items, several extensions – known as sequential pattern mining – were proposed which also considered the sequential order of items. With respect to the task of finding linguistic characteristics of brand associations, those techniques are especially promising to uncover regularities in structural linguistic characteristics, as the sequential order of linguistic categories (i.e. PoS-tags) is preserved which is important for understanding associations. One drawback in sequential pattern mining is that algorithms derive a large number of patterns being included in other patterns. This leads to high efforts for researchers to screen and clean up meaningless patterns. Therefore, so-called frequent closed-sequential patterns (FCSP) were proposed (Yan et al. 2003) to reduce the number of mined sequential patterns without losing information (Fournier-Viger et al. 2017). To investigate hidden structural linguistic properties and patterns of brand associations, we apply the BIDE algorithm (Wang and Han 2004) to discover FCSPs. We must note, that also other algorithms might be applied, as the task of FCSP mining is deterministic and thus every algorithm will retrieve the exact same set of FCSP for a given
threshold (i.e. support) while differing e.g. in runtime. We chose BIDE (Wang and Han 2004), as it is considered being very efficient (Fournier-Viger et al. 2017), and therefore also allows to uncover patterns from large data sets as UGC. In our study, we use discovered FCSPs to compare if the same patterns are found in reactive and non-reactive data sources to provide additional indications on concurrent validity (APA 1954).

4 Study 1: Associations from UGC

In the first study, we use our methodology to elicit and investigate linguistic characteristics of brand associations emerging from UGC. As a representation for UGC we use online user-generated product reviews where consumers describe their perception of products, services, or brands in general. We collected user-generated product reviews from the now-defunct consumer product review website Epinions.com using a Web crawler. On Epinions.com consumers were able to publish reviews on a broad scope of products, services, and brands. In 2014, Epinions.com removed the functionality for consumers to publish reviews while the already-published consumer reviews were still publicly accessible. In total we collected 1.63 Mio. user-generated product reviews which covered a broad scope of 589 categories and 385,224 products, brands, or services within a time span of approximately 15 years between 1999 and 2014. To extract consumer associations from the reviews, we make use of one important and specific characteristic of Epinions.com. On Epinions.com consumers had the ability to describe their experience using unstructured free text, like on other platforms such as Amazon.com. Next to their unstructured text review, authors also had the ability to provide a semi-structured summarizing free-text statement (key points) on the perceived advantages and disadvantages of the reviewed entity (denoted as pro and con statements). We choose those semi-structured pro and con statements published on Epinions.com as elicitation basis for our study because as well as (Decker and Trusov 2010) and (Gensler et al. 2015) we agree that those pro and con statements, that represent an aggregated view on the most important information contain consumers’ top-of-mind product and brand associations. This is of relevance because those associations are most quickly and easily extracted from memory. Those top-of-mind associations that (Lee and Bradlow 2012) synonymously refer to as the “most important product perceptions” are the strongest brand associations and represent the first things that comes to a consumers mind when confronted with a certain brand or product. Thus, we consider these semi-structured pros and cons as an adequate representation of a non-reactive data source incorporating consumer associations to products and brands and therefore particularly suitable to test whether UGC can serve as a non-reactive data source for the elicitation of brand associations.

