

Predicting wine properties with language: a machine learning approach

Ilja Croijmans - Faculty of Social and Behavioral Sciences, Utrecht University, Utrecht, the Netherlands

Iris Hendrickx - Centre for Language Studies, Faculty of Arts, Radboud University, Nijmegen, the Netherlands

Els Lefever - Language and Translation Technology Team (LT3), Department of Translation, Interpreting and Communication, Ghent University, Ghent, Belgium

Asifa Majid - Department of Psychology, University of York, Heslington, York, United Kingdom

Antal van den Bosch - Meertens Institute, Royal Netherlands Academy of Science and Arts, Amsterdam, the Netherlands

Predicting smell and flavor profiles from descriptions or ratings can inform what dimensions underlie perceived olfactory space. However, most endeavors have yielded disappointing results. This is not surprising since odors, and consequently flavors, are difficult to name for most people. Nevertheless, wine experts seem able to name smells and flavors in wine distinctively. This appears to be a contradiction, and wine expert descriptions have frequently met criticism. Are wine reviews written in a consistent, distinctive manner, and if so, what domain-specific terms feature in those reviews? To answer these questions, two analyses involving machine learning were performed on a corpus of 73329 wine reviews. A classification paradigm, predicting color, grape variety, and origin, was trained on wine reviews for which those properties were known, and subsequently predicted those properties for new reviews from a different author. Color and grape variety were predicted with high accuracy. By comparing how words were used in wine reviews compared to other English texts, 146 wine-specific terms were found. These experiments show that wine reviews, written by experts, contain distinctive information about wine. Given the right set of data, it is possible to predict properties of the odor space that underlies olfactory experiences.