FOOD AND LANGUAGE: PRICE-BASED LEXICAL VARIATION IN TURKISH ONLINE RESTAURANT REVIEWS

by

Maria Joy Cupery

A Thesis

Submitted to the Faculty of Purdue University
In Partial Fulfillment of the Requirements for the degree of

Master of Arts

Department of Linguistics
West Lafayette, Indiana
December 2017

THE PURDUE UNIVERSITY GRADUATE SCHOOL

STATEMENT OF COMMITTEE APPROVAL

Dr. Felicia Roberts
Brian Lamb School of Communication
Dr. Margie Berns
Department of English
Dr. Elaine Francis
Department of English

Approved by:
Dr. Mary Niepokuj
Head of the Graduate Program

To my parents, who fostered my love of languages and of Truth

ACKNOWLEDGMENTS

I would like to thank my advisor, Dr. Felicia Roberts, for her patience and wisdom in guiding me through the thousand details of writing a thesis, and Dr. Margie Berns and Dr. Elaine Francis for their questions and insights during the writing process.

I am grateful for the help of Taner Sezer, of Mersin University, both for the extensive and professional work he has done on collecting Turkish corpora, and for his quick assistance when I had questions about using his tools.

And finally, I would like to thank the friends who studied next to me, brought me tea, listened to me talk about this, or convinced me that I needed to get up from the computer. You have distracted me and motivated me in turn, and you all light up my life with your love, humor, wisdom, and charm.

TABLE OF CONTENTS

LIST OF FIGURES

ABSTRACT

Author: Cupery, Maria, J. MA
Institution: Purdue University
Degree Received: December 2017
Title: Food and Language: Price-Based Lexical Variation in Turkish Restaurant Reviews
Major Professor: Felicia Roberts

Previous research has demonstrated price-based lexical variation in American online restaurant reviews and advertising (Freedman and Jurafsky, 2011; Jurafsky, Chahneau, Routledge & Smith 2014). According to the theories of Bourdieu (1986, 1987), and Douglas and Isherwood (2002), this type of variation may be the result of different social classes using differing cultural capital to evaluate their consumer experiences. As reviewers' lexical choices reflect the criteria they apply to that good, the reviews of more expensive restaurants were expected to use significantly different lexicons than the reviews of less expensive restaurants. In comparing lexical, grammatical and etymological choices across four price rating levels in Turkish online restaurant reviews from Yelp, this study extends the research done previously in American corpora to see if price-based lexical variation occurs as expected in a new language and culture.

The results showed multiple examples of price-based variation in the Turkish reviews, and these variations created lexical and etymological patterns that were unique to the Turkish corpora. When reviewing inexpensive restaurants, Turkish reviewers stressed health, homey authenticity and price. As prices increased, the reviewers used different terms for wait staff. The grammatical analysis showed no significant differences, but the etymological analysis showed a growing preference for words with Germanic and Romance roots as the price increased. In conclusion, price-based lexical variation was

measurably present in the reviews, though the patterns differed widely from the ones in the American corpora. The measurable lexical differences supported the hypothesis that review criteria vary by the perceived class ranking of the good. Finally, as socioeconomic class is often a neglected variable in sociolinguistics and corpus studies, this study was also valuable in demonstrating a method for identifying class-based register differences in corpus studies.

CHAPTER 1: INTRODUCTION

Study Overview

Using several different corpus linguistic analyses, this study demonstrates how Turkish online restaurant review writers make different lexical choices when reviewing inexpensive restaurants than when reviewing expensive restaurants. The dataset is 986 Turkish-language reviews of 100 restaurants on Yelp.com. A lexicon-based analysis of the reviews indicate that the reviewers use markedly different patterns than American reviewers; they do not pattern similarly regarding food metaphors, authenticity or uniqueness lexicons. When reviewing inexpensive restaurants, Turkish reviewers stress qualities like health, homey authenticity and price. As prices increase, the reviewers use different terms for wait staff. A grammatical analysis showed no significant differences across corpora, and an etymological analysis showed a growing preference for words with European roots as the price increased. With the theoretical framework of Bourdieu's social and cultural capital, and Douglas and Isherwood's culture making through consumption, it appears that statistically measurable price-based lexical variation should appear in the Turkish context. Restaurants at different price levels are judged using different criteria, and these differences are observable through lexical analysis, supporting the theory that consumers review goods differently according to the price rating. The tastes that lead to a value judgment can vary depending on the perceived class of the good offered.

Previous corpus analyses of price-based lexical variation have concentrated on English-language online restaurant reviews (Jurafsky et al., 2014) and the language used in potato chip advertising (Freedman & Jurafsky, 2011). In the first paper, Jurafsky et al (2014) found that the metaphors used for food in expensive restaurants are strikingly different from the metaphors used for food in inexpensive restaurants. In the second paper, Freedman and Jurafsky (2011) discovered that the language and lexicons used by advertisers to sell potato chips varied by the price of the bag – the more expensive the chips, the longer the words and the more that qualities like uniqueness and health were emphasized. These results are examples of statistically demonstrable lexical differences between different price levels. This study extends the previous studies to a new cultural and linguistic setting. Since previous research in this field has been mostly descriptive, this study also considers the implications of the data for the theories of cultural capital and culture making through consumption, and the value of using corpus linguistics to conduct socioeconomic research. Though the trends in the Turkish reviews differ from the ones in the American reviews, the measurable lexical differences found in this study support the idea that reviewing a consumer experience is not just about the characteristics of the experience; it is also a form of culture-making where the reviewer establishes their cultural capital by critiquing the experience. When a reviewer writes a review, they are practicing and displaying their ability to judge an experience based on their understanding of what the experience should be. Others read the reviews to see if the restaurant will live up to these social expectations. Through writing and reading online reviews, consumers enforce and challenge each other’s cultural opinions about what is and is not appropriate in a restaurant of a certain price. Thus, cultural standards for taste and judgment are both refined and maintained in the constant cycle of reviews.

Justification for Research

Bourdieu (1986, 1987), and Douglas and Isherwood (2002) describe how class and social power are tightly interconnected with how humans evaluate their experiences; however, given the grand scale of their theories, it can be difficult to connect the theories to the specifics of sociolinguistic data. The advent of online reviews and the tools of corpus linguistics give us the ability to measure written text in ways that were previously impossible. This pairing gives us both access to a large corpus of informal text written by a wide variety of people, and the capacity to look for patterns in these large corpora. So far, these corpora have mostly been mined by companies who are seeking to improve their products and understand their customers. However, less work has been done on using this data to extend, question or defend sociolinguistic theories.

These reviews are suitable data for sociolinguistic study because they were written informally and for the public. There has been a push in sociolinguistics to study language that is "socially realistic," that is, "based not… on samples elicited from language users in laboratory type settings, but on samples collected, *in situ*, from within existing speech communities," (Block, 2013, p. 74). This is natural language that is open to analysis; unlike much of corpus data, there is no issue with the observer skewing the data (as in studies were participants can be influenced by the presence or questions of the researcher), or with the accessibility of the data to outside researchers (as there is with recording private conversation).

The existing research on online reviews provides intimations of what this research could mean for sociolinguistic theory. For example, Freedman and Jurafsky (2011) found that the advertising for both cheap and expensive potato chips touted authenticity as an important value; but for cheaper chips, they advertised a homey, "mom and pop" authenticity, while for expensive ones there was more talk of following exotic recipes and the distant locales of the chips' ingredients. Advertising intentionally targets the perceived desires of the intended audience; but price-based variation has been found in the discourse of consumers, as well as in advertising copy. Jurafsky et al. (2014) also discovered that we tend to talk about expensive food using sensual and sexual metaphors, while metaphors of drug addiction are more common for inexpensive food. Their findings show that people informally reviewing products chose different language to describe inexpensive and expensive experiences. However, these two papers leave several questions unanswered. First, how common are these trends in lexical choice? Do they hold across cultures and languages, or will each language and culture have its own price-based lexical differences in evaluating a consumer experience? And secondly, how does this data reshape our understanding of the theories about the interaction of cultural capital and language? Does the data fit in with the models proposed by Bourdieu, Douglas and Isherwood?

Research Questions and Hypotheses

My research seeks to answer the following questions: Are there measurable lexical differences between the online reviews of inexpensive restaurants and expensive restaurants in Turkish? Do these differences follow the trends in the American studies, or do they pattern along different lexical divisions? How does this data extend the social theory of culture-making through consumption?

This thesis will propose the following. The basis for these hypotheses will be developed in the literature review that follows.

H1: The lexical choices and structures (such as word origins, grammatical patterns and lexicons) in Turkish online restaurant reviews from Yelp will vary with the price rating of the restaurant.

H2: It will be possible to measure these price-based lexical differences using corpus analysis tools.

CHAPTER 2: LITERATURE REVIEW

These research questions arise from the goal of connecting larger social theories with the data and analyses available through corpus linguistics. Thus, it is necessary to understand the social theories that inform this study. Bourdieu’s two books discuss how social class shapes taste (1986, 1987); Douglas and Isherwood (2002) extend that to discuss goods as indicators of social class. Combining these two views, it is possible, in modern capitalist societies, for different individuals from a variety of socioeconomic backgrounds to consume more than one social class of good. The same person may review an inexpensive restaurant and an expensive one; but, in light of Bourdieu’s and Douglas & Isherwood’s theories, they should be judging these experiences on different criteria. When they eat at a lower-class restaurant, they will be basing their review on their expectation of what constitutes a quality experience for a restaurant of that social class; for example, they will not be expecting valet service or cloth napkins when they go to a local hamburger joint; and the absence of those qualities will not decrease their rating. Because modern democratic societies allow their citizens to sample a wide variety of social experiences, if they have the funds to pay for them, citizens of these societies should adapt the criteria they are using to make social judgments according to the experience they believe they should be getting. The price ranking system in online restaurant reviews provides a clear way of setting expectations dependent on the average price of a meal at that venue. Thus, the work of Bourdieu, Douglas and Isherwood provides some background as to why I expect to see price-based variation, as in Hypothesis 1.

This argument begins with the two interrelated claims in Bourdieu's *The Forms of Capital* and *Distinction: A Social Critique of the Judgement of Taste* (1987): that material (economic) capital is closely connected with social and cultural capital, and that this cultural capital is expressed in our value judgements, in our tastes, or in how we judge experiences and products. Douglas and Isherwood's *The World of Goods* (2002) then argues that not only are our values and judgements class-based; the experiences and objects we consume, and the way we talk about these experiences and objects, are important ways of establishing our class standing and increasing our cultural capital. In this literature review, I will discuss how these theories account for the modern phenomenon of online restaurant reviews and predict price-based lexical variation. Then, I will discuss the results of previous research done on American online restaurant reviews. This work is helpful both in that it demonstrates the existence of price-based variation in another corpus, and in that it gives some indication of patterns to look for in the Turkish review corpus. Thus, the work on English provides the foundation for Hypothesis 1. Finally, I will discuss the tools that make Hypothesis 2 possible. It would be impossible to find what characteristics were unique to price-based variation if it were not possible to assemble a corpus of Turkish reviews to contrast with the data from the American reviews. At the same time, there are several analysis tools, such as a part-of-speech tagger and frequency list generators, that have been designed specifically for the Turkish language and which will be essential for measuring the differences.

Price-Based Taste Variation: Theoretical Grounding in Bourdieu's *The Forms of Capital*

In 1986, Pierre Bourdieu argued that the forms of capital go beyond property and money; cultural and social capital are also forms of collecting and establishing social power. By cultural capital, he meant the abilities, education, and properties that give a person high cultural standing. It is easiest to acquire these when one's family has enough economic capital to invest in their children through actions like hiring tutors, going to artistic events, and paying for quality education. Thus, while the results (an appreciation of fine music or gourmet food, a quick intellect, a well-developed physique) are often seen as the results of individual talents and dedication, they are often made possible by the economic standing of the individual or his or her family. Cultural capital comes in three forms. Embodied cultural capital is the knowledge of what counts as culture; for example, an appreciation of art. For our study, the knowledge of what criteria can be used to evaluate restaurants is a form of embodied cultural capital. Objectified cultural capital is the objects (e.g. art pieces, new technology) that give social standing, and institutionalized cultural capital is the formal recognition of that capital, for example a university degree, or a professional title like 'certified public accountant.'

