In the Eye of the Beer-Holder.

Lexical Descriptors of Aroma and Taste Sensations in Beer Reviews

Els Lefever, Liesbeth Allein and Gilles Jacobs

LT[3], Language and Translation Technology Team

Email: els.lefever@ugent.be, liesbeth.allein@ugent,be, gillesm.jacobs@ugent.be

Abstract—Western languages do not dispose of a well elaborated vocabulary for describing smell and flavour sensations. We investigate whether beer experts share a common vocabulary to describe beer properties. We collected an English text corpus of beer reviews and analyzed the lexical descriptors used by beer-tasting experts to describe aromas and flavours. The informativeness of beer reviews was investigated by running a machine learning experiment for predicting the colour of a beer based on the review text. This preliminary experiment shows promising results, with average accuracy figures of about 60% for automatic beer colour prediction. Our experimental results show that beer experts share a common vocabulary to describe beer characteristics in a consistent way, allowing to automatically predict beer properties based on the review text.

Keywords–Corpus linguistics; Natural language processing; classification; beer reviews

I. INTRODUCTION

It appears that people in the western world are not very good at describing smells and flavours. Research has shown that western languages have few words to describe smells and flavours, in contrast to visual phenomena, for which we dispose of a well elaborated vocabulary [1]. The description of taste, smell, and sight remains evaluative and non-specific for non-experts as [2] has shown in studies conducted with wine experts, coffee experts, and novices. Both coffee and wine experts tend to use more specific source-based terms (metaphors or *it smells like* + source), while novices use more evaluative terms (e.g., *nice, bad, good*). Consequently, we can assume that reviews written by beer experts should contain more source-based descriptions and fewer evaluative terms in order to describe smell, taste, and sight.

Previous research has shown that the perception of foods and drinks depend on both the visual and orthonasal sensory inputs, especially before the tasting [3]. The colour and look of a drink influence both the perception of smell and the way the taster describes his/her perceptions [4]. Furthermore, people link certain smells with certain colours [5]. It can therefore be concluded that perceptions of sight are closely related to perceptions of smell. Consequently, certain sight descriptions in the reviews will often be accompanied by the same smell descriptions and certain smell descriptions will be accompanied by the same sight descriptions. For this reason, an automatic prediction system could predict missing sight properties based on the smell descriptions in the reviews it is often accompanied by, and vice versa. For example, "gold" could refer to the colour of a beer, and in the corpus, reviews about light-coloured beer often contain the words *herbs*, *spicy*, and *butter* to describe the aroma of that beer. Therefore, an automatic prediction system could predict the colour of the beer in a review, which lacks sight or colour descriptions, based

solely on the aroma description. Such a system could tell if a beer is gold even though the reviewer does not mention any colour or sight property.

Not only descriptions of smell and sight, but also descriptions of smell and taste/flavour are claimed to be closely connected. According to [6], the flavour of food is described by both gustatory and olfactory stimuli. There are two olfactory stimuli: orthonasal smell and retronasal smell, which are respectively the smell we sniff before the food or drink is tasted and the smell that is pulsed out after the food or drink had been swallowed. It is the combination of retronasal aromas and gustatory cues that defines a flavour and leads to descriptions such as *fruity* and *malty* [6]. In [7], the authors consider odour-taste synaesthesia (smelling tastes/tasting smells) a factor of the link between smell and taste. When people smell an odour, they recognize the smell as a taste and describe it as such, e.g., something *smells sweet*. This is also due to the co-occurrence of retronasal odour simulation and oral stimulation and the result of a unitary perception [7].

For this research, we have compiled a corpus of American beer reviews (See Section II). Our hypothesis is that the reviewers working for this website will be subject to odour-taste synaesthesia and describe flavours and aromas as such. As a consequence, certain smell descriptions in the reviews should often be accompanied by the same taste descriptions and vice versa.

This leads us to our first research question (RQ1): is it indeed the case that taste and smell descriptions are closely linked, and do beer reviewers, by consequence, use the same lexical descriptors for taste and smell?

The second research question (RQ2) is the following: are the expert beer reviews meaningful providers of information considering the limited vocabulary and ways of describing sensory perceptions such as smells, flavours, and sight? Therefore, sensory experiences should be worded in a consistent manner.In [8], the authors have shown this is the case for authors of wine reviews. They built classifiers to predict colour, grape variety, country of origin, and price of a wine, based on the experts' wine reviews. The experimental results showed promising F-scores, demonstrating that wine reviews really are informative.