To extract associations from the pro and con statements for formalizing and describing their linguistic characteristics, we apply preprocessing of (1) data cleaning and (2) tokenization. Data cleaning is necessary as we observe that review authors often use placeholder text (e.g. “nothing”, “everything”, “none”) and referrals (e.g. “see review”) because they don’t want to leave the pro and con statement online forms empty. To clean up those placeholders and referrals, we manually inspected 2,000 of the most frequent pro and con statements and identified 567 placeholder phrases (such as “none”, “see review” etc.). These amount to 27.5% of the 2,000 most frequent pro and con statements, which might be a hint that sometimes consumers have difficulties expressing their top-of-mind associations. Based on those 567 placeholder phrases and referrals, we composed a dictionary that serves as a stop-word list. We apply this stop-word list to the whole data set of 2.67 Mio. pro and con statements to exclude those statements consisting solely of placeholders or referrals as we consider those noise that does not add meaningful associations to the following analysis. Data cleaning finally removed 206,562 placeholders and referrals, which resulted in a final data set of 2.46 Mio. pro and con statements (1.3 Mio. pro; 1.16 Mio. con). In the second step of preprocessing we extract associations from the pro and con statements using tokenization. Similar to Gensler et al. (Gensler et al. 2015; Liu et al. 2005), we observe that consumers explicate multiple associations within each (pro/con) statement. This aligns with evidence from cognitive science as for each memory probe usually more than one association is retrieved from memory (Mitchell and Olson 1981). In our collected consumer reviews and in alignment with Gensler et al. (Gensler et al. 2015; Liu et al. 2005), we observe that review authors tend to separate their associations using conjunctions (e.g. “and”, “but”) and punctuation characters (e.g. “comma”, “semicolon”, or “period”) within their pro and con statements. To receive single associations, we thus tokenize each pro and con statement using punctuation characters and conjunctions as the tokenization rule. Thus, a pro statement like <nice brand and cool appearance> would be tokenized into the two associations <nice brand> and <cool appearance> while the con-statement <crappy design, price, and sound quality > would be tokenized...
to four single associations <crappy design>, <price>, <sound quality>, and <color>. As a result of prepro-
cessing, 5.06 Mio. associations are received which are linguistically characterized using the approach de-
scribed in the methodology section.