Social capital, according to Bourdieu, is embodied in the social connections and mutually dependent relationships that provide social and economic benefits to their participants. These relationships are developed and maintained through different social groups like families, schools, local communities, work communities, clubs, and so forth. By continuously reaffirming each other's worth and belonging through activities, exchanges and interactions, these groups have a multiplier effect: their participants are not only reaffirmed in what social capital they possess; they also increase the social capital of those around them. Such social capital can be institutionalized in the titles of nobility. A modern example would be universities, where students and alumni benefit from and work to increase their institution's prestige in business, politics, and society.

Bourdieu's theory of capital allows us to observe culture and social relationships through the eyes of economic theory. When people participate in cultural events or join social networks, they are not doing so out of a merely disinterested enjoyment of that group; they are also enacting their class roles and seeking to affirm and increase their cultural and social capital. This theory, in turn, can explain how individuals decide what cultural and social experiences are valuable; Bourdieu examined the consequences of his theory of capital on issues of taste and aesthetic judgement in his 1979 book, *Distinction: A Social Critique of the Judgement of Taste*.

Because obtaining social and cultural capital requires free time, wealth and supportive families, and because these processes usually begin at a young age, those with economic power tend to dominate the aesthetic values of a culture. They not only control society's monetary wealth; they also determine what good taste means, and, by extension, which experiences and objects are aesthetically valuable and which are cheap or tawdry. Often, the working class are unable to properly understand or appreciate the upper class's aesthetics because they lack the access to the social and cultural capital, such as education, the vocabulary required to discuss an experience, the free time to learn a skill, or the social connections to join in these experiences. Middle class and working class individuals will tend to develop their own aesthetics, based on the activities and values they find most important, but the upper class may not appreciate a working-class aesthetic because they do not understand the traditions or skills necessary to create that aesthetic. Generally, societies often treat working class aesthetics as inherently less valuable than the upper class aesthetics.

These large-scale theories affect our understanding of every social arena from education to cinema; Bourdieu himself understood the value of testing his theories against specific data sets to see how they worked out in the particularities of his culture. Several sections of *Distinction* are devoted to statistical analyses of the surveys on personal taste he supervised in France in the 1960s. He grouped his respondents by their cultural capital and social origin; these two factors were measured by their educational standing and their father's occupation. He then searched for correlations between these two factors and their opinions on both "legitimate" and "personal" matters of taste. "Legitimate" matters were categories such as art, music, politics and cinema, and he found his participants' opinions on these matters most closely correlated with their educational standing. "Personal" matters had to do with more daily choices: furniture, food, and clothing. In this category, the participants' aesthetic choices were closely linked with their social origin. Bourdieu found that three major groups emerged in the responses, with strong distinctions between the working class and the upper class, and with the middle class falling in between. Working class participants tended to stress cleanliness, value for money, and simplicity in cuisine; middle class participants were more concerned with individualization in clothing and in creating welcoming homes, and were split on food, some preferring simple foods, others exotic and original cuisines; and upper class participants were most interested in harmonious furnishings, individualized clothing and in exotic, original foods.

For the current study, the most salient conclusion from Bourdieu's work is that the value of an experience or an object was not simply determined by its inherent value. All food experiences are not judged on the same scale of bad to good; depending on the class it is associated with, it will be judged according to different attributes. For example, in Bourdieu's findings, a working class meal will be considered a good meal if it is simple, filling, affordable, and served quickly. An upper class meal will be considered a good meal if it contains exotic ingredients, original presentation, and good service from the wait staff. Thus, the language we use to talk about an inexpensive food experience and an expensive food experience should correlate with the different values we hold for different levels of food, because what counts as 'cultural capital' varies at different price ranges.

Some fifty years have passed since Bourdieu conducted his first surveys in France, and 21st century communities differ in many ways. However, Bourdieu's analysis of social and cultural capital is still a relevant explanation when looking at how modern individuals make value judgements. Bourdieu looked for patterns by gathering information about his participants' demographics and correlating that with their judgements. However, since we now have access to large quantities of data where people are making value judgments on experiences with clear economic rankings (assigned price levels), we can study their language use to see how it differs based on the price level. If Bourdieu's theories still hold, then rating a dining experience is not just about the actual food consumed; it is also a chance to display one's cultural capital through the process of evaluating the experience. A reviewer must draw on his or her knowledge of taste to decide if a restaurant meets his or her expectations. Depending on the class category of the experience being reviewed, different criteria will be used to evaluate.

How We Evaluate Goods: Douglas and Isherwood's Explanation of Taste

While Bourdieu provides a framework where experiences, abilities, tastes and goods are effective modes of communicating cultural capital, Douglas and Isherwood concentrate on how goods, in particular, are used to demonstrate and increase social power. By "goods," they are referring to both objects and experiences; anything that can be purchased. Their theories overlap with Bourdieu's, but provide a more specific framework in which to look at how and why people evaluate the experience of eating in a restaurant.

According to Douglas and Isherwood (2002), goods are primarily "needed for making visible and stable the categories of culture… all material possessions carry social meanings" (p. 38). Specifically, in the context of eating, "Food is a medium for discriminating values, and the more numerous the discriminated ranks, the more varieties of food will be needed… The choice of goods continuously creates certain patterns of discrimination, overlaying and reinforcing" (p. 44). It is not merely the quality of food that demonstrates the rank of the consumer; it is also the process of discriminating which foods are quality and which ones are not that allows one to create, reinforce or exercise cultural capital.

Value judgements on these goods are never made in a vacuum; as Douglas and Isherwood (2002) argue, creating value judgments on a product or experience necessitates a group; judgements are more valid when supported by many people: "Goods are endowed with value by the agreement of fellow consumers. They come together to grade events, upholding old judgments or reversing them" (p. 51). Though online review culture was not in existence when they were writing, it is now a common venue for consumers to demonstrate their abilities to make judgments and grade events.

Because the value of a product is dependent on the reviews and opinions of the society, that value is also constantly shifting. Goods are not inherently as valuable as the social relations they represent: Thus, for the individual making a value judgment, "he continually needs to maintain his synthesis or adapt it in the light of rival views. The risk for him comes from an alien view that is more comprehensive in scope than his own" (p. 53). This theory can be extended to explain the popularity of online reviews. By writing and reading online reviews, individuals have immediate access to a large database of other peoples' judgments and judgment criteria, and this allows them to refine their own opinions through the group's opinions.

Baron and Isherwood also highlight the connection between the economic theory of supply and demand and the anthropological theory that goods are used to communicate status. This affects how we judge quality: the value of a good increases as it becomes rarer and is seen or experienced less often. They apply the term "periodicity of consumption" specifically to food in to illustrate this point. Food that is consumed often, such as potatoes, bread or rice, does not mark social standing and is cheap. Food that is rarely consumed, such as imported foods, wine or caviar, is more expensive and does indicate higher social standing. Thus, the value of the food does not come merely from its taste; the more economically difficult it is to obtain a food, the more it confers prestige to its consumer, and the more it is assumed that that food must be consumed for pleasure, not because of its ability to provide sustenance. If their theory is valid, this definition of value should apply in the matter of what criteria are manifested in online restaurant reviews. Less expensive restaurants should be rated on their ability to provide decent sustenance well, whereas more expensive restaurants should be rated on their ability to provide pleasure or a unique experience or rare product.

Finally, Baron and Isherwood (2002) state that, "Goods are now to be seen as the medium… Attention is directed to the flow of exchanges, the goods only marking out the pattern" (p. 152). As mentioned in the introduction, one helpful effect of their concentrating on goods as an indicator of social class is that it allows one class to participate in the consumption of more than one social class of good. The same person may review a more and a less expensive restaurant; but the critiques they offer should vary. Thanks to Yelp's price ranking system, the different levels of the restaurants are clearly delineated. The reviewers are judging the restaurants based on what kind of service they would expect for a certain amount of Turkish lira; and if Bourdieu's theory of taste holds true, the reviewers' lexical choices should vary systematically with the price levels, as stated in the hypotheses.

Jurafsky’s Work in the American Context: Grounding the Study Hypotheses

Having set up the theoretical context for the study of price-based lexical variation in discussing food, it is possible to now turn to the research that has been done on this area and has been mentioned in the introduction. The first hypothesis is grounded in studies that have found evidence of lexical patterns that vary by price rating in English-language Yelp reviews and food advertising. While the lexical patterns might change, it is possible to assume that such patterns will also exist in other languages, including Turkish. The first article, “Narrative framing of consumer sentiment in online restaurant reviews,” (Jurafsky et al., 2014) demonstrates that in English, reviewers use shorter words and the language of craving and drug metaphors to talk about inexpensive restaurants, while they use longer words, more complex sentences and sensual or sexual language to describe expensive restaurants. The second article, “Authenticity in America: Class distinctions in potato chip advertising” (Freedman & Jurafsky, 2011) makes the further discovery that advertising targeting the wealthy used language of distinction, health and exoticism, while advertising targeting the lower and middle classes used language of tradition, heritage and family values.

For the first article, the authors created a database of over 850,000 reviews from restaurants in seven major cities in the United States. They studied the words that were most strongly associated with negative and positive reviews, and with each of the four levels of price rating. Their major findings were that the 1-star reviews featured more negative adjectives, more first personal plural pronouns, more past tense verbs and narrative sequencers and more third person pronouns that the other reviews. This cluster of features had been found to indicate trauma-processing narratives in previous research. (Biber 1995, Stone and Pennebaker, 2002). Thus, they were able to make a meaningful connection between grammatical features and content. Second, they studied the association of metaphors of addiction with price rating; the lower the price rating on a restaurant, the more likely it was to include the lexicon of addiction, with words such as “addict,” “crave,” “binge,” “drug,” and “crack.” Finally, they tracked two lexical features that increased for the more expensive restaurants: word length and the use of the lexicon of sensuality. Expensive restaurants tended to have reviews with higher word counts, and the writers employed longer words in their reviews. They also made more use of the lexicon of sensuality, employing terms like “temptation,” “sinful,” “seductive,” “sexy,” “romantic,” etc. Reviewers primarily used these words to describe either the desserts or the ambiance of a restaurant.

Jurafsky et al. (2014) were mostly concerned with how online data could be used to critique and extend previous work on narrative framing and computational extraction of social meaning. However, their study has clear connections with the work of Bourdieu and Douglas and Isherwood. When writers are using different lexicons to evaluate similar experiences, it indicates that they are using different paradigms to arrive at their judgments. By their shifting lexical choices, writers are switching paradigms depending on the price of the food. There are two possible, overlapping explanations of this. One is that each price level attracts a unique group of writers from a correlating socioeconomic group; working class people are reviewing the inexpensive restaurants while upper class people are reviewing the expensive restaurants. In this case, each group is creating taste judgments according to its group’s values. However, it is also possible that the same people are writing reviews for both more and less expensive restaurants, and that they are switching their criteria – and their lexicons – accordingly. In the case of the online restaurant reviews, it is a combination of both elements. As someone’s socioeconomic status increases, they are more likely to go to and to review more expensive restaurants, and thus the reviews would reflect the language of the class level that most frequents that level of restaurant. However, people also go to and review restaurants across a range of price levels, and change their lexical choices according to what they find appropriate for that level. While we do not know if the reviewers themselves changed by price level, or if the same reviewers were reviewing across price levels and changing their vocabulary, Jurafsky et al. found a distinct change in lexicons used according to the price level of the restaurant. As Bourdieu suggested, each class has its own aesthetic and makes its judgments accordingly, whether the people are entering in and out of the classes or are permanently using only one set of values to judge all their experiences. As Douglas and Isherwood explained, purchasing and evaluating a restaurant experience is part of how individuals prove their social standing.

The second paper, by Freedman and Jurafsky, is about advertising discourse, not restaurant reviews, but it is nonetheless helpful in learning how to create lexicons that are relevant to restaurant review data. In 2011, Jurafsky and Freedman looked at how the advertising on potato chip bags varied by the price of the potato chips. They found that the advertising language for more expensive chips had longer words, more mentions of healthiness, more use of distinction language (words like “only,” and “unique,” and more superlatives), and lexicons of exoticism. The advertising for cheaper chips employed shorter words, fewer words in general, and the lexicons of authenticity, tradition and value.

Again, in the light of Bourdieu’s theory of taste, these variations in lexicon are not surprising. Advertisers are aware that customers of different socioeconomic classes are judging a product based on different values, so they tailor their descriptions accordingly. The wealthier customers place more value on health, uniqueness and exotic experiences. Those buying the less expensive chips place more importance on thriftiness, simplicity, following tradition and remaining faithful to their culture. As Jurafsky pointed out, the ingredients for all of the chips were nearly identical; what changed was not the product, but the way in which the customers were expected to make their taste judgements about that product.