In this research, we investigate whether beer experts share a common vocabulary to describe beer properties. The consistency of the descriptions will be verified by building a machine learning system to automatically predict beer properties for new beer reviews on the basis of smell/aroma and taste/flavour descriptions from experts' reviews. For these preliminary experiments, we build a system that predicts beer colour labels on the sole basis of smell and taste properties. This means that the system can assign colour properties to the beer in

the review even though colour descriptions are lacking. The automatic prediction system then bases its colour property predictions on the smell and taste descriptions present in the review. The general hypothesis is then that beer experts, just like wine experts, are capable of describing beer properties in a sufficiently consistent manner, which allows beer properties to be automatically predicted on the basis of experts' reviews. Training automatic systems to predict beer characteristics could be a first step to develop content-based recommender systems for beer. Whereas current recommender systems only take metadata like beer style (e.g., IPA) and user-based filtering or subjective ratings into account, beer recommendations based on review content and aroma and taste descriptions could be very useful.

The remainder of the paper is structured as follows: Section II describes the beer review corpus we used for these experiments. Section III elaborates on the colour classification experiments for automatic prediction of colour based on smell and taste descriptions in the reviews (RQ2), while Section IV presents the lexical analysis of the language that is used for describing smell and taste in beer reviews (RQ1). Finally, in Section V, we draw conclusions and present prospects for future research.

II. Corpus

For our experiments, a corpus of online beer reviews written by experts is composed. Experts are widely considered to be more accurate and detailed in their smell, taste and sight descriptions, which is important for the construction of an automatic prediction system. In [9], the author claims that experts are biologically superior to novices when it comes to distinguishing tastes.

We have collected 2205 beer reviews from the American website Tastings.com [10] and automatically extracted per review the following structured beer properties: *name, category, alcohol, country, style, aroma, flavor, bitterness*, as well as the review text written by the expert. Figure 1 shows an example of the structured beer properties that are listed for the different beers.

TASTING INFO

	Style:	Spicy & Complex & Malty
	Aroma:	toasted banana-raisin muffin and spicy vanilla custard
	Flavor:	long, elegant
	Bitterness:	Low
	Enjoy:	Enjoy on its own
	Pairing:	Beef Stew, Peking Duck, Morbier
	Bottom Line:	A fantastic flavor ride and an archetype of the style.

Figure 1. Example of the structured beer properties .

Example 1 lists the review text accompanying the structured properties of the *Westmalle Trappist Double* beer:

(1) *Hypnotic reddish mahogany color. Rich aromas of toasted banana-raisin muffin and spicy vanilla custard with a satiny, fruity-yet-dry medium-full body and a long, elegant finish with notes of caramelized nuts and dried fruits, peppery spice, and earth. A fantastic flavor ride and an archetype of the style.*

III. Classification Experiments

The task of predicting the colour of the beer was conceived as a supervised classification task. Two sets of bag-of-words features were extracted as information sources from the expert review texts: unigrams (single words) and bigrams (sequences of two words). As we want to investigate the viability of automatically predicting the colour of the beer based on the sole review text, sentences containing a colour description were automatically removed from the review. The reviews were further preprocessed by removing all punctuation marks and by lower-casing all words contained in the review.

Unlike the other beer properties, colour descriptions could not be automatically extracted from the website, because they were only present in the written review itself and not in the structured information about the beer in the website. Therefore, the word preceding the word *colour* in the review was extracted as the colour description. If the review contains for instance *Cloudy golden color with a high head*, the word preceding *color*, being *golden* in this case, is automatically extracted as the colour label and the entire sentence is removed from the review text. Altogether, 49 different colour labels were extracted following this procedure. This high number of classes, however, makes the colour classification of beers from new reviews less accurate. Therefore, colour terms referring to the same colour category were grouped together and 7 new classes were formed (*very light, yellow, amber, brown, black, red/rose*, and *green*). Other colour terms could not be grouped within these classes, and were kept as individual colour classes. This was the case for the colours: *deep, hazy, cloudy*, and *oak*. Some of the latter colour labels do not refer to a specific colour, but are artifacts introduced by the automatic extraction of the label. For the experiments, colour labels only occurring once in the corpus were not considered (*brilliant, violet, indigo, gray, platinum, nickel, wood*). This way we ended up with 11 different colour categories or classes, which are shown in Table I. Beer reviews where no colour label could be extracted, were removed from the database, resulting in a reduction of the corpus from 2205 to 2121 instances.

TABLE I. COLOUR CLASSES.