4.1 Analysis on Linguistic Characteristics

First, we investigate the number of associations and discover that consumers on average express 2.35 pos-
itive associations (standard deviation: 1.3) and 1.7 negative associations (standard deviation 0.94). Thus, as
consumers have to both report pros and cons, on average 4-6 associations are reported with a slight tendency
to report more positive than negative associations. This finding is completely in line with research on cogni-
tion psychology where participants were exposed to a certain stimulus (e.g. product or brand) and had to
report whatever came to mind. In the course of such a study Graesser and Mandler (Graesser and Mandler
1978) found that participants report 3-5 members of a category (e.g. product or brand) at a time, while Olson
and Muderrisoglu (Olson and Muderrisoglu 1979) found that participants retrieve 5-6 information chunks
for each product category or brand probe and Koll and Wallpach (Koll and Wallpach 2014) report 3.94
(standard deviation 1.35). Second, we investigate the length of associations. On average, each of the 5.06
Mio. associations consist of 3.6 words (e.g. “gives plants a home” or “easy to use”) with a standard deviation
of 2.7. Positive associations on average consist of 3 words (e.g. “great tasting coffee”) with a standard devi-
ation of 2.3. Negative associations on average consist of 4.5 words (e.g. “only few colors available”, “smaller
than it should be”) with a standard deviation of 3. Hence, consumers express associations in short phrases
on average consisting of 4 words (2-6 words when considering the standard deviation) with a slight tendency
to express positive associations with less words. This seems comprehensible, as negative associations often
contain added negation operators (e.g. “not” a good design). Third, we investigate the associations’ struc-
tural linguistic characteristics by means of PoS-sequences (as described in the methodology section). These
describe a Zipf-distribution where few PoS-sequences (Top-20) can be used to linguistically construct more
than half of all observed consumer associations (51.46% respectively 2.6 Mio.). To construct the other half
of associations, 473,245 (99.996%) distinct PoS-sequences would be necessary. Hence, we observe a high
consistency in how consumers verbally express their associations. Furthermore, the top 20 PoS-sequences
are rather short, ranging from 1-4 PoS-Tag components, which is in line with our investigation of the length
of associations. To determine the most important PoS-sequences, we plot similar to John et al. (2006) the
PoS-sequences by occurrence frequency and visually find the inflection point (elbow criterion) where the
difference in occurrence frequency drops. Following this approach, one would consider the elbow to be
located at the eighth most frequent PoS-pattern (<DET><NOUN>, e.g. “no batteries”) (see Figure 2). From
this follows that the top eight PoS-sequences would be the most important linguistic structures allowing the
construction of 42.19% of all associations. Taking a closer look at those PoS-sequences reveals that most
often associations are represented by <ADJ, NOUN> sequence (13.1 percent; 0.6 Mio), which allows the
construction of associations like <great coffee> or <nice design>. This construction is also often used in
opinion mining research when product features are linguistically described from a noun prepended by an
adjective transferring the sentiment to the product feature (e.g. (Liu et al. 2005)). This is supported by con-
sumer behavior theorists which widely agree that evaluations of product features are central to consumer
knowledge structures (Lawson 1998). Among the second to fourth most frequent PoS-sequences, we find
single nouns and compound nouns (e.g. <metal carafe> or <battery life>) representing concepts and named
entities. Furthermore, single adjectives are found. Those describe general attitudes (e.g. nice) or attributes
(e.g. sleek) which are association types that were described to be important for brands.
and Muderrisoglu (Olson and Muderrisoglu 1979) found that participants retrieve 5-6 information chunks
for each product category or brand probe and Koll and Wallpach (Koll and Wallpach 2014) report 3.94
(standard deviation 1.35). Second, we investigate the length of associations. On average, each of the 5.06
Mio. associations consist of 3.6 words (e.g. “gives plants a home” or “easy to use”) with a standard deviation
of 2.7. Positive associations on average consist of 3 words (e.g. “great tasting coffee”) with a standard devi-
ation of 2.3. Negative associations on average consist of 4.5 words (e.g. “only few colors available”, “smaller
than it should be”) with a standard deviation of 3. Hence, consumers express associations in short phrases
on average consisting of 4 words (2-6 words when considering the standard deviation) with a slight tendency
to express positive associations with less words. This seems comprehensible, as negative associations often
contain added negation operators (e.g. “not” a good design). Third, we investigate the associations’ struc-
tural linguistic characteristics by means of PoS-sequences (as described in the methodology section). These
describe a Zipf-distribution where few PoS-sequences (Top-20) can be used to linguistically construct more
than half of all observed consumer associations (51.46% respectively 2.6 Mio.). To construct the other half
of associations, 473,245 (99.996%) distinct PoS-sequences would be necessary. Hence, we observe a high
consistency in how consumers verbally express their associations. Furthermore, the top 20 PoS-sequences
are rather short, ranging from 1-4 PoS-Tag components, which is in line with our investigation of the length
of associations. To determine the most important PoS-sequences, we plot similar to John et al. (2006) the
PoS-sequences by occurrence frequency and visually find the inflection point (elbow criterion) where the
difference in occurrence frequency drops. Following this approach, one would consider the elbow to be
located at the eighth most frequent PoS-pattern (<DET><NOUN>, e.g. “no batteries”) (see Figure 2). From
this follows that the top eight PoS-sequences would be the most important linguistic structures allowing the
construction of 42.19% of all associations. Taking a closer look at those PoS-sequences reveals that most
often associations are represented by <ADJ, NOUN> sequence (13.1 percent; 0.6 Mio), which allows the
construction of associations like <great coffee> or <nice design>. This construction is also often used in
opinion mining research when product features are linguistically described from a noun prepended by an
adjective transferring the sentiment to the product feature (e.g. (Liu et al. 2005)). This is supported by con-
sumer behavior theorists which widely agree that evaluations of product features are central to consumer
knowledge structures (Lawson 1998). Among the second to fourth most frequent PoS-sequences, we find
single nouns and compound nouns (e.g. <metal carafe> or <battery life>) representing concepts and named
entities. Furthermore, single adjectives are found. Those describe general attitudes (e.g. nice) or attributes
(e.g. sleek) which are association types that were described to be important for brands (Keller 1993), and
those make up almost 22% of all associations. The remainder of the top eight PoS-sequences mostly describe
linguistic variations. For example, the PoS-sequence <ADJ, NOUN, NOUN> (e.g. “long battery life”) is
similar to the <ADJ, NOUN> (e.g. “long life”) but incorporates a compound noun instead of a singular noun.
Similarly, the PoS-Sequence <DET, NOUN> (e.g. “no batteries”) represents singular nouns prepended by a
determiner. Furthermore, we find single verbs <VERB> used in descriptive and subjective associations (e.g.
<entertaining>, <slow>, <challenging>, <overpriced>) referring to both attributes and benefits. Finally the
PoS-sequence <ADJ, PRT, VERB> allows the construction of associations like <easy to use>, <comfortable
to wear> or <difficult to assemble> which relate to benefits that (Keller 1993) considers as important brand
associations in his conceptualization. Finally, our results on linguistic characteristics from associations in
UGC are in line with Keller (1993) and Lawson (1998) who argue that brand associations are usually repre-
sented by concepts describing general characteristics, related products and topics, product uses, and summa-
rizing evaluations.
5 Study 2: Associations from Survey-based Research