The lexical choices used in advertising will not precisely align with the lexical choices of reviewers because the concerns are different. As advertising is concerned with selling its product, the lexical choices will tend to be overwhelmingly positive. When writing reviews, the writers discuss how the product failed to meet their expectations as well as the ways in which they found it satisfactory. However, Jurafsky's lexical categories provide an excellent starting point for analyzing reviews when the lexemes in the lexicon are extended to include both positive and negative words. For example, it will be important to search for the unhealthiness lexicon (lexemes like "fat," "oily," "sugary") as well as the health lexicon developed by Freedman and Jurafsky.

Corpus Linguistics in the Turkish Context

While social theories of taste and consumption provide a helpful background for analyzing restaurant reviews, the work in corpus linguistics reported so far has been limited to one cultural and linguistic context. All the restaurant reviews that Jurafsky et al. looked at were in English and were based on American restaurants. While their data supports the work of Bourdieu, Douglas and Isherwood, the theories must be tested against a broader range of cultural settings. To test a different culture, it is also helpful to test a different language; other English-language-based countries, such as the UK or

Australia, tend to have many cultural similarities with the United States. Thus, for this project, it is most helpful to pick a non-Western country with a non-Indo-European language. Since I spent many years in Turkey and am fluent in Turkish, I chose Turkey as the location for the restaurants and Turkish as the language to analyze.

Furthermore, the restaurants are all from the city of Istanbul. Istanbul is home to a diverse range of Turks of a variety of socioeconomic backgrounds, and this is reflected in the restaurant culture. With over 15 million inhabitants, it is the largest city in Turkey. At the same time, Turkey in general is highly technologically literate and both younger and older people are active on social media, which means that the Western models of online restaurant reviews have been accepted and widely adopted. The main source of online reviews is yelp.com.tr. Currently, the website has over 17,000 restaurants listed for Istanbul, with anywhere between 1 and 70 reviews for each restaurant.

One further benefit of selecting Turkish as a language is that there has already been substantial work done in creating usable corpus linguistics tools for analysis of Turkish. These corpora and these tools are what make Hypothesis 2 possible. There are several available corpora in Turkish: The Turkish National Corpora, the METU Turkish Corpus, the BOUN Corpus, and multiple general and specialized corpora in the TSCorpus collection. There have also been several programs created for tagging Turkish corpora and disambiguating morphologically ambiguous lexemes (Durrant, 2013; Sak, Güngör & Saraçlar, 2011; Sezer & Sezer, 2015).

While the large number of available reviews and the benefits of testing the social theories in a new cultural setting make Turkish reviews a productive field for study, the cultural and linguistic milieu require some changes in methodology and in expected results for this study. The syntax and morphology of Turkish differs from English syntax and morphology in ways that affect how corpus analysis is conducted. Turkish is SOV and is highly agglutinative. This means the default sentence order is subject-object-verb (though that order is highly flexible), and that words are often composed of a stem followed by several morphemes that provide semantic and syntactic information. These morphemes include a wide variety of noun cases, and verb endings that inflect for aspect, tense, mode and person. It also uses bleached verbs to make verbs out of nouns, and adjectives and adverbs are often formed from or converted into nouns or verbs. The morphology after the stem is subject to vowel harmony, both front/back and rounded/unrounded.

To give a brief example, consider the following line from one of the reviews:

(1) Dur-ma-dan gid-ebil-ir-im Çiya zincir-ler-i-ne
stop-NEG-ADV go-can-PRS-1s Çiya chain-PL-of-ACC
'I can go continuously to the Çiya chains'

The first word is formed from the verb root for 'stop,' with negating and adverbial morphemes added to create an adverbial expression that literally means 'without stopping.' The second word is a verb followed by an aspect morpheme, a tense morpheme, and a person morpheme. The noun also has three morphemes, one marking plurality, the next its relation to 'Çiya', and the last marking it as the object of the verb. The vowels in all the added morphology of all three words are determined by the vowel of the stem, according to the rules of vowel harmony.

This means that while the corpus analysis can still be done on word stems, the way English corpus analysis is done, a lexical stem must not be confused with a part of speech. Counting the number of times that a stem is used would not give information as to whether that stem appeared as part of a verb, as a noun, or as an adjective. So, part-of-speech analyses will have to be done by part-of-speech taggers and not word stem or lexeme searches. A second issue is the phonological rules that dictate vowel harmony and other word-internal changes; patterns related to suffixes will have to include all the possible spelling variations and be distinguishable from the stems to which they are attached. For example, to measure the use of the future tense across the corpora, I would have to search for both “-ecek” and “acak” morphemes to account for vowel harmony, and then “-eceğ” and “-acağ” to account for a phonological rule that eliminates the /k/ phoneme between vowels.

Secondly, as the cultural milieu is different, the basis of lexical distinctions between classes may vary as well. Jurafsky’s findings make sense in American culture; the upper class in the United States does tend to concentrate on their health, seek out exotic or international experiences, and use longer, Latinate words as a sign of their education. The working class is typically associated with a more substantial desire for fast, cheap food and for traditional flavors. Addiction is not considered desirable, and thus is only used for inexpensive food, while in American culture it is permissible to use sexual language to discuss food.

However, a less diverse, majority Muslim, Middle Eastern culture may demonstrate socioeconomic variation in other values. In English, the educated Latinate words tend to contain more characters than the Germanic words that are considered more basic; thus, it is expected that reviews of more expensive restaurants will contain words that are measurably longer. For Turkish, however, the language has a strong Turkic base in structure and vocabulary, but it has borrowed extensively from Arabic, Persian, French and English, among other languages. Very little research on how the connotations of the words differ depending on their origin exists. There is preliminary research on the use of English-origin words as a marker of education and Western sympathies (Selvi, 2011) and in my experience, use of obviously Arabic-derived words can either be a sign of classical education or of more conservatively religious opinions. While word length does not correlate conveniently with origin as it does in English, word origin is still a sociolinguistically relevant factor.

The common metaphors for food may also show culturally dependent variations. It is still taboo to use sex as a metaphor in the Turkish context, and addiction may not provide the positive association it has in the American context. Thus, it is an open question as to what kinds of metaphors Turks will use in place of sensual or addiction language; there has been very little work done on metaphors in Turkish, and none on metaphors for food experiences.

The cultural analyses are only possible because of previous work done on corpus linguistics in Turkish. The best set of corpus linguistics tools for Turkish available are developed by Taner Sezer, and available at his website, tscorpus.com. The Natural Language Processing Toolkit there includes a Part-of-Speech Tokenizer that identifies the parts of speech for each individual word stem and agglutinated morpheme, and a Frequency Calculator that calculates the number of tokens, the number of unique tokens, and the frequencies for each token.

Summary

The yelp.com.tr reviews allow us to test the theories of cultural capital and culture making through consumption in a unique way. Because the reviews are written informally, by a wide range of individuals, they create publicly available data that represents a cross-section of voices from Turkish society. If, as Douglas and Isherwood argue, culture is made through the consumption and evaluation of goods, then restaurant reviews should provide an excellent venue for discovering what values shape the judgements in 21st century Turkish culture and how the reviews reinforce or test the judgment-making process. If Bourdieu’s theory of taste makes the correct predictions, then we will see different value paradigms for the dining experience, correlating with the price-level of the restaurant. These different value paradigms should be expressed in different lexical choices, similar to the results found by Jurafsky et al. Specifically, the most common lexicons for each price-level (the lexicon of health, the lexicon of uniqueness, or other culture-specific lexicon)– should show a gradation between the values and aesthetics associated with the upper class and values and aesthetics associated with the working class. The word origin of borrowed words – whether more of the words are borrowed from English, Arabic, Farsi, or another language – may also vary by price level.

The research on American restaurant reviews and American advertising gives us an indication of what categories to look for in the Turkish data. At the same time, the tools developed specifically for Turkish make it possible to tweak the methodologies to support both smaller scale analysis and analysis of a language with morphological and cultural differences.

CHAPTER 3: RESEARCH DESIGN

Gathering Data from Yelp

Why Yelp Restaurant Reviews?

Yelp.com.tr allows their users to rank a variety of consumer experiences, such as cafes, barbershops, pastry shops, cinemas and bars. However, this study chose to concentrate on the reviews of restaurants for two reasons. First, it allows us to draw clear parallels with the previous work done on restaurants in America. Second, it is one of the few selections that is well represented at a variety of price levels by a variety of consumers. For example, there is much less price variation for cafes and pastry shops; and bars will tend to be in the higher price range in Istanbul. At the same time, concentrating on these venues would give undue representation to certain parts of Istanbul's demographics; conservative Muslims would tend to avoid bars, and cafes tend to attract younger and more Western populations. However, restaurants are frequented by religious and secular, younger and older, alike, so their reviewers are a more diverse audience. However, yelp.com.tr lists over 17,000 restaurants, so to create manageable and representative corpora, the data set had to be narrowed.

The possible price ratings on Yelp.com.tr for the restaurants are ₺, ₺₺, ₺₺₺, or ₺₺₺₺, based on the symbol for the Turkish lira. A rating of one lira sign means the meal costs below 20₺; a rating of two means between 21-50₺; a rating of three means between 51-120₺; and a rating of four means above 120₺ per meal. To develop a representative sample of price ratings, the data set consisted of four separate corpora: each corpus contained all the available Turkish-language reviews for 25 restaurants at one of the four price ratings. For ease of reference, the corpora will be referred to as C1 through C4, with C1 being the corpus of the least expensive restaurants, and C4 the most expensive.

To be included as one of the 25 restaurants, the restaurant had to have a minimum rating of 3 stars and at least three reviews in Turkish. Because online reviews are highly skewed towards positive reviews (Jurafsky et al. 2014), restaurants with very poor overall ratings are uncommon and, as they are outliers, could potentially cause the results for their corpus to be skewed in comparison with the other corpora. Ensuring that each restaurant had at least three reviews also guaranteed that no one review or opinion dominates the restaurant's rating. Since the study is concerned with differences in how large groups tend to evaluate restaurants based on their price level, it is important that the ratings are not based on one individual's preferences, but on a group consensus. Again, this prevented outliers from skewing the language in one of the corpora.

Because the low- and mid-level restaurants tend to have more reviews than the most expensive restaurants, the corpora for the lower price ratings are larger than the corpora for the higher price ratings; in fact, at 25,493 words, C1 is substantially larger than C4, which has 15,328 words. However, because the data from the different corpora were compared as words per 10,000 words, not as raw frequencies, this did not prevent making comparisions. It ensured that each corpus covers the same variety of resturants, and that each restaurant was represented through the full range of opinions. Creating corprora that each contained the same number of reveiws as the other would have led to skewed corpora where C4 would contain a much larger number of restaurants while the data in C1 and C2 would be based on a few venues.

Setting Parameters for Reviews

In setting the parameters for the search, there are two different Turkish words that are used for restaurants – "lokanta" and "restoran." The first one was borrowed from Italian, the second from French. However, the first one is an older borrowing, and feels more nativized because it follows the rules of vowel harmony. The second one is more recent, and it feels like a borrowed word because it breaks vowel harmony; the “e” is a front vowel and “o” and “a” are back vowels. Thus, connotatively, “restoran” tends to imply fancier or more exotic restaurants, and “lokanta” refers to homier venues. However, there is extensive overlap in their usage, and the website does not seem to differentiate widely in their use. Searching for one or the other within a price rating leads to a list of the same restaurants; the only difference is that ones with “lokanta” in their names occur earlier on in the list when the search is “lokanta,” and the ones with “restoran” in their names appear first when the search is “restoran.” For this study, the word “lokanta” was chosen to set the parameters, as this tended to lead to more restaurants that were targeting Turkish audiences, and meant that there were fewer foreign-language reviews to delete when collecting and processing the reviews.