Colour category	Colour labels
very light	silver
gold	gold, yellow, golden, sunburst, straw, light, bright brassy, sunset, white, sunrise
amber	amber, copper, bronze, orange, maroon, penny
brown	brown, mahogany, medium-brown, walnut dark-brown, sienna
black	black, ebony, cola
red/rose	ruby, garnet, pink, red, salmon
green	green, emerald
cloudy	cloudy
deep	deep
hazy	hazy
oak	oak

As a classification algorithm, we used Support Vector Machines as implemented in the LIBSVM toolkit [11]. As evaluation measures, we report (ten-fold cross-validated) (1)

Precision, (2) Recall and (3) F_1-score per colour class, calculated as follows:

$$Precision = \frac{Number\ of\ correctly\ predicted\ labels}{Total\ number\ of\ predicted\ labels} \quad (1)$$

$$Recall = \frac{Number\ of\ correctly\ predicted\ labels}{Total\ number\ of\ gold\ standard\ labels} \quad (2)$$

$$F - score = \frac{2(Precision\ *\ Recall)}{Precision\ +\ Recall} \quad (3)$$

In addition, we report accuracy, which simply divides the number of true predictions (both positive and negative class) by the total number of instances of the complete data set. The scores for colour labels occurring in less than 5 training instances are not reported in the results. Due to lack of sufficient training data, these rare labels are never predicted by the classifier (and thus result in 0% performance for all evaluation measures).

A. Experimental results

For the presented colour prediction experiments, LIBSVM was applied with two different kernels. The first experiment was run with the linear kernel of LIBSVM, resulting in an average cross-validation accuracy of 56.05%. For the second experiment, the default (RBF) kernel was optimised by means of a grid search on one training fold, resulting in an optimised c parameter value of 2.0 and an optimised g parameter value of 0.0078125. This second experimental setup yielded the best results, with an average accuracy of 59.88%. We compared these results to two classification baselines: (1) a majority baseline predicting the most frequent class for all instances, being *amber* and (2) a random baseline predicting labels uniformly at random. Table II shows the results of the two baselines and the two variations of our colour prediction system.

TABLE II. ACCURACY SCORES FOR TWO BASELINES AND TWO VERSIONS OF THE COLOUR PREDICTION SYSTEM.

System	Accuracy
Baseline 1	39%
Baseline 2	1%
linear kernel	56%
optimised RBF	60%

Table III presents the detailed results per individual colour class for the first experiment, while Table IV reports the results for the optimized classifier. As can be noticed, the F-scores for the more frequent classes (i.e., *yellow, amber, brown*) are higher in the optimised version, but the performance of *black*, which has 148 training instances, drops considerably (from 35.9% to 20.9% F-score). A second observation that can be made is that colour labels with few training instances are never predicted by the classifier, resulting in an F-score of 0%. Hence, to improve the prediction accuracy for all colour classes, more training data need to be collected to have a more balanced corpus. In future research, we will investigate alternative experimental set-ups, including an ensemble of binary classifiers trained for each colour separately.

In addition, a shallow qualitative analysis revealed that the classifier is often confused between the colour labels *amber, gold, brown* and *black*. The confusion matrix for *amber*, for

TABLE III. CLASSIFICATION SCORES REPORTING THE NUMBER OF INSTANCES (AND CORPUS DISTRIBUTION), PRECISION, RECALL AND F-SCORE ON THE POSITIVE CLASS.

Colour category	Nr of instances (distribution)	Recall	Precision	F-score
very light	9 (0,5%)	0.0	0.0	0.0
gold	683 (32,5%)	59.8	62.7	61.2
amber	819 (39%)	55.7	62.8	59.0
brown	366 (17%)	52.4	49.7	51.1
black	148 (7%)	41.2	31.8	35.9
red/rose	41 (2%)	0.0	0.0	0.0
green	18 (1%)	0.0	0.0	0.0
oak	18 (1%)	0.0	0.0	0.0

TABLE IV. SCORES FOR THE OPTIMIZED CLASSIFIER.

Colour category	Nr of instances (distribution)	Recall	Precision	F-score
very light	9 (0,5%)	0.0	0.0	0.0
gold	683 (32,5%)	64.5	63.8	64.2
amber	819 (39%)	58.6	72.8	64.9
brown	366 (17%)	55.5	57.9	56.7
black	148 (7%)	55.9	12.8	20.9
red/rose	41 (2%)	0.0	0.0	0.0
green	18 (1%)	0.0	0.0	0.0
oak	18 (1%)	0.0	0.0	0.0

instance, illustrates that the classifier often predicts amber beers as *gold* (221 times) and to a lesser extent as *brown* (75 times):

TABLE V. CONFUSION TABLE OF AMBER.