To address research question 1 and validate whether UGC might serve as an adequate but cost-effective data source for brand association extraction, we compare linguistic characteristics of reactively collected brand associations (e.g. from surveys, focus groups, or interviews) to those from the first study (i.e. UGC as a non-reactive data source). To compose a data set of reactively elicited brand associations we performed a comprehensive structured literature review on empirical studies that collected brand associations using reactive approaches. As brand associations are a phenomenon widely studied in marketing research (Keller 2003), we focus on marketing journals as a literature base to gather reactively-collected brand associations. For this purpose we searched the top 71 leading marketing journals of the fourth version of the VHB-Jourqual ranking (see Schrader and Hennig-Thurau (Schrader and Hennig-Thurau 2009)). To identify relevant articles we formulated the following search pattern to search within the article titles, abstracts, and keywords: (Brand OR Product OR Consumer OR Customer) AND (Association Network* OR Concept Map* OR Association* OR Brand Association* OR Brand Perception OR Knowledge Structuring* OR Cognitive Structuring OR Memory Association* OR Perception). The search query was based on synonyms and differences in spelling and terminology of the main search phrase brand association and was derived by brainstorming and collecting keywords from initially identified articles. As a result of the structured literature research we collected 2,260 brand associations, form 28 articles (Table 1) out of 6,424 search results (articles) that collected associations via traditional reactive approaches such as surveys, focus groups, or interviews, and reported them.

![Figure 2. 20 Most Frequent PoS-Sequences in Associations from Product Reviews](image)

<table>
<thead>
<tr>
<th>Elicitation Method</th>
<th>Articles</th>
</tr>
</thead>
<tbody>
<tr>
<td>Quantitative</td>
<td>Bong Na et al. 1999, Dillon et al. 2001</td>
</tr>
</tbody>
</table>

Table 1. Literature Research on Empirical Brand Association Studies

When investigating the length of consumer associations we find an average length of 2.36 words (standard deviation of 3.97) in Study 2. Thus, in comparison to Study 1, consumer associations from survey-based research are much shorter on average (2.36 vs. 3.6) while having a higher standard deviation (3.97 vs. 2.7). The shorter average length might be explained as it’s not uncommon in various reactive research instruments (e.g. Christensen and Olson 2002; Henderson et al. 1998; 2002; John et al. 2006) to incorporate a stage where the researcher aggregates associations to resolve synonyms or to homogenize associations which lead to linguistic changes (such as the length). So, associations in Study 2 might have been post-processed, which we will control for in Study 3. However, the higher standard deviation of association length in Study 2 might also have different reasons. First, it might be the result of the higher diversity of elicitation methods applied in the collected articles. For example, some of the 28 investigated studies used sentence completion tasks...
such as “for me, the brand XY is …”. The participant’s answer might be dictated by the structure of the question, which implies a certain structure of the answer. It is unlikely that a participant would complete a sentence like the one above using a conjunction. Second, no common guidelines on when and how to perform post-processing on associations exist, which might result to a higher diversity in performed aggregations.