Inclusion Criteria for Reviews

The next step was determining which restaurants to include in the 25 for each corpus. On Yelp, it is possible to search by the regions of Istanbul. To ensure that the data covered as wide a cultural range as possible, it is important to ensure that as wide a range of locations as possible is represented. As a result, in order to build the corpora, the reviews included restaurants from a minimum of five neighborhoods for each price rating, and no more than ten restaurants from any one neighborhood. In order to ensure the diversity of locations, for each corpora, the searches included the neighborhoods of Kadıköy, Üsküdar, Fatih, Beşiktaş and Şişli. These five neighborhoods were chosen because they have different sociocultural atmospheres. Kadıköy, on the Asian side, is one of the most modern, westernized neighborhoods, drawing younger Turks interested in art and the counterculture. Üsküdar, also on the Asian side, is considered more traditional and conservative and is both a business and shopping center. Fatih is on the European side, a mix of tourism and local business. Beşiktaş is also on the European side and is a center for local business and shopping. Şişli is one of the vast suburbs that has sprung up around the old city, home to some of the thousands of recent immigrants to the city. Because these neighborhoods each have a distinct profile, they gathered a wider sample of writers than concentrating on any one neighborhood. The five neighborhoods do not all together have 25 restaurants for each price rating; but when the search includes a specific location, Yelp automatically broadens the location out to nearby neighborhoods to find more restaurants in that price rating. Thus, while each corpus includes at least these five neighborhoods, some spread farther out to include nearby neighborhoods as well.

Variables Included with Each Review

Each review comes with an assortment of information that was processed and included with the review. Each review was listed along with the name, neighborhood and price rating of the restaurant, the author’s projected gender, the number of stars they gave the restaurant, and the feedback the review received (readers can rate each review as ‘helpful’ ‘funny’ or ‘cool’). The authors’ projected gender is apparent because the reviewers must use a first name and a photo to post reviews. Some reviewers belong to the “Yelp Elite,” which means that they have posted reviews often enough and had their reviews voted helpful often enough to receive the special designation. Yelp also checks each review for fairness before they “recommend” it, so some restaurants have reviews that have not yet been recommended. Each review, therefore, included information as to whether the review was written by a Yelp Elite or had been recommended.

The restaurant names were used to identify which keywords referred to a specific restaurant and the neighborhood information was used to ensure that the corpus was balanced across several districts. The information about gender, feedback and Yelp Elite made it possible to check for confounds, particularly to see if any one reviewer or demographic occurred too frequently. However, in the final corpus, no one reviewer dominated any of the price levels, and as these variables did not correlate significantly with the lexicon and etymology results, they were not considered confounds.

Therefore, the first step for this project was to assemble four corpora composed of the reviews for 25 restaurants at each price level.

Once the data was assembled on an Excel spreadsheet, the next step was to put all the reviews through a parts-of-speech tagger to create corpora that have searchable morphological tags. This was done using the POS Tagger on TSCorpus.com. One benefit of using this tagger is that it has been trained on a corpus of Turkish tweets from Twitter, so that it has categories for internet abbreviations, emoticons, internet slang, common English borrowings and abbreviations. As these appeared in the reviews, the tagger was equipped to handle them.

Creating and Comparing Lexicons

Once the dataset was organized and cleaned, the next step was to create the lexicons for each category; that is, word lists clustered around that category's topic. I took a two-pronged approach to this, one part brainstorming and one part corpus data, to create lexicons for "sensuality," "addiction," "health," "uniqueness," "exoticism," "tradition" and "local authenticity". To use "health" as an example, I listed all the words I could think of in Turkish related to healthy eating, and then passed these lists on to native speakers and asked for their contributions. Once I had incorporated their suggestions into the lexicons, I expanded the lexicons by taking their lexemes, entering them into AntConc and looking at the list of common collocations. When their collocations referred to the same semantic concepts, they were added to the concept. When certain lexemes led to noise in the data, they were either eliminated or changed to forms that only referred to the subject under consideration. For example, “fayda” (“benefit”) was originally included in the healthiness lexicon, as it is often used to refer to the health benefits of food. However, a look at the lexeme’s KWIC data (Keyword in Context; i.e. the 10 words before and after the keyword) showed that it was never used in that sense; the reviewers used it as part of a different expression used to offer recommendations. This lexeme was removed from the healthiness lexicon; and its frequent occurrence led to the idea of creating a new lexicon of advice terms.

The KWIC results for each lexicon were examined to make sure that the results did refer to the concepts, and when there were only a few exceptions for a lexeme, that those exceptions were deleted instead of removing the lexeme. For example, “uygun” (“appropriate”) carries in the same sense as “reasonable” in the English expression “reasonable prices.” There were a few instances where it was used to describe the restaurant, as in “appropriate for birthday parties.” These instances were deleted before the lexicon’s frequency was calculated. Then, as lexicons – or their components – notably decreased or increased as the price rating increased, these concepts were marked for correlating with the price.

Frequency, Length, Grammar and Etymology Analyses

Several other analyses were done to test for how language use changes with the restaurants’ price ratings. The four analyses were word-and sentence-length comparisons; lexeme frequency and keyword list comparisons; parts-of-speech comparisons; and comparative etymological analysis. The first analysis consisted of calculating the average *word length* and *sentence length* for each level of price rating on the restaurants to see if higher price ratings correlate with longer words and sentences in Turkish, or if the morphological and historical differences between the two languages mean that that is not a helpful criterion for distinguishing between levels.

The second analysis was to see if the restaurant reviews vary by lexicons other than those developed in previous literature. The *frequency* lists for each corpus was extracted from AntConc; the frequency list is the list of all the lemmas in the corpus ordered from most frequent to least. I then followed Jurafsky et al. (2014) in attempting to posit categories or explanations for highly frequent lemmas. For example, Jurafsky traced the abnormal mentions of “crack” and “drugs” in the reviews of inexpensive restaurants to the fact that writers were using addiction metaphors to describe the inexpensive food that they liked. For this study, examining the lists led to the creation of new lexicons like the “pricing lexicon” and the “personnel lexicon”, which involved building up full lexicons for these subjects and seeing how they varied across the corpus.

The third kind of analysis was *grammatical analysis*. The Jurafsky et al. paper covered one grammatical point when they noted that use of past tense verbs tended to correlate with negative reviews.

However, they did not discuss any grammatical features that correlated with price level. Once the Turkish corpora were tagged with part of speech tags, it was possible to check if characteristics like the number of adjectives and adverbs, grammatical tenses and aspects, and noun cases varied by the price rating.

Finally, the *keyword etymologies* were analyzed. The top 120 keywords from the lemma lists of each corpus was used to determine if word root languages varied by price level. The keyword list for each corpus was tabulated by measuring its lemma list against a comparison corpus composed of the lemma lists of the three other corpora. Thus, the keywords are not the most frequent lemmas for a corpus, but rather the lemmas that occurred significantly more often in that corpus, compared to the other corpora. Of course, a large section of the keywords (from 28 of the C4 keywords to 70 of the C2 keywords) consisted of restaurant names and food names that occurred frequently. While place and food names were discarded from other analyses, they were kept here because they reflect the fixed vocabulary that the restaurant owners or the culture has selected for places and items at that price level. However, to differentiate between the reviewers' independent word choices and the words they do not choose, the name lexemes were extracted and calculated separately.

To calculate the keyness of the words, Antconc's log likelihood (LL) method, based on the principles described in Dunning 1993, was used instead of the chi-square method, because the LL method is considered a better estimate of keyness for smaller corpora with less frequent lemmas (Rayson & Garside, 2000). The Dunning statistic is a method for identifying the words that are unusually frequent (or key) in a particular corpus, using a reference corpus as the baseline. In this study, each corpus was contrasted with the three other corpora. For example, for C3, the reference corpus was a combined corpus of the texts from C1, C2 and C3. Each word in the corpus is given an LL co-efficient, which is a statistic indicating how key the word is; the higher the number, the more unusally frequent it is compared to the reference corpus. To establish that the difference in keyword frequencies between two smaller, similar corpora is actually significant, words with an LL co-efficient of 6.63 or above have a 99% confidence rating (or $p < 0.01$) that they vary significantly, not as the result of chance (Rayson, 2003; Wilson, Archer & Rayson, 2006). Once duplicate lemmas were discarded, C4 had 118 lemmas with an LL co-efficient above 6.63; thus, for the three other corpora, only the top 120 keywords were sorted by root language, to ensure that a similar number of words was compared across corpora. A chi-square test was conducted on the 478 words to see if the associations between corpora and root language were significant.

CHAPTER 4: RESULTS

Introduction

Among the analyses conducted, the most productive was the lexicon analysis. As expected, the *frequent lemmas and keywords* mainly reinforced the trends found in the lexicon analysis. The lexicon analyses showed that the Turkish reviews did not follow the price-based trends that Jurafsky identified in the American corpus, though there were other price-based lexicon variations. The etymology study also produced unique indicators of the price level, and these results were also supported by the keyword analyses. The *grammatical* and *sentence and word length* analyses did not lead to significant differences. Thus, after describing these analyses that emphasized the corpora's similarities, this results section will discuss the lexicon analyses and the etymology analysis and conclude with how the keyword lists supported those results. While not every type of analysis found measurable differences across corpora, there were a number of significant differences that show that, as hypothesized, price-based variation exists and can be measured in the reviews.

Unproductive Analyses: Grammar and Word Length

Grammatical features show very little variation, particularly across verb structures. No price level favors a significantly higher use of a certain person, number, tense, or any of the many aspect markers that are attached to the verb; none of these categories varied across the corpora by more than 3%. This also held true for the other parts of speech; the corpora had very similar rankings when it came to the percentage of the total words that each part of speech tag identified. For every POS tag in the corpus, the variation in the percentage of lemmas with that tag never varied by more than 3%. for example, for adjectives, 16.7% of the lemmas in C2 had an adjective tag, and 18.5% of the lemmas in C1 had an adjective tag. One of the important findings from this similarity is that writers do not change how they refer to themselves or to their audience in the plural vs. the singular by price range. Turkish does have different levels of formality in address, indicated either by the formal plural you or the informal singular you (Kerslake & Goksel, 2014, p. 287). Thus, the grammatical analysis indicates that writers are not consciously adapting more formal language or demonstrating more respect to their audience when they review more expensive restaurants.

The word length analyses also showed little variance across the corpora. All four corpora fell within the range of 7.19 letters per word (C1) and 7.29 letters per word (C3). Similarly, the number of words per sentence remained similar, from 15.29 (C3) to 16.15 (C1) words per sentence. As the variance was so small, no statistical tests were conducted to measure significance. While the informal nature of the writing means that the numbers themselves may not be accurate (reviewers differ in if they write certain question words as one word or two or if they use punctuation grammatically), the similar results make it clear that reviewers are not choosing longer words or increasing their sentence language when they talk about more expensive dining experiences.

Lexicon Analyses

The lexicon-based searches were more revealing. To begin, the lexicons of metaphoric words used to describe food do not parallel the lexicons found in Jurafsky's research. Lexemes related to addiction and attraction only appeared in Corpus 1 and 2, and they appeared minimally; there were five instances of words related to addiction and only one word about attraction. The addiction metaphors simply referenced "addiction" ("bağım") without using related terms; for example, "[the meal] creates addiction," "I am addicted to this place," and "try it once, and you'll be an addict." Similarly, the one instance of attraction was, "[the meal] made me fall in love." These metaphors did not appear in the other corpora. The lexicon search included variations of "feel" and "feeling," to catch possible romantic or sensual feelings. In C1 and C2, these were mostly used to describe feeling at home, and for the more expensive restaurants, the feeling of peace, or of entering a home, a historic locale, or another country. Feelings and sensations were mostly reserved for the ambiance of a locale rather than the sensations of a particular dish, and they did not have the romantic connotations found in the American corpus.

However, the lexicon search revealed a novel common food metaphor that peaked in C3: expressions that involved the root “bayıl” (‘faint’). To “faint for” something is a term expressing delight in Turkish, and it appeared in all four corpora. It shows up in various forms: “I am the type of person who faints for onions,” “They made it well but I didn’t faint for it,” “I fainted for this place,” “I fainted most for the squid,” “I ate it for the first time and fainted for the taste.” It is applied to specific dishes, to the atmosphere, and to objects in the restaurant (like cats or a teapot). While common in C1 and C2 (at 7.85 and 7.15 occurrences per 10,000 words), it jumps in C3 to 13.10 occurrences per 10,000 words before almost disappearing in C4, with 1.30 occurrences per 10,000 words. As a more informal term, and a very expressive one, reviewers do not seem to find it appropriate for describing experiences and food in C4.

The Turkish reviews also do not reflect the trends seen in American food advertising by Freedman and Jurafsky. To begin, Freedman and Jurafsky found that the more expensive the chips, the more mentions of health on the packaging (Freedman & Jurafsky, 2011). In the Turkish Yelp reviews, however, the lexicon of healthiness was highest in C1. The lemmas “doyur, doğal, sağlık” (“filling, nature, health”) were most frequent in C1; “fresh” was the only one that was higher in another corpus (C3) (Figure 1; refer to the appendix for the numerical data represented by each figure). Thus, Turkish consumers do not see healthiness as primarily a concern when speaking of expensive food; health is a more frequent consideration when choosing and describing daily food, and it shrinks to irrelevance when evaluating expensive food. “Fresh” is the only health attribute that increased with the food’s price.

