Food Quality and Preference 47 (2016) 34–44

Contents lists available at ScienceDirect

Food Quality and Preference

journal homepage: www.elsevier.com/locate/foodqual

Analyses of open-ended questions by renormalized associativities and textual networks: A study of perception of minerality in wine

Pascale Deneulin [a,b,*], François Bavaud [b]

[a] University of Applied Sciences and Arts, Western Switzerland, Changins – Viticulture and Oenology, Switzerland
[b] Department of Language and Information Science, University of Lausanne, Switzerland

ARTICLE INFO

Article history:
Received 1 November 2014
Received in revised form 12 May 2015
Accepted 18 June 2015
Available online 26 June 2015

Keywords:
Automatic lemmatization
Correspondence Analysis
Minerality
Modularity
Open-ended questions
Renormalized Markov associativities
Term co-occurrences
Textual statistics
Verbal-based survey

ABSTRACT

Verbal tasks are increasingly used in food science, but still often suffer from time-consuming manual preprocessing procedures. Also, traditional visualization techniques are not always successful at clearly revealing the structure of term co-occurrences. The present study proposes a few statistical innovations in the analysis of textual data resulting from an open-ended survey on minerality perception, without tasting phase. First, we use dedicated, amenable software aimed at producing term lemmatization and construction of contingency table, enabling minimum manual verification and correction. Furthermore, co-occurrences are treated as a textual weighted network, which can be further iterated and renormalized in a flexible way, filtering out rare terms and their associations. In addition, visualization and clustering techniques, initially developed in social networks studies, reveal meaningful and well-defined terms communities, corresponding to distinct conceptions of minerality. Results are exclusively based upon statistical methods, without resorting to semantic nor linguistic considerations. Altogether, they demonstrate the polysemy and ill-definiteness of the concept of minerality among wine professionals.

1. Introduction

Many original proposals have recently appeared in sensory science, aiming to improve the description of sensory perceptions and consumers' preferences. They characteristically seek to be faster for participants, less time-consuming for analysts and overall cheaper. Typically, expert panel with long and expensive training period tends to be replaced by a few shorter training periods, or even by direct evaluations by consumers. The main difficulty arising from working directly with untrained judges is to assess if the same word or descriptor carries the same signification for all participants. Moreover, using synonyms does not guarantee to refer to the same perception.

Verbalization tasks are increasingly used, as reflected by the number of related publications in food science journals, which increased from 1990 to 2012 by a factor twelve (Bécue-Bertaut, 2014). Open-ended questions have been addressed in sensory science by Kleij and Musters (2003) who asked consumers to explain why they gave a particular liking score on different mayonnaises. Analysis of these responses, spontaneous and unstructured, was time-consuming and required a subjective preprocessing task, not entirely reproducible.

Despite the difficulties, those studies provided useful results, reinforcing quantitative findings. To distinguish between positive and negative likes in consumers comments, Symoneaux, Galmarini, and Mehinagic (2012) separated answers in two parts. The transcription and the interpretation made the understanding of consumers' comments easier, but the analysis was still time-consuming. Free comments were also used to describe milk desserts (Ares, Giménez, Barreiro, & Gámbaro, 2010), wines (Bécue-Bertaut, Álvarez-Esteban, & Pagès, 2008; Lawrence et al., 2013; Sauvageot, Urdapilleta, & Peyron, 2006), and textures of cheeses (Hanaei, Cuvelier, & Sieffermann, 2015). In other activity fields, Chung and Pennebacker (2008) used a writing exercise to study the personality of students and Lowe et al. (2013) highlighted main themes of alcohol-related behavior by asking consumers to explain what they thought about drinking, as well as when and why they drank or not. All these studies underlined the difficulties associated with subjective and time-consuming pre-treatments. In most cases, words were replaced by their standardized form and synonyms, considered to carry the same meaning, were aggregated together in the final dataset.

Among alternatives aiming at overcoming those pitfalls, let us cite Antmann et al. (2011) who used a free-listing to generate vocabulary about texture. This technique, introduced by Hough and Ferraris (2009) consists in asking participants to "list all the

* Corresponding author.
E-mail address: pascale.deneulin@changins.ch (P. Deneulin).

http://dx.doi.org/10.1016/j.foodqual.2015.06.013
0950-3293/

X terms they know about", where X is the topic of the study. Libertino, Ferraris, López Osornio, and Hough (2012) applied it to the study of menus among different income-level populations. An alternative to free-listing is to limit the list to a few words, as in Guerrero et al. (2010) where the concept of traditional food is handled by limiting participants to three relevant words, or as in Ares et al. (2010) with four words qualifying milk desserts.

After pre-treatments, textual data are typically expressed as term × product (or respondent) contingency tables, amenable to overall or cell-wise tests of independence (Symoneaux et al., 2012), and further decomposable and visualizable through Correspondence Analysis (Benzécri, 1973; Lebart, Salem, & Berry, 1998). It is also possible to mix a contingency table with quantitative data (consumers' preferences or descriptive analyses for instance) using a Multiple Factor Analysis (Bécue-Bertaut & Pagès, 2008).