5.1 Analysis of Linguistic Characteristics
With regard to structural linguistic characteristics by means of PoS-sequences, similar to Study 1, a Zipf-distribution can be observed. The 20 most frequent PoS-sequences (out of 276 = ~7%) are able to construct almost 84% (Study 1: 51.46%) of all collected brand associations. Applying the elbow criterion to find out the most important PoS-sequences reveals, that the top 6 (Study 1: top 8) allow for the construction of 75% (1,695) of all associations, while the other 270 PoS-sequences are needed to construct the remainder (565). When comparing the PoS-sequences between Study 1 and 2, we find that all PoS-sequences from Study 2 are also found in Study 1, both within the top 6 most important PoS-sequences of Study 2 and on the whole data set. Meanwhile, within the top 20 the coverage remarkably drops to 65%. To additionally find and compare hidden linguistic patterns, we now investigate both data sets for frequent closed sequential patterns (FCSPs) hidden in brand associations using the BIDE algorithm (Wang and Han 2004). We used a minimum support of 1%, which resulted in about 500 FCSPs for each data set.

<table>
<thead>
<tr>
<th>FCSPs (Survey)</th>
<th>Example</th>
<th>Support</th>
<th>FCSPs (UGC)</th>
<th>Example</th>
<th>Support</th>
</tr>
</thead>
<tbody>
<tr>
<td>NOUN</td>
<td>restaurant</td>
<td>1,512</td>
<td>66.93%</td>
<td>NOUN</td>
<td>quality</td>
</tr>
<tr>
<td>ADV</td>
<td>expensive</td>
<td>990</td>
<td>45.82%</td>
<td>ADV</td>
<td>funny</td>
</tr>
<tr>
<td>ADJ NOUN</td>
<td>good quality</td>
<td>438</td>
<td>19.39%</td>
<td>ADJ NOUN</td>
<td>nice features</td>
</tr>
<tr>
<td>VERB</td>
<td>refreshing</td>
<td>423</td>
<td>18.73%</td>
<td>VERB</td>
<td>lasts</td>
</tr>
<tr>
<td>ADV (1)</td>
<td>rarely</td>
<td>211</td>
<td>9.34%</td>
<td>ADP (1)</td>
<td>on</td>
</tr>
<tr>
<td>ADP (1)</td>
<td>in</td>
<td>206</td>
<td>9.12%</td>
<td>ADV (1)</td>
<td>too</td>
</tr>
<tr>
<td>VERB NOUN (4)</td>
<td>promotes equality</td>
<td>185</td>
<td>8.19%</td>
<td>DET (4)</td>
<td>no</td>
</tr>
<tr>
<td>NOUN NOUN</td>
<td>ice cubes</td>
<td>182</td>
<td>8.06%</td>
<td>NOUN NOUN</td>
<td>battery life</td>
</tr>
<tr>
<td>ADP NOUN (1)</td>
<td>for kids</td>
<td>178</td>
<td>7.88%</td>
<td>DET NOUN (4)</td>
<td>no batteries</td>
</tr>
<tr>
<td>VERB ADJ (6)</td>
<td>smells good</td>
<td>127</td>
<td>5.62%</td>
<td>ADP NOUN (1)</td>
<td>at home</td>
</tr>
<tr>
<td>DET (4)</td>
<td>no</td>
<td>126</td>
<td>5.58%</td>
<td>VERB NOUN (4)</td>
<td>thickens hair</td>
</tr>
</tbody>
</table>

Table 2. 11 Most Frequent Closed Sequential Patterns.

Table 2 shows the 11 most outstanding FCSPs by means of their support (ratio PoS-sequences where the FCSP occurs). The number next to the FCSPs shows the difference in ranking in the respective other data source, while the bold font indicates FCSPs that are not found in the top 11 FCSPs. The result shows that not only nearly the same FCSPs were found, but also that FCSPs are ranked similarly. The application of Spearman’s rank correlation coefficient reveals a rank correlation of 0.83 among the top 50 FCSPs. Sequential pattern mining not only reveals nearly the same FCSP of brand associations from UGC and survey-based data sources, but also that the ranking is linearly correlated. Study 1 and 2 reveal a high similarity by means of PoS-sequences, similar to Study 1, a Zipf-distribution can be observed. The 20 most frequent PoS-sequences (out of 276 = ~7%) are able to construct almost 84% (Study 1: 51.46%) of all collected brand associations. Applying the elbow criterion to find out the most important PoS-sequences reveals, that the top 6 (Study 1: top 8) allow for the construction of 75% (1,695) of all associations, while the other 270 PoS-sequences are needed to construct the remainder (565). When comparing the PoS-sequences between Study 1 and 2, we find that all PoS-sequences from Study 2 are also found in Study 1, both within the top 6 most important PoS-sequences of Study 2 and on the whole data set. Meanwhile, within the top 20 the coverage remarkably drops to 65%. To additionally find and compare hidden linguistic patterns, we now investigate both data sets for frequent closed sequential patterns (FCSPs) hidden in brand associations using the BIDE algorithm (Wang and Han 2004). We used a minimum support of 1%, which resulted in about 500 FCSPs for each data set.