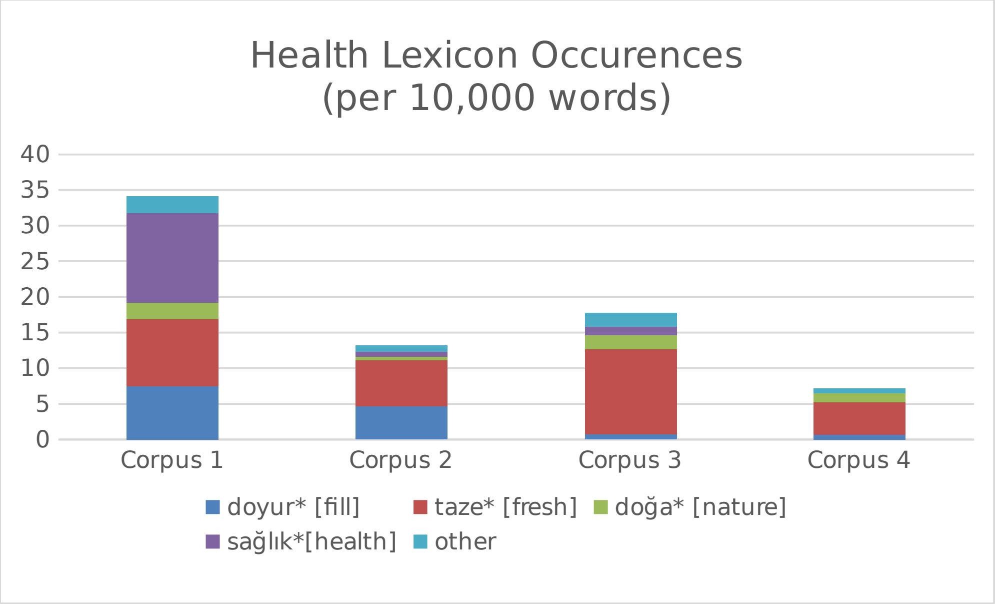

Figure 1: Health Lexicon Occurrences

The lexicon of unhealthy food terms yielded similar results. Reviewers used terms like “ağır, rahatsız, fast food,” (“heavy” which for food means fatty, oily, or over-seasoned; “discomfort,” and “fast food,” which refers to cheap American-style fair) most often in C1 and C2 (Figure 2). The most common lemma was “mide” (“stomach”), as this was always used in expressions that refer to the unhealthiness of a dish: “after leaving, the food made my stomach uncomfortable,” “If you eat here, the food can touch your stomach,” (a polite way of saying it will make you nauseous); “the oily food made my stomach burn.” The KWIC results of this lexicon demonstrate that when discussing the unhealthiness of a meal, Turkish reviewers are concerned with a different unhealthiness than Americans. In Freedman and Jurafsky’s research, potato chip advertisers mostly stressed the absence of fat, sodium and cholesterol content. (2014). Turkish reviewers are concerned with whether a food is well prepared (as in “ağır”), and if it will lead to discomfort or illness (the “mide” tokens). The tokens “vitamin,” “protein,” and “nutrition” in the healthy lexicon, and “prepared,” “too sweet,” and “fatty,” in the unhealthy lexicon, returned zero hits. Finally, the low numbers of unhealthy and healthy lexemes per 10,000 words (Figure 2) highlight that this is a less important concern for the reviewers. While it is worth mentioning if inexpensive food is filling and healthy, neither of the lexicons feature heavily.

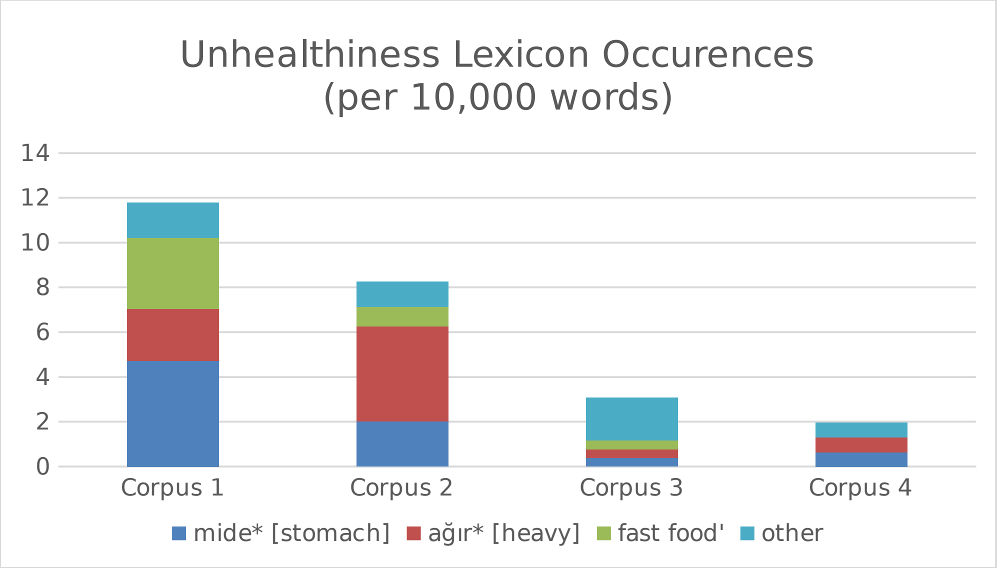

Figure 2: Unhealthiness Lexicon Occurences

The second lexicon that showed price-based variation in Freedman and Jurafsky's study was uniqueness. The more expensive a food, the more claims there were that it was superior or exceptional. The lexicon for this category was composed of words like "özel, rakipsiz, eşsiz, özgün, has, tek" ("special, without competitor, unequaled, unique, particular [to], only"); words that were used to describe dishes or places as unlike any other. The trend of increasing claims of uniqueness as price increased was not supported in the Turkish Yelp data. Here, the highest rates of words for uniqueness occurred in C2, with 15.19 per 10,000 words belonging to the uniqueness lexicon. C1, C3 and C4 showed little variance, with levels from 10.98 to 11.94. The word "özel" ("special") occurred most frequently, and was more common in C3 than C2. C4 had 9 occurrences of "özel" that were not included because they refer to "special days" or "special occasions" rather than dishes or atmosphere. A sub-corpus of “Rakipsiz, eşsiz, has, özgün, bulunamaz, tek,” the words that specifically imply incomparableness, as opposed to something special, was created, to see if that semantic subcategory had different results. However, this lexicon occurred very sporadically, with only 3.08 to 7.15 words per 10,000, thus making it too small to measure for significant variation.

These results indicate that rarity and individuality are not as important a part of high-end dining in Turkey as they are in the United States, and that an increase in describing uniqueness does not correlate with more expensive cuisine; in fact, the category of restaurants that reviewers reviewed most frequently (C2) had the most claims of uniqueness. Turkish reviewers are more likely to talk about distinctive dishes and atmospheres when describing the advantages of a middle-level restaurant.

The final lexicons that Freedman and Jurafsky studied were the words for food that describe it as either traditional or as authentically exotic. Whereas tropes like family recipes and local customs were emphasized on the less expensive food’s advertising, the more expensive food was described as authentically foreign: exotic locales, exotic traditions. However, the lexicon of “tradition” in this study combined both types of authenticity, as most of the individual lexemes are used in either sense in Turkish. It was necessary to refer to the lexicon’s KWIC concordance to understand how this lexicon was employed.

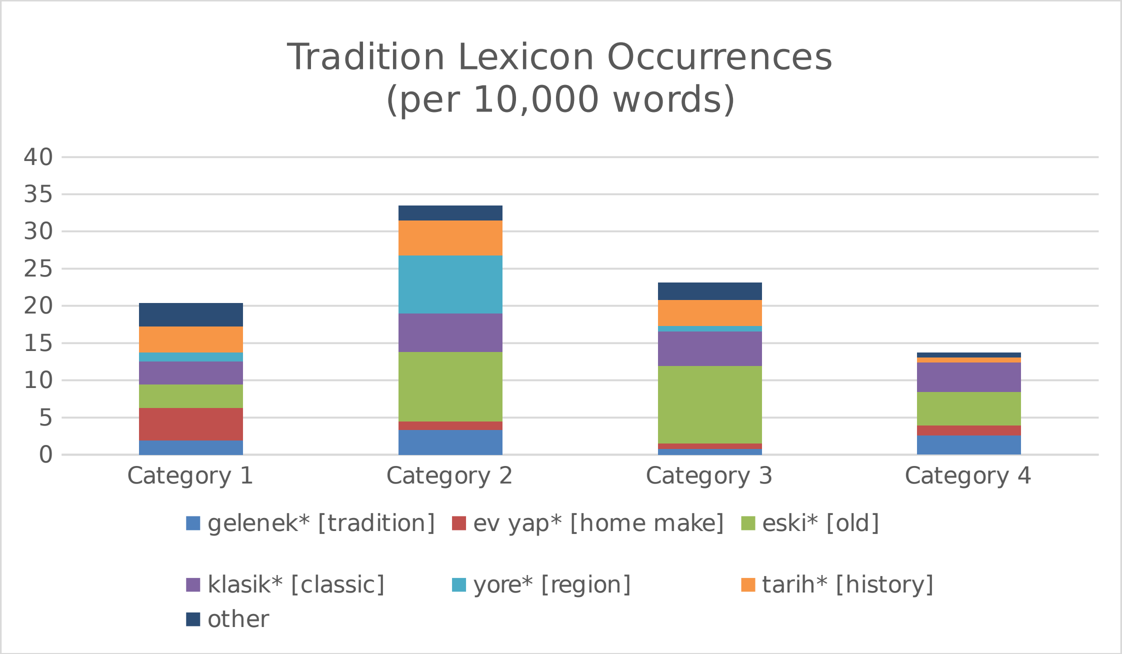

Figure 3: Tradition Lexicon Occurrences

The tradition lexicon peaks in C2 and occurs least frequently in C4. Most of these lexemes, in context, represent homey authenticity: "gelenksel, ev yapımı, klasik, yöresel, tarihi" ("traditional, homemade, classic, regional, historic"). Though some of the restaurants serve Italian, Japanese, French and even Scandinavian, cuisine, the authenticity lexemes did not appear there. Aside from three mentions of "a classic pizza menu," "a classic schnitzel," and "classic drag queens," all the KWIC results described Turkish dishes or Middle Eastern-style décor. When "historic" and "regional" appeared, it was clear from context that the reviewers were always implying "historic Turkish" or “regional Turkish.” “Eski” (“old”), which predominates in C3, describes authenticity of décor. This means that for the reviewers, while there is some concern for traditional authenticity, this applies primarily to C2 dishes. The authenticity of regional dishes to their Turkish province of origin is primarily a method of evaluation for C2 dishes. In the upper restaurants, where there are more imported cuisines, the concern with Turkish authenticity is not replaced by a desire for authenticity in exoticism; the authenticity of foreign dishes is unimportant. There is, however, a small increase in the concern for authentically historic or old décor.

The next two lexicons under analysis were not analyzed in previous research, but they correlated more clearly with the Turkish restaurants’ price levels. The first one, the pricing lexicon, is composed of both neutral words for price, such as “fiyat, bahşiş, hesap, lira,” (“price, tip, total, lira”), and words denoting the reviewer’s attitude towards the price, such as, “pahalı, ucuz, uçmuş, soygun, uygun” (“expensive, cheap, sky-high, robbery, suitable [price]”). There is a steady decrease in pricing words per 10,000 from C1 to C4 (Figure 4). Most words within this lexicon follow that trend, with the exceptions, like “pahalı” (“expensive”) and “yüksek” (“high”), referring to high prices. Words like “ucuz” (“cheap”) decrease markedly after C1, and are usually followed by a negator (67% of “ucuz” occurrences in C3 were negated, and in C4, of the five occurrences, three were negated and the other two referred to the quality, not the cost).

Thus, reviewers use the words one would expect for each price level, and they talk less about pricing as the price increases. The first trend is interesting only in that it indicates that individual descriptions of dishes and location mirrors the reviewers’ overall price ratings. Reviewers do describe the restaurants as more expensive at higher levels of the corpus, and less expensive at lower levels. The second finding demonstrates that one of the clearest lexical indicators of a rising price level is that reviewers decrease their use of all price-related words. When it comes to expensive restaurants, neither the costs of the dishes and services, nor the reviewers' attitudes towards that price, is a key factor in their evaluations.