The present study presents a few innovative point of views and treatments of textual data obtained from open-ended questions about *minerality in wine*, a new concept which emerged a few years ago in oenology. Our approach is limited to quantitative analysis, without semantic nor linguistic considerations.

First, we propose automatic pre-treatments with free software (Section 2), original in sensory science. Standard Correspondence Analysis is then performed (Section 3.1), yielding meaningful term clustering after manual removal of rare terms and first factorial dimensions. In the second place, we present how term co-occurences define a textual network of *Markov associativities* (Section 3.2), as well as higher-order association networks obtained by iteration. Edge weights can be further renormalized by taking into account the frequencies of the associated pair of words, automatically reducing the impact of rare terms (Section 3.4). In this picture, term clustering becomes a network clustering issue, and strategies based upon the *modularity* and, indirectly, *normalized associativity* criteria (Shi & Malik, 2000; Girvan & Newman, 2002) are recalled and compared in a formal perspective (Sections 3.2 and 3.6).

Finally, results (Section 4) illustrate the enhanced interpretability and the flexibility of the approach. In particular, in spite of the apparent polysemy of "minerality" at the intra-respondent level (a single answer possibly refereeing to various topics), the emergence of distinct communities (or clusters) of terms can be observed and quantified.

2. Data collection and preprocessing stages

A survey addressed to French-speaking wine professionals from Switzerland and France was diffused on-line in 2012 and 2013. It consisted of three open-ended questions, generating three sub-corpora further referred to as *evocation, definition* and *synonyms*. In this paper, we illustrate the method on the sub-corpus *evocation* only, resulting from the free answers to the question *"If I speak to you about minerality in wine, what does come to your mind?"*. A total of 1898 answers (in French) have been collected, representing a total of 52,316 terms (tokens), 4353 of which being distinct (types) before lemmatization. As answers were mandatory, some respondents wrote particular marks such as "?" or "...", indicating lack of familiarity with the concept of minerality, which were further regrouped into a single symbol (Deneulin, Gautier, Le Fur, & Bavaud, 2014).

Answers were widely variable regarding their lengths, as well as the presence of abbreviations, syntactic and spelling mistakes. Further preprocessing did consist of:

1. Semi-manual correcting syntactic and spelling mistakes, with the help of an automatic spelling software.
2. Semi-automatic term lemmatization, consisting in transforming terms into their normalized forms (infinitive verbs, masculine singular adjectives, masculine singular nouns), with the software TreeTagger (Schmid, 1994), available for different languages. TreeTagger considers morphosyntactic criteria such as grammatical forms. As a result, poorly structured sentences and ambiguities may generate mistakes in the lemmatization, which must be manually verified.
3. Removal of functional terms (stop words). In our study, we decided to keep only common and proper nouns, verbs and adjectives, and to remove all others grammatical forms. This step, as well as the construction of the final respondent-term contingency matrix, was performed with the free and flexible software Textable (Xanthos, 2014).

Before analyzing data, negations have been considered. Inspection of collocations reveal that terms "not" and "little" (respectively "pas" and "peu" in French), appearing 291 times, respectively 127 times in this corpus, have little or no specific connection with the other terms. They have been removed, without altering the overall lexical meaning of the corpus.

The resulting 25,908 tokens, distributed along 2634 types, are further regularized by removing terms occurring less than five times, often specific to individual respondents, leaving a final set of 642 types. Also, the frequent terms "to be" and "wine", whose associations are not specific nor informative, have been removed.

Recall that one of the main targets of the study was to reveal the expected semantic variability of the concept of minerality, together with its synonyms, and irrespectively of the consumer characteristics. In particular, and contrarily to others studies (Hanaei et al., 2015; Lawrence et al., 2013; Kleij & Musters, 2003), we did not group together terms considered as synonyms.

3. Textual network, renormalized associativities, and clustering

After preprocessing, textual data consist of the term × respondent contingency matrix $N = (n_{ik})$, counting the emissions of term i by respondent k. Those data can be processed by Correspondence Analysis, but can also generate textual networks and Markov transitions between nodes. Specifically, textual contingency matrices enable to define in turn

(a) term dissimilarities, such as the chi-square or the Hellinger (square) distances (Section 3.2). Both constitute squared Euclidean distances, thus enabling to factorially embed the term configuration in an Euclidean space (Section 3.3).
(b) textual networks, whose nodes represent the terms and whose valued edges constitute a measure of association between terms (Gaume, 2004; Bavaud & Xanthos, 2005). Symmetric measures of term associations constitute term similarity indices and define an unoriented network (Section 3.2).

Both representations generate in turn further associated treatments, such as traditional distance-based textual clustering or spectral clustering for textual networks. Depending upon initial definitions, the two frameworks may or may not be mathematically equivalent.

In particular, edge weights can be *absolute*, that is proportional to the term frequencies of the two nodes at stake, or *relative*, that is size-independent. Unfortunately, the question of the absolute versus relative nature of edge weights is often neglected in the literature on applied networks. This lack of precision entails damaging consequences, in our opinion, for further methodological progress and unification. By contrast, this paper insists upon this important issue: we introduce a generalized family of *renormalized edge*