6 Study 3: Associations from Free-elicitation Study
When comparing the linguistic characteristics of consumer brand associations between Study 1 and 2 we find high similarities in linguistic characteristics, but also differences. The latter is especially found with regard to the length and the distribution of PoS-sequences (associations from reactive elicitation are shorter and less diverse). One explanation for these differences might emerge from the reactive elicitation instruments themselves. These often incorporate a stage where associations are post-processed by researchers to homogenize associations, resolve synonyms, or aggregate associations to higher classes of meaning. This eventually might lead to less diversity in linguistic patterns and shorter associations. To control for these distortions in linguistic characteristics we conduct Study 3: We collect consumer associations ourselves via
a traditional survey-based reactive instrument and compare collected raw and unprocessed associations’ linguistic characteristics to those from Study 1 and 2. To collect brand associations ourselves we use the free elicitation method (Olson and Muderrisoglu 1979) an established instrument that is considered as one of the most powerful methods to elicit brand associations (Supphellen 2000). In free-elicitation participants are asked for their initial associations regarding a defined stimulus and focuses on easily accessible verbal associations from semantic memory (Krishnan 1996). In our study, we recruited 50 participants using the Amazon Mechanical Turk platform. All participants came from North America and were 25 to 40 years old, 64.5% were male, 50% finished high school or college, while the other 50% received a Bachelor’s or Master’s degree. In our study, participants were asked in an online survey to express things that initially come to their minds when they think of the brand Google within the topic of autonomous driving. We used this question as the stimulus, as it fosters participants to retrieve both product, brand, and general associations, and thus a wide variety of associations.

6.1 Analysis of Linguistic Characteristics

The results of Study 3 show that the frequency distribution of PoS-sequences is similar to Study 1 and 2, and also all PoS-sequences are found as in respective other studies. Furthermore, from comparing the length of associations as well as the linguistic structure, we find evidence for our tentative explanation that differences in linguistic characteristics from Study 2 might be a result of post-processing, which is typically performed in reactive association elicitation instruments. In Study 3, we find identical to Study 1 (UGC, online consumer reviews), an average length of associations of 4, while Study 2 (survey-based research) reported a smaller length of 2 words. Furthermore, we observe that the top 20 most frequent PoS-sequences account for 63% of all expressed associations in Study 3, which is still 12% higher than in Study 1 (online consumer reviews), but 21% lower than in Study 2 (84%; survey-based research). Finally, our results from Study 3 support that there is consistency in how consumers linguistically express brand associations independent from the medium and elicitation situation. Thus, from the perspective of linguistic patterns, UGC might be an adequate addition or alternative to manually conducted surveys in eliciting brand associations.

7 Study 4: Ordinary Written Speech – The Google NGram Corpus

The linguistic characteristics we have derived from consumer brand associations so far might be meaningless if they are simply dictated by the syntax of language. Thus, neither the construction of information systems to extract consumer associations from natural language text would be possible, nor the indications on concurrent validity between UGC as a non-reactive data source and reactive data sources would be valid; any comparison of textual content based on language syntax would lead to the same results. To control for this situation we conduct Study 4, which aims to discover whether linguistic characteristics (by means of PoS-sequences) differ between ordinary speech and consumers’ brand associations. To do so, we use NGram analysis and compare n-gram frequencies of ordinary written speech to the combined set of consumer associations from user-generated product reviews (Study 1) and survey-based research (Study 2 and 3). As a reference corpus to ordinary written speech, we use the Google Books NGram Corpus (Michel et al. 2011).