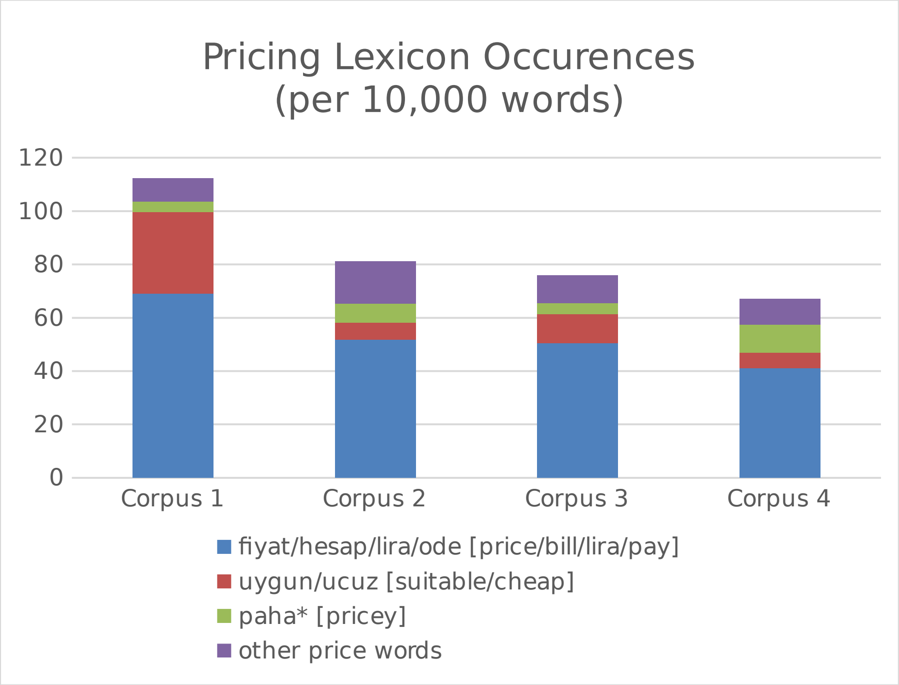

Figure 4: Pricing Lexicon Occurrences

The personnel lexicon also changes by price level (Figure 5). The frequency of personnel-related lemmas does not vary dramatically; between 32.56 to 40.87 words per 10,000 are used to name and describe the restaurant staff. What does change across the corpora are the terms used for the staff. Terms for the staff shift towards European terms and more specific, formal terms as price level increases.

In C1, "çalışan" ("worker") and "sahib" ("owner)" predominate. There are also a few mentions of "abi," and "amca," ("big brother" and "uncle"). In C2, "garson," ("waiter"), and "hizmet" (a general term for those serving) prevail. "Personel" ("personnel") is highest here, though it is not the most common lemma in any corpus. C3 is similar to C2, but "eleman" (employee) also rises. In C4, "şef" ("chef") spikes upward.

Across the fours copora, there is a slow shift from “çalışan” and “sahib,” to an increased use of “garson,” “eleman” and “şef”.

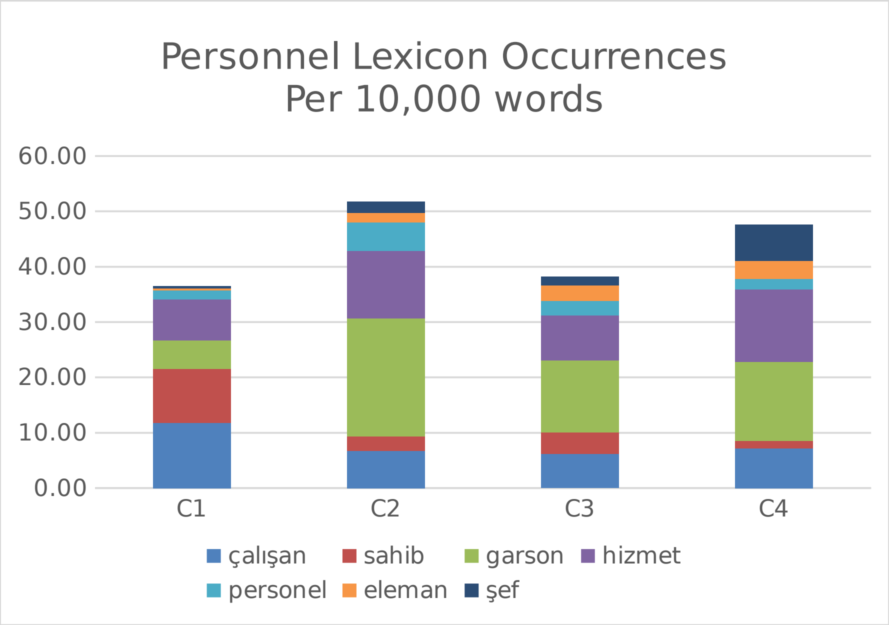

Figure 5: Personnel Lexicon Occurrences

The two final lexicons both measure the attitude of the reviewers. The first one consists of the reviewers’ attitude toward the audience in giving advice. The “-meli” verbal suffix in Turkish expresses obligation, like the English “must” or “should” (Kerslake & Göksel, 2014, p. 199). By searching its phonetic variations in AntConc, the frequency of this verb affix in the corpora was calculated. As seen in Figure 6, this affix occurs frequently throughout the corpora, but it predominates in C3, and this pattern holds regardless of the person affix of the verb (which indicates the object of the command). There two common expressions are equivalent to “it is beneficial to” in Turkish: “-da yarar var” and “-da fayda var.” As with the commands, these expressions occurred in all the corpora, but were most common in C3. “Dene” (“try”) is the final lexeme in the advice lexicon that appeared frequently, though this was as common in C2 as C3. Thus, reviewers were most likely to offer explicit advice to their audience in C3.

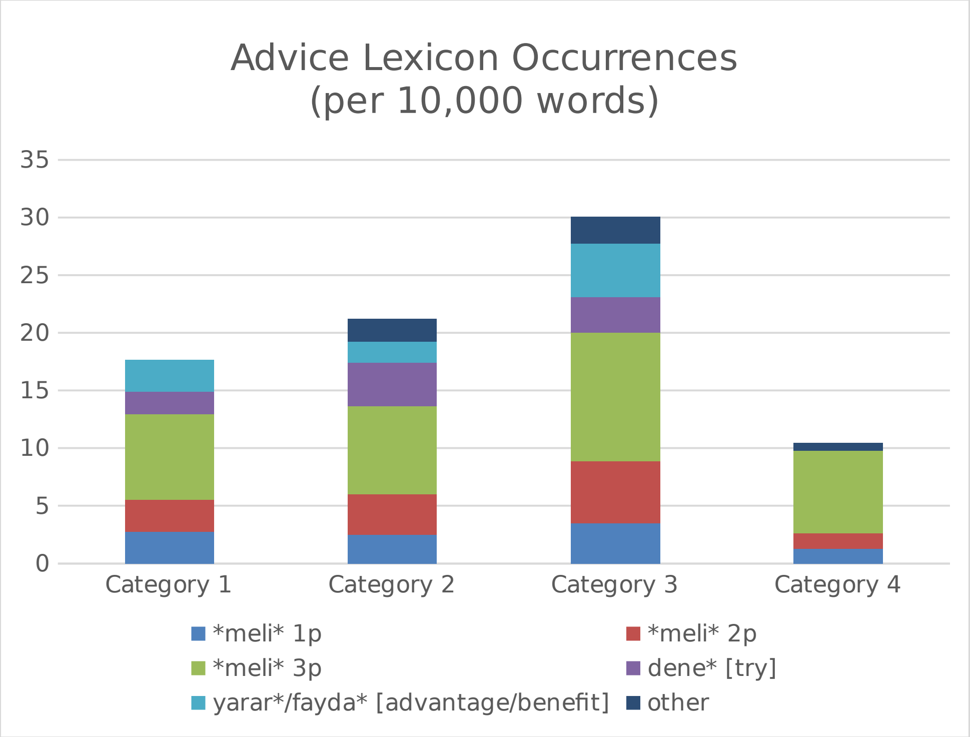

Figure 6: Advice Lexicon Occurrences

The second attitude lexicon is the lexicon of lexemes used to indicate strong opinions: “mutlaka, kesin, asla, lazim, vazgeçme” (“certainly, definitely, never, necessary, don’t pass up”). These parallel the advice words, in that they are highest in C3; however, this is a slight increase; C1, C2 and C4 all have between 19.57 and 24.12 certainty words per 10,000, so the variation between those corpora and C3 with 27.74 words per 10,000 is negligible. An analysis of a lexicon of uncertainty words also showed similarity across the corpora, further confirming this finding. Thus, while the reviewers are most likely to issue explicit advice in C3, the certainty expressions that accompany their opinions do not vary.

Summary of Lexicon Analyses

These lexicon analyses demonstrate that the Turkish Yelp review lexicons pattern differently than in the American reviews. Instead of speaking of addiction or sensuality, the predominant metaphor for delicious food was fainting, and it was common in all the corpora except for C4. The healthiness and uniqueness lexicons showed an opposite pattern from what Jurafsky found in American advertising; health and unhealthiness were more common concerns in the inexpensive restaurants. Reviewers criticized the unhealthy food for its unpleasant effects, not for inducing weight gain. In terms of authenticity, lexemes for exotic authenticity were vanishingly rare, even in the upper corpora. However, similar to the American findings, homey authenticity was emphasized most in C1 and C2 and was less important in C3 and C4. Uniqueness as a criterion for evaluating restaurants was negligible and consistent across corpora. The lexicons that showed consistent variation were those that had to do with pricing and personnel. Reviewers mentioned pricing and their attitudes towards price less as the restaurant price increased; and they slowly changed their terms for the owners, chefs and wait staff as the price increased. Finally, reviewers were most likely to give direct advice in C3.

Etymology Analysis

After the lexicon analyses, the other analysis that demonstrated significant price-based variation was the analysis of word roots. Of the 120 keywords from each corpus, certain root languages appeared more often in certain levels of the corpora (Figures 7 and 8). A chi-squared test that compared the found that the etymological languages differed significantly by price level, with χ^2 (15, N = 273) = 30.94, $p < 0.01$ for the general words, and χ^2 (15, N = 204) = 46.58, $p < 0.001$ for the words that were restaurant or food names.

Lemmas with Turkish roots are more preponderant in C1 than in any of the following corpora, both in the freely chosen words and in the names. The Western languages – Romance languages (mostly French and Italian, but also Greek) and English – increase steadily and then spike sharply in C4. For example, in C4 the reviewers use the lemmas “şov, show, brunch, teras, kostüm, modern, dining, mönü,” (“show, show, brunch, terrace, costume, modern, dining, menu”) in their reviews, and the place names were clearly Western-influenced, as in “The Kev Café” (“the Cave Café”) and “Mia Mensa” (“my table” in Italian), while the menus included items like “kokteyl” (“cocktail”).