Gold standard label	Predicted label	Nr of instances
amber	amber	514
amber	black	9
amber	brown	75
amber	gold	221

This can be explained by the fact that the different variations of these colours present a continuum, rather than a clear-cut distinction ("pale amber" resembles "deep gold", "amber brown" resembles "brown" and "deep brown" is very similar to "black" in reality). As a result, one can assume that the beer experts are not 100% consistent in naming these similar beer colours and might use similar lexical descriptors in the review text for colour variations that are alike.

B. Most informative lexical descriptors for colour

To gain insight in which n-grams are most characterizing of colour-labelled dataset, we performed Mutual Information (MI) feature scoring. Feature selection filter metrics, such as MI, can be used to characterize both the relevance and redundancy of variables [12]. The mutual information between two random variables is a non-negative value which measures the dependency between the variables. It is equal to zero if and only if two random variables are independent, and higher values mean higher dependency. We used the MI implementation in the Scikit-learn toolkit [13] which relies on nonparametric methods based on entropy estimation from k-nearest neighbors distances as described in [14] ($k = 3$).

We used the MI-score ranking as an approximation of most characterizing features for the target colour labels in which higher MI-scores are more dependent on the 11 colour target classes. The vast majority (81.11%) of the n-grams

TABLE VI. MUTUAL INFORMATION SCORE RANKING TOP 20.

Rank	MI score	ngram
1	1.89E-01	chocolate
2	6.80E-02	stout
3	6.54E-02	of chocolate
4	5.98E-02	coffee
5	4.97E-02	cider
6	4.69E-02	dark
7	3.68E-02	lemon
8	3.58E-02	light-to-medium body
9	3.56E-02	light-to-medium
10	3.43E-02	cherry
11	3.28E-02	porter
12	3.21E-02	dark chocolate
13	3.11E-02	nuts
14	3.03E-02	mocha
15	2.70E-02	apple
16	2.47E-02	chocolate and
17	2.36E-02	medium-to-full body
18	2.35E-02	roasted
19	2.33E-02	cocoa
20	2.21E-02	toffee

over the 90th percentile of scores ($P_{90} = 5.85e\text{-}3$, $n = 254$) pertain unambiguously to smell or odour semantic classes. This shows that the most characteristic n-grams for the colour labels largely pertain to semantic classes of smell and odour. As illustration of this conclusion, Table VI gives an overview of the 20 highest ranked MI descriptors.

IV. LEXICAL DESCRIPTORS FOR TASTE SMELL

In the introduction, it was hypothesized that beer reviewers are subject to odour-taste synaesthesia and by consequence use the same lexical descriptors for taste and smell. To verify this premise, a corpus analysis was performed for the lexical descriptors used in the structured "aroma" and "flavor" labels and a frequency list of all lexical descriptors assigned to both categories was compiled. The full aroma frequency ranking consists of 3121 terms, of which 1993 are unique terms (63.86% of the aroma corpus) and the full flavour ranking contains 2466 terms, of which 1456 unique terms for flavour description (59.04% of the flavour corpus). Figure 2 illustrates, however, a rapid stagnation of both lines starting from the top-500 most frequently used terms. A closer examination reveals that indeed, in the aroma ranking, 1832 terms are only used once, 389 twice and 189 three times in the entire training corpus. In the flavour ranking, 1416 terms are only used once, 310 twice and 160 three times. These are terms that are rather uncommon for describing both aroma and flavour.
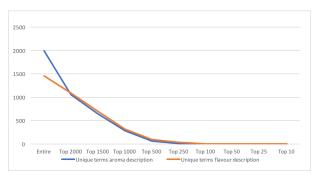

Figure 2. Distribution of unique terms for aroma and flavor labels.

Table VII lists the number of unique *aroma* and *flavour*

terms (and their corresponding percentage of the respective aroma and flavour corpus). The ten most frequently used terms to describe smell properties (i.e., aroma) are also used to describe taste properties, which means that none of these terms are unique for aroma description. Only one term (*roasted nuts*) is unique in the top 25 and even top 50 most frequently used terms in the aroma ranking. In its top 100 most frequently used terms, only two terms (*roasted nuts* and *danish*) are unique for aroma description. For the description of taste properties (i.e., flavour) only one term (*tangy*) in the top 10 and top 25 most frequently used terms is unique. Two terms (*tangy* and *grassy*) are unique in the top 50 of the flavour ranking and in its top 100, five terms (*tangy*, *grassy*, *driven*, *radish sprouts* and *bitter greens*) are solely used for the description of flavour.

TABLE VII. UNIQUE TERMS FOR AROMA AND FLAVOUR DESCRIPTIONS.