In its second version, the Google Books NGram Corpus project digitized 8,116,746 books from the last 500 years, which amounts to 6% of all books ever published (Lin et al. 2012). As the second version of Google Books NGram Corpus now also covers n-grams on PoS-Tags using the universal PoS-Tag-set (Petrov et al. 2012), we are able to compare PoS-Tag n-grams derived from verbally explicated brand associations with the n-grams that represent an average of written English-language. As the basis for our comparison analysis, we used the Google Books NGram corpus for the English language covering books between 1998 and 2008.

7.1 Comparison of linguistic characteristics

When comparing the relative frequency of unigrams from our associations, we find that nouns and adjectives are by far more represented, exceeding the frequency of nouns and adjectives within the Google NGram Corpus by 9.5% and 10%. Slightly higher is also the frequency of adverbs (+3.8%) and verbs (+1.18%). Probably nouns are much more frequent in verbally explicated associations because they are used to describe concepts, which seems important when explicating brand associations. The stronger presence of adjectives (+10%) and adverbs (+3.81%) within verbal associations with respect to the frequencies within the Google n-grams hints to the strong emotional and subjective component of brand associations. On the other hand,
we find the unigram adpositions (-2.88%), determiner (-2.15%), pronouns (-1.13%), cardinal numbers (-2.13%), and conjunctions (-2.61%) occur less often within associations. The fewer number of conjunctions can be explained by our approach in Study 1 where we used punctuation characters as well as the “and” conjunction to separate multiple associations. The fewer number of cardinal numbers underpins the fuzziness of perceptions since cardinal numbers are used to quantify while associations have a more subjective character where quantifications might be not critical. When comparing the top five bigrams, we observe that the PoS-sequences <adjective, noun> and <noun, noun> occur more than two times more often within associations than in written speech, which aligns also with the result that nouns occur more frequently in associations. However, prepending a determiner to the noun is more characteristic for written speech as well as using adpositions after the noun. Within the <noun, verb> PoS-sequence only minor differences in frequencies can be observed (-0.14%). Finally, when comparing 3grams we see that only the <adjective, noun, noun> PoS sequence occurs three times more often within associations. As the <noun, noun> component in this PoS-sequences usually describe compound nouns, the PoS-sequence can be seen as a specialization of the <adjective, noun> PoS-sequence, which also occurs two times more often for consumer associations. Compared to ordinary written speech, especially the most-frequent PoS-sequences (that allows the construction of most associations) differ in usage frequency and lead to major differences between ordinary written speech and verbal expressions of consumers’ brand associations. Therefore, we conclude that PoS-sequences found in brand associations differ from ordinary speech and thus are not dictated by the syntax of language itself, which should enable the construction of algorithms to identify brand associations in natural language.