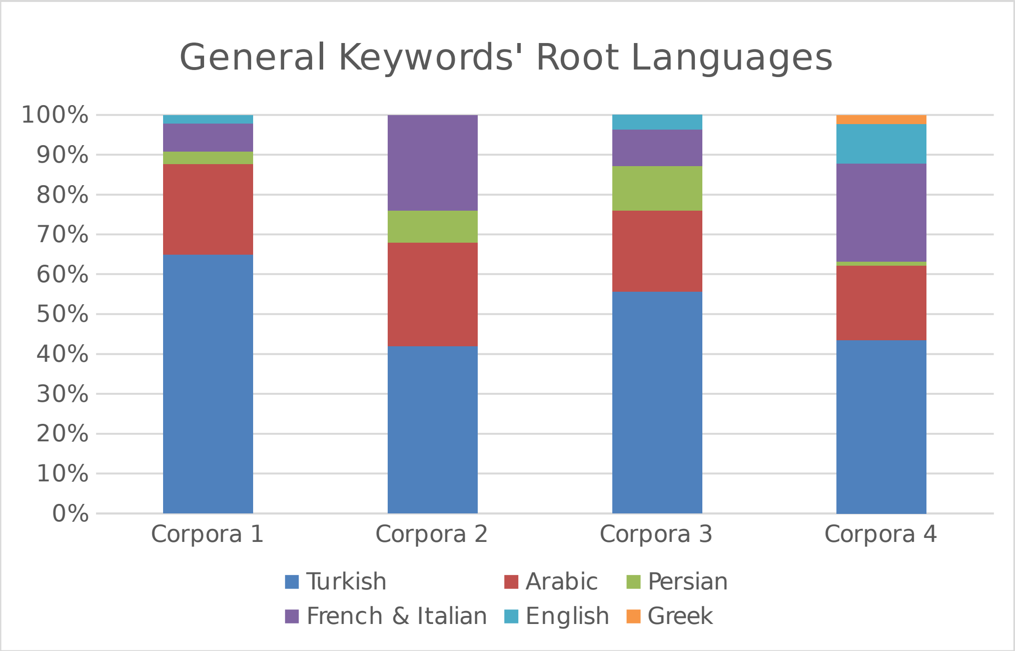

Figure 7: General Keywords’ Root Languages

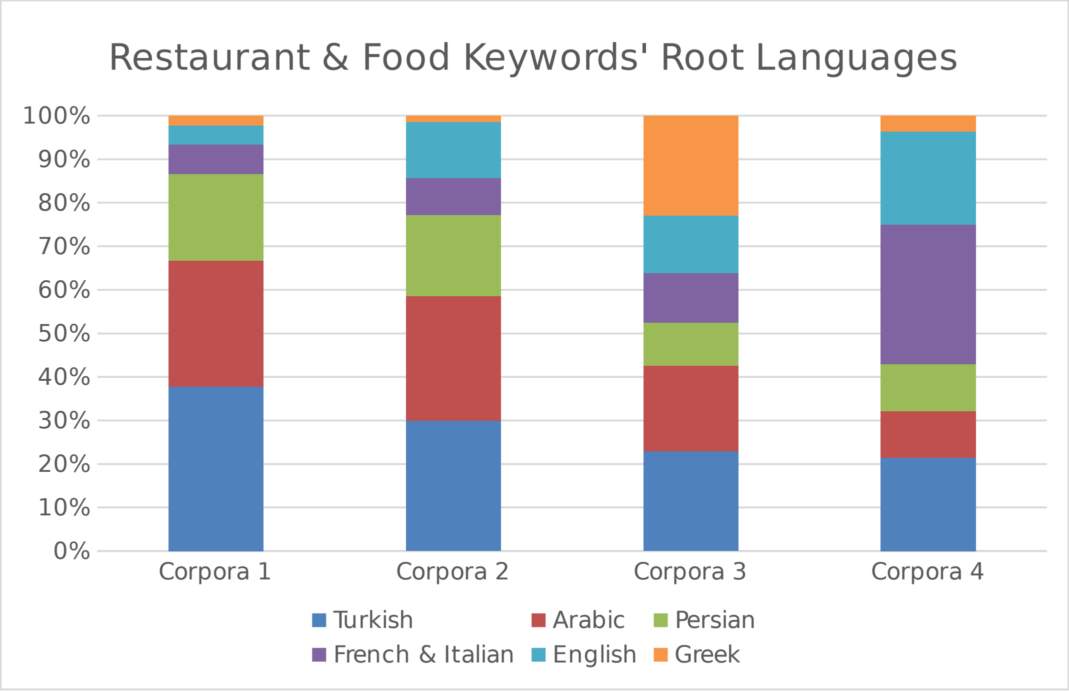

Figure 8: Restaurant & Food Keywords’ Root Languages

When it came to the restaurant and food names, Greek roots showed a large increase in C3 names only, from 1.4% of the lemmas in C2 to 22% in C3. This is explained by the surge of seafood dishes in C3, which are borrowed from Greek cuisine. Examples of this are “lakerda, levrek, lüfer, midye” (“bonito, bass, bluefish, mussel”). The Arabic roots do not vary significantly between the corpora, as they vary nonlinearly from 18.89% to 26.00%, but the place names with Persian and Arabic roots do decrease from 48.89% to 21.42%. Arabic and Persian roots seem to be so widespread and unnoticed in the language that the speakers do not pay attention to their use of words, but they decrease in the names because they are crowded out by Western and European names that have more foreign glamour.

Thus, as price increases, the reviews are marked by an increase in Romance languages and English-based terminology. This is both a conscious decision on the part of those who choose the names for the establishments and their dishes, and evident in the words that reviewers choose, unconsciously or consciously, to describe their experiences at these locales. Though the Arabic and Persian restaurant and food names decrease as price increases, reviewers do not reflect this in their review language; words with Arabic and Persian roots are acceptable for describing a high-end experience.

Support from Keyword Analysis

While the frequency lists for individual lemmas were used to create the original lexicons, the keyword analysis based on Dunning’s log-likelihood statistic also confirmed the lexicon and etymology findings. The top five keywords for each corpus are below (Figure 9), and they fit with previous findings.

Top Keywords for Each Corpus			
C1	**C2**	**C3**	**C4**

ev [house]	kahvaltı [breakfast]	rezervasyon [reservation]	manzara [scenery]
sağlıklı [healthy]	yöresel [local]	fasıl [fasil]	bar [bar]
alıyorsunuz [you take]	türk [Turkish]	deniz [sea]	sahne [stage]
ucuz [cheap]	pazar [Sunday]	restoran [restaurant]	şov [show]
sebze [vegetable]	kart [card]	Cumartesi [Saturday]	show [show]

Figure 9: Top Keywords for Each Corpus

The first, second, fourth, and fifth key lemmas for C1 are “ev” (“home,” from “homemade foods”), “sağlıklı” (“healthy”), “ucuz” (“cheap”), and “sebze” (“vegetable”), following the lexicon trends. “Alıyorsunuz” (“you take”) is a result of the way food service is set up in lower-priced restaurants. For C2, “yöresel” (“local”), and “Türk” (Turkish) reflect the finding from the lexicon analysis that locality and authenticity are more important in the lower-priced restaurants. “Kahvaltı” (“breakfast”) and “Pazar” (“Sunday”) are a result of the custom of going out for breakfast on Sundays, which usually falls in this price range, and “kart” (“credit card”) is mentioned frequently because reviewers are warning their audience about the mid-level restaurants that still do not accept credit cards. In C3 and C4, five of the words are connected to the increase in European word roots (“rezervasyon, restoran, bar, şov, show”). “Fasil” (a type of traditional music) and “Cumartesi” (“Saturday) predominate because several of the restaurants play fasil music on Saturdays. “Manzara” (“scenery”) and “sahne” (stage”) are elements associated with expensive restaurants. Thus, the keywords can be explained by practical details related to the restaurants or by the trends found in the etymological and lexical analyses.

CHAPTER 5: IMPLICATIONS AND CONCLUSION

Theoretical Implications

It is evident from the lexicon and etymology analyses that there are significant lexical variations between the corpora, both in terms of the lexicons that appear and the etymologies of the keywords. Two themes emerge: that the significant differences between corpora do not follow the trends in the American research, and that several indicators (grammar, word length, some lexicons) show little to no variation. Thus, the trends in Jurafsky and Freedman's research are particular to American culture; while a corpus analysis does reveal price-based variations in this new language and context, the most significant variations appear in novel lexicons, or appear in the opposite direction of those found in the American research.

A review of the main findings demonstrates that the data lends some support to the theories advanced by Bourdieu and Douglas and Isherwood. Yelp reviewers do not change their overall styles when reviewing products that are more expensive: their grammatical features, word and sentence lengths, and predominant metaphors for tasty food and concern for uniqueness all showed little variation. However, several lexicons did show significant variation: healthiness, homey authenticity, price and attitudes to price were more important in inexpensive restaurants. The appropriate terms for the wait staff shifted, and reviewers gave the most direct advice when describing the C3 restaurants. An increase in English, French and Italian borrowings also accompanied descriptions of more expensive experiences. While these are subtle shifts, and the C1 terms rarely disappeared in C4, and vice versa, there are indications that the reviewers were evaluating their more and less expensive experiences on different criteria, and thus were establishing symbolic cultural capital through the act of writing a review.

One of the hypothesized differences was that reviewer would judge the inexpensive restaurants on providing sustenance, while judging expensive restaurants on their ability to provide pleasure and a unique or rare experience. The fact that instances of the health lexicon and the unhealthiness lexicon occurred more frequently in C1 and C2 support this claim, as people were concerned that the lower-priced meals should be filling and good for their health. Though this is different from the American corpora, where concern with health was more frequent in the upper corpora, it makes sense in the Turkish context. Instead of an upper-class concern on healthy eating, customers are more concerned about the sanitation and nutrition of their cheaper meals. These concerns decrease in the upper corpora, as the reviewers take sanitation for granted and do not consider nutrition to be an important criterion for a quality meal. At the same time, at the inexpensive level, reviewers are more likely to correlate the price with the quality, whereas at the upper levels, price again is less of a criterion.

Unlike the America studies, C3 and C4 did not have increased levels of the uniqueness lexicon. In fact, of the lexicons for different semantic concepts that were traced throughout the reviews, there were none that occurred significantly more often in the pricier corpora than in the less expensive ones. Turkish reviewers did not seem to care how authentic the non-Turkish dishes were, and they did not use more sensual language for the more expensive dishes. There are two possible explanations for this; first, that no set of vocabulary predominates in the upper levels because the major lexical mark of C3 and C4 reviews is a more diverse spread of concerns; or second, that there are lexicon(s) that appear more often, but that they were not among the lexicons considered in this study. Thus, the patterns in the upper corpora do not follow the trends in American corpora, but what the Turkish trends are is an open question.

The hypotheses were that lexical choices and structures in the reviews would vary with the price rating, and that these variations would be visible through corpus analysis. The results support that, though only on some of the several analyses performed. While price-based lexical variation is evident in the corpora, it is generally manifested in changing percentages. Instead of one lexicon or etymological language going from nonexistent to predominant, the analyses demonstrated consistent shifts across the corpora. The price-based lexical variations were also highly culture-specific; the Turkish and American corpora had few overlapping trends. In line with the theories of Bourdieu and Douglas and Isherwood, reviewers at different class levels did shift what values they used to review the food. They demonstrated their cultural capital by employing more European words when reviewing more expensive restaurants, but the evidence was perhaps subtler than the social theories would have led us to believe.

However, the lack of dramatic results does not mean that we should neglect studying price-based variation, as this research combines an understudied area within linguistics with the newer trend of studying registers. Researchers often neglect social class and its effect on language in favor of studying the relationship between identity markers and language. David Block, in his summary of social class in sociolinguistics, notes that, instead of denial of social class, there has been “social class *erasure*, as social class has tended to receive little or no attention in publications that deal with issues around identity and social life,” (2015, p. 2). Thus, it is essential that as linguistics moves to more analysis of corpus data, we reintroduce analyses that examine the relationship between class and language. Big data makes it possible to do so from lexical and grammatical angles. Corpus studies like this one also make it possible to examine socioeconomic lexical changes without pinning that to speakers of a demographic, which aligns with the push in sociolinguistics to view language use as a matter of register and stance, not just as a matter of associating particular sounds and expressions with a class, gender or ethnic variable. Eckert explains that variation studies originally “established broad correlations between linguistic variables and the macrosociological categories of socioeconomic class, sex, class, ethnicity and age.” Now, however, linguists acknowledge that speakers command more than one register, and each register “does not simply reflect, but also constructs, social meaning.” (2012, p. 87). As a result, language users (or, in this case, Yelp reviewers) are “stylistic agents, tailoring linguistic styles in ongoing and lifelong projects of self-construction and differentiation” (p. 98). Thus, even if the same reviewer is writing in multiple levels of the restaurant corpora, their shifts in lexical choice and root languages model a shift in register. A register-based approach makes sense within the Bourdieu context as well, as it understands that speakers may shift to the linguistic repertoires that gives them the most social credibility and capital in different situations.

Connecting Corpus Results to Cultural Phenomena

The results allow for some speculations into the nature of language and society in Turkey and the reasons for the trends observed. Given my own knowledge of Turkish society, I can postulate reasons for the patterns found in the health, pricing and personnel lexicons, though sociological research are necessary to move them from the realm of possibilities to true hypotheses. Health and value for price appear most frequently as evaluation criteria for the more inexpensive restaurants. This is unsurprising, as Turks talk frequently about their weight and health and the benefits of different foods. However, they do not often use the specialized scientific words (carb, protein, calorie, etc.) that are frequent in American discourse. At the same time, people usually go to the expensive restaurants to celebrate a special occasion, and celebrations are an important showcase of one’s generosity and taste. Expensive, high-quality food in Turkey tends to be unhealthy (with more meat, oils and sugars than in ordinary food), but as an expensive meal is supposed to indicate one’s generosity, not one’s health-consciousness, healthiness should matter much less than quality, service and presentation when evaluating a pricier restaurant.