	Nr of unique aroma terms	% of corpus	Nr of unique flavour terms	% of corpus
top 10	0	0.000	1	0.041
top 25	1	0.032	1	0.041
top 50	1	0.032	2	0.081
top 100	2	0.064	5	0.203
top 250	12	0.384	35	1.419
top 500	60	1.922	98	3.974

The low number of unique terms for both aroma and flavour descriptions confirm our initial hypothesis of odour-taste synaesthesia [7], which states that people recognise and describe aromas and flavours, respectively smell and taste properties, similarly. The fact that all 10 most frequently used terms for aroma description are present in the top 100 most frequently used terms for flavour description and that only two of the top 10 most frequently used terms for flavour description are unique, shows that indeed many lexical descriptors are used for both aroma and flavour descriptions.

V. CONCLUSIONS

This paper presents preliminary research investigating the sensory descriptors used by expert beer reviewers. The performed lexical analysis confirms the odour-taste synaesthesia hypothesis, as the most frequently used descriptors are shown to be used for describing both aroma and flavour properties of beers.

In addition, we wanted to examine whether expert beer reviewers succeed at describing sensory experiences in a consistent manner. To this goal, we conducted a machine learning experiment aiming at automatically predicting beer properties, being the colour of the beer for the present research. By relying on the fact that perceptions of sight are closely related to perceptions of smell and flavour, consistency of the beer property descriptions has been shown, because the review text was the sole information source used in colour prediction. Our classification experiment showed promising results, with an average accuracy score of about 60%. Analysis of the results, however, revealed that the classifier was only successful at predicting the most frequent colour classes. Moreover, a statistical analysis by means of Mutual Information showed that the most informative review terms for colour prediction largely pertain to the semantic classes of smell and odour.

In future research, we want to collect more data in order to have a more balanced corpus and start collecting data for other languages as well. This way, we can carry out multilingual

analyses of the lexical descriptors that are used to express smell and taste sensations. In addition, we will add more advanced (semantic and syntactic) features to improve the classification accuracy and perform experiments aiming at predicting other beer properties in an automated way.

REFERENCES

[1] S. Levinson and A. Majid, "Differential ineffability and the senses," Mind & Language, vol. 29, no. 4, 2014, pp. 407–427.

[2] I. Croijmans and A. Majid, "Not All Flavor Expertise Is Equal: The Language of Wine and Coffee Experts," PLoS ONE, vol. 11, no. 6, 2016, e0155845. doi:10.1371/journal.pone.0155845.

[3] C. Spence and B. Piqueras-Fiszman, The Perfect Meal: The Multisensory Science of Food and Dining. Oxford, UK: John Wiley and Sons, 2014.

[4] G. Morrot, F. Brochet, and D. Dubourdieu, "The Colors of Odors," Brain and Language, vol. 79, no. 2, 2001, pp. 309–320.

[5] M. Dematte, D. Sanabria, and C. Spence, "Cross-Modal Associations Between Odors and Colors," Chemical Senses, vol. 31, no. 6, 2006, pp. 531–538.

[6] C. Spence, "Multisensory Flavor Perception," Cell, vol. 161, no. 1, 2015, pp. 24–35.

[7] G. Calvert, C. Spence, and B. Stein, The Handbook of Multisensory Processes. Cambridge: MIT Press, 2004.

[8] I. Hendrickx, E. Lefever, I. Croijmans, A. Majid, and A. van den Bosch, "Very quaffable and great fun: applying nlp to wine reviews," in Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2016, pp. 306–312.

[9] L. Bartoshuk, "Comparing sensory experiences across individuals: recent psychophysical advances illuminate genetic variation in taste perception," Chemical Senses, vol. 25, no. 4, 2000, pp. 447–460. [Online]. Available: + http://dx.doi.org/10.1093/chemse/25.4.447

[10] "Tastings.com," 2017, URL: https://www.tastings.com/ [accessed: 2017-01-02].

[11] C.-C. Chang and C.-J. Lin, "LIBSVM: A Library for Support Vector Machines," ACM Transactions on Intelligent Systems and Technology (TIST), vol. 2, no. 3, 2011, pp. 27:1–27:27, ISSN: 2157-6904.

[12] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," Journal of machine learning research, vol. 3, no. Mar, 2003, pp. 1157–1182.

[13] F. Pedregosa et al., "Scikit-learn: Machine learning in Python," Journal of Machine Learning Research, vol. 12, 2011, pp. 2825–2830.

[14] B. C. Ross, "Mutual information between discrete and continuous data sets," PloS one, vol. 9, no. 2, 2014, p. e87357.