8 Discussion and Conclusion

Consumers’ brand associations can be drivers for marketing strategies and thus are of practical relevance for marketers (Keller 1993, 2003; Faircloth et al.2001). Therefore, marketers are interested in fast cost-efficient brand image elicitation approaches. To enable the construction of automated marketing intelligence systems that can elicit a brand’s image fast and cost-efficient, it first has to be investigated whether brand associations can be formalized in a way that allows an automated elicitation (like product features) and if cost efficient data sources like UGC are utilizable. Thus, the article at hand addressed two research questions. First, we investigated whether UGC might serve as an adequate data source for brand association elicitation to foster concurrent validity to established reactive data sources (e.g. interviews and surveys) (1). Second, we investigated the ability to structurally describe and formalize characteristics of brand associations to foster the development of information systems to elicit brand image with respect to construct validity (2). We approached both research objectives in formalizing and describing linguistic characteristics of how consumers express brand associations. In doing so, we performed a comprehensive quantitative analysis on PoS-sequences of 5.06 million brand associations from UGC using 1.63 million user-generated online product reviews (Study 1) and compared those to 2.226 associations collected from 28 empirical studies (Study 2). To control for modified linguistic characteristics due to aggregation steps, which are often performed in reactive elicitation instruments, we conducted a free elicitation control study (Study 3) to elicit raw and unprocessed associations. Furthermore, to rule out that found linguistic characteristics are simply dictated by the syntax of language, we conducted another control study to compare linguistic characteristics of brand associations to those of ordinary speech based on the Google Books NGram corpus (Study 4). With regard to research question 1, our results show that in UGC consumers on average express 4-6 associations for a stimulus, which is in line with previous research on the number of associations consumers typically retrieve in reactive research studies (Graesser and Mandler 1978; Olson and Muderrisoglu 1979). Furthermore, the comparison of linguistic characteristics between reactive- and non-reactive data shows high similarity and consistency by means coverage (all associations’ PoS-sequences occur in both datasets), distribution (both PoS-sequence frequencies follow similar power-law distributions), and high rank-correlation of frequent closed sequential patterns. Thus our studies provide strong indications on concurrent validity between non-reactive and reactive data sources and we conclude, that UGC might serve as an adequate data source for brand association elicitation. With regard to research question 2, the formalization and description of brand associations’ linguistic characteristics shows that associations are retrieved from consumers’ memory as short information chunks ranging from 2 to 4 words. Associations predominantly describe product attributes and consist of single nouns (e.g. “design”), nouns prepended by a determiner (e.g. “no batteries”), as well as the adjective-noun construction (e.g. “great screen”). Furthermore, non-product-related attributes are characterized by single adjectives (e.g. “traditional”) and the adjective-noun (e.g. “great brand”) construction. General benefits
are found being described by single verbs (e.g. “satisfying”), adverbs relating to adjectives (e.g. “very English”) and the adjective-particle-verb construction (e.g. “easy to use”). However, the bulk of associations seem to be foremost descriptive product-related attributes using adjectives and nouns and especially the <adjective, noun> construction. The latter is in line with previous research on consumer behavior, which claims that product features are most often the foundation of consumer knowledge structures and can be retrieved from memory most easily (Lawson 1998). As a result, our studies demonstrate that structural attributes of brand associations can be described by linguistic characteristics, that as Study 4 proves are unique to consumers’ verbalized associations and not being dictated by a language’s syntax itself, because differences in n-gram frequencies between ordinary speech and associations could be shown. Thus linguistic characteristics offer fundamental information to be considered in the construction of automated algorithms that are targeted towards the automatic extraction of brand associations from unstructured natural language text. However, they prompt additional research, because even though most associations are expressed by an <adjective, noun> construction also often used for discovering product features, other frequent and differentiating brand association patterns unfortunately solely consist of single items such as nouns, adjectives, and verbs. Extracting all nouns and adjectives from unstructured text would not lead to meaningful brand associations; the brand-relevance cannot be inferred from the presence of a noun, adjective, or verb alone. We conclude that the automatic extraction of brand associations’ product attribute components can be addressed using opinion mining approaches, but the elicitation of other brand association types such as general benefits or attitudes (Keller 1993) has to be considered as being very challenging and requiring additional research on how to detect brand-relevance within unstructured natural language text. Beside the already-stated challenges that our research emphasizes, general limitations also exist. We only investigated the most salient verbally expressed consumer associations. However, there are also associations that are psychologically more complex (e.g. self-esteem) in a way that consumers themselves are often not aware. Those associations are usually elicited by asking for justification chains (e.g. using laddering techniques). As eliciting those associations requires interaction with the researcher, we claim that neither our studies, nor UGC in general, will cover those. Nevertheless, this research helps towards understanding how consumers verbally describe the content of knowledge structures in terms of brand associations. Thus this research is a first step to create the preconditions for the development of novel algorithms and information systems aiming to determine brand-relevance in unstructured text data to extract presumably syntactically simple, but hard to detect, types of brand associations (such as attributes, attitudes, and benefits described by single nouns, adjectives, or verbs). This would enable brands and researchers to elicit a brand’s image both in a novel non-reactive way (from UGC) and also from reactive data sources (like interviews and surveys) to tackle the increasing demand for frequent, cost-efficient, valid, and reliable research instruments.

Acknowledgements: This research is supported by DFG grant 1452/1 GZ: SCHO 1321/3-1/
9 References