Similarly, Turks talk freely about money; for example, asking a friend what their income is or what they pay for rent is generally not taboo. However, because of the importance of generous hospitality, the host of a party or event is expected to spare no expense. They also usually pay for everyone’s food; friends might take turns paying for different occasions, but it is viewed as poor form to split a bill multiple ways, Thus, for the expensive restaurants in particular, the reviewers most likely either paid for the entire dinner party, or were the recipient of their host’s generosity and never saw the food prices. Thus, it is not cheap or crude to evaluate a lower- or middle-priced dish based on the price you paid for it. However, complaining about the price if you are paying for your friends or receiving their generosity would be rude; reviewers will feel more justified in complaining about the quality or the experience.

Finally, service is an essential part of the restaurant experience at every level in Turkish society. Because wages for workers are generally lower in Turkey than in the United States, Turkey has not experienced the push towards automated service and self-service that is evident in the more inexpensive American restaurants. Even local soup-and-rice restaurants, which serve some of the most affordable food available, still have staff that greet the customers, seat them, take their orders, bring their food, stop by during the meal to ask what they need and socialize, and clean up after they leave. Thus, it is not surprising that while mentions of service personnel occurred with similar frequency across the price levels, the reviewers' relationship with the personnel and attitude towards them shifted with the price level. In less expensive restaurants, the customer's relationship with the staff is informal and even familial; if customers frequent a restaurant often, customers and staff are likely to learn each other's names and background information. In a more expensive restaurant, however, the relationship is much more formal and distant; the staff will behave deferentially to the customer and there are more expectations about the staff's speediness, attention to detail and good manners. The replacement of more general, even familial terms with more precise, European terms may reflect this change in the social relationship.

Methodological Implications

The most original aspect of this study was the methodology created to identify price-based lexical differences in a corpus. Lexicon studies and statistical methods for comparing etymological languages are rare, and to my knowledge neither had previously been conducted in the Turkish language. The lexicon methodologies developed in English by Jurafsky et al. (2014) were only possible to employ when working on a large scale with skilled computer scientists and computer-collected corpora. The methods employed in this paper are more appropriate for a detailed analysis of a smaller, manually collected corpus, as it combines automated searches with personal examination and evaluation of the results. These methodologies were built on trial and error and with the input of several native speakers and colleagues. Through the process, I identified the steps necessary to building lexicon lists that accurately represented those concepts: accounting for spelling variations, seeking outside input, including common collocations, reading KWIC concordance lines to determine the appropriateness of the words, and deleting irrelevant results before arriving at the final counts for each lexicon. This methodology can provide an example for future researchers looking for reliable and comprehensive methods for tracking the relative occurrences of semantic concepts within corpora. While the particular morphological and syntactic issues in this paper are specific to Turkish, the process itself is applicable to other under-studied languages.

In this study, to identify possible correlations between root languages and price ratings, the novel methodology of tabulating the etymologies of keywords was employed, and a chi-square test was used to determine the strength of the relationship. It might be beneficial in the future to supplement this with the etymologies of the most frequent words as well. However, keyword etymologies allow the researcher to skip over identifying the etymologies of the most frequent content words that appear in every corpus and will show little variation, and to concentrate instead on the words that make the corpus distinctive.

Corpora of online reviews make it possible to study price-based register shift at the level of hundreds or thousands of language users in a way that was previously impossible. In the future, it would be helpful to extend this type of price-based corpus analysis to a category other than restaurants, or a new language and culture combination. In the American reviews, sentence length proved to be relevant; in Turkish, etymological language was more important; and each demonstrated different lexicon-based variations. Other languages might pattern along other linguistic features or criteria of evaluation. Capturing these patterns requires methodological experimentation and innovation; explaining them will push the boundaries of current theoretical and sociolinguistic inquiry. A more complete catalogue of the types of price-based variations in review languages would play an important role in understanding how speakers use language and register variation to gather, demonstrate and increase cultural capital. This field is just emerging, and class-based lexicon studies across a wider variety of corpora will no doubt reveal more of the subtle ways in which language users, consciously or unconsciously, project class awareness in how they evaluate experiences.

REFERENCES

Biber, D. (1995). *Dimensions of register variation: A cross–linguistic comparison*. Cambridge: Cambridge University Press.

Block, D. (2013). *Social Class in Applied Linguistics*. Hoboken: Taylor and Francis.

Block, D. (2015). Social Class in Applied Linguistics. *Annual Review of Applied Linguistics 35*, 1-19.

Bourdieu, P. (1987). *Distinction: A social critique of the judgement of taste.* (R. Nice, Trans.) Cambridge: Harvard University Press.

Bourdieu, P. (1986) *The forms of capital*. In J. Richardson (Ed.) Handbook of Theory and Research for the Sociology of Education. New York: Greenwood, 241-258.

Douglas, M. & Isherwood, B. (2002). *The world of goods: Towards an anthropology of consumption.* London and New York: Routledge.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics, 19*(1), 61-74.

Durrant, P. (2013). Formulaicity in an agglutinating language: The case of Turkish. *Corpus Linguistics & Linguistic Theory, 9*(1), 1-38.

Eckert, P. (2012). Three waves of variation study: The emergence of meaning in the study of sociolinguistic variation. *Annual Review of Anthropology, 41*. 87-100.

Freedman, J. & Jurafsky, D. (2011). Authenticity in America: Class distinctions in potato chip advertising. *Gastronomica 11(4)*, 46-54.

Jurafsky, D., Chahneau, V., Routledge. B. & Smith, N. (2014). Narrative framing of consumer sentiment in online restaurant reviews. *First Monday 19*(4), online.

Kerslake, C., & Goksel, A. (2014). *Turkish: an essential grammar*. Oxfordshire: Routledge, Taylor & Francis Group.

Rayson P., & Garside R. (2000). *Comparing Corpora Using Frequency Profiling*, Proceedings of the workshop on Comparing Corpora, Hong Kong, 1-6.

Rayson, P. (2003), Matrix: A statistical method and software tool for linguistic analysis through corpus comparison. Unpublished PhD thesis, Lancaster University.

Sak, H., Güngör, T., & Saraçlar, M. (2011). Resources for Turkish morphological processing. *Language Resources and Evaluation*, *45*(2), 249-261.

Selvi, A. (2011). World Englishes in the Turkish sociolinguistic context. *World Englishes*, *30*(2), 182-199.

Sezer, T., and Sezer, T. (2015). TS Wikipedia LDC2015T15. Web Download. Philadelphia: Linguistic Data Consortium.

Stone, L. and Pennebaker, J.W. (2002). Trauma in real time: Talking and avoiding online conversations about the death of Princess Diana. *Basic and Applied Social Psychology*, *24*(3), 173-183.

Wilson, A., Archer, D., & Rayson, P. (2006). *Corpus Linguistics Around the World* (Language and Computers - Studies in Practical Linguistics, 56). Amsterdam: Editions Rodopi.

APPENDIX A. LEXICON RAW DATA

Table 1: Metaphor Lexicon

Corpora	Total tokens	Per 10,000	Tokens bayıla* [faint]	Per 10,000	Tokens bağım* [addict]
1	24	9.41	20	7.85	3
2	33	7.37	32	7.15	2
3	35	13.48	34	13.10	0
4	2	1.3	2	1.30	0

Table 2: Health Lexicon

C	Total tokens	Per 10,000	Tokens doyur* [fill]	Per 10,000	Tokens taze* [fresh]	Per 10,000	Tokens doğa* [nature]	Per 10,000	Tokens sağlık* [health]	Per 10,000
1	87	34.13	19	7.45	24	9.41	6	2.35	32	12.55
2	59	13.18	21	4.69	29	6.48	2	0.45	3	0.67
3	46	17.72	2	0.77	31	11.94	5	1.93	3	1.16
4	11	7.18	1	0.65	7	4.57	2	1.30	0	0.00

Table 3: Unhealthiness Lexicon

Corpora	Total tokens	per 10,000	tokens mide [stomach]	per 10,000	tokens ağır* [heavy]	per 10,000	Tokens ‘fast food’	Per 10,000
1	32	11.77	12	4.71	6	2.35	8	3.14
2	38	8.26	9	2.01	19	4.24	4	0.89
3	8	3.08	1	0.39	1	0.39	1	0.39
4	3	1.96	1	0.65	1	0.65	0	0.00

Table 4: Uniqueness Lexicon

Corpora	Total tokens	Per 10,000	Tokens özel [special]	Per 10,000	Tokens for incomparable sub-lexicon	Per 10,000
1	28	10.98	13	5.10	13	5.10
2	68	15.19	31	6.92	32	7.15
3	31	11.94	23	8.86	8	3.08
4	18	11.09	8	5.22	8	5.22

Table 5: Tradition Lexicon

Corpora	Total tokens	per 10,000	Tokens gelenek* [tradition]	per 10,000	Tokens ev yap* [home make]	per 10,000	Tokens eski* [old]	per 10,000
1	52	20.40	5	1.96	11	4.31	8	3.14
2	150	33.50	15	3.35	5	1.12	42	9.38
3	60	23.12	2	0.77	2	0.77	27	10.40
4	21	13.70	4	2.61	2	1.30	7	4.57

Table 6: Tradition Lexicon Con.

Corpora	Tokens klasik* [classic]	per 10,00	Tokens yore* [region]	Per 10,000	Tokens tarih* [history]	Per 10,000
1	8	3.14	3	1.18	9	3.53
2	23	5.14	35	7.82	21	4.69
3	12	4.62	2	0.77	9	3.47
4	6	3.91	0	0.00	1	0.65

Table 7: Pricing Lexicon

Corpora	Total Tokens	Per 10,000	Tokens fiyat* [price]	Per 10,000	Tokens hesap* [bill]	Per 10,000	Tokens ode* [pay]	Per 10,000
1	286	112.19	118	46.29	14	5.49	28	10.98
2	363	81.06	152	33.94	19	4.24	45	10.05
3	197	75.90	81	31.21	25	9.63	16	6.16
4	103	67.20	54	35.23	2	1.30	3	1.96

Table 8: Pricing Lexicon Con.

Corpora	Tokens uygun [suitable]	Per 10,000	Tokens ucuz [cheap]	Per 10,000	Tokens paha* [pricey]	Per 10,000	Tokens lira* [lira]	Per 10,000
1	42	16.48	36	14.12	10	3.92	16	6.28
2	17	3.80	11	2.46	32	7.15	16	3.57
3	19	7.32	9	3.47	11	4.24	10	3.85
4	4	2.61	5	3.26	16	10.44	4	2.61

Table 9: Advice Lexicon

C	Total tokens	Per 10,000	Tokens -meli* all	Per 10,000	Tokens -meli* 2p	Per 10,000	Tokens -meli* 1p	per 10,000	Tokens -meli* 3p	Per 10,000
1	45	17.65	33	12.94	7	2.75	7	2.75	19	7.45
2	87	19.43	62	13.85	16	3.57	11	2.46	34	7.59
3	76	29.28	56	21.58	14	5.39	9	3.47	29	11.17
4	15	9.79	15	9.79	2	1.30	2	1.30	11	7.18

Table 10: Advice Lexicon Continued

Corpora	Tokens dene* [try]	Per 10,000	Tokens yarar* and fayda* [benefit, advantage]	Per 10,000
1	5	1.96	7	2.75
2	17	3.80	8	1.79
3	8	3.08	12	4.62
4	0	0	0	0

Table 11: Certainty Lexicon

C	Total Tokens	Per 10,000	Tokens mutlaka [must]	Per 10,000	Tokens kesin* [certain]	Per 10,000	Tokens lazım [necessary]	per 10,000
1	58	22.75	29	11.38	10	3.92	6	2.35
2	108	24.12	64	14.29	9	2.01	16	3.57
3	72	27.74	41	15.80	10	3.85	10	3.85
4	30	19.57	12	7.83	9	5.87	3	1.96

Table 12: Origin Language of Keywords

Word roots	**C1**	**C1 names**	**C2**	**C2 names**	**C3**	**C3 names**	**C4**	**C4 names**
Turkish	61.33%	37.78%	42.00%	30.00%	50.85%	22.95%	43.33%	21.43%
Arabic	21.33%	28.89%	26.00%	28.57%	18.64%	19.67%	18.89%	10.71%
Persian	8.00%	20.00%	8.00%	18.57%	10.17%	9.84%	1.11%	10.71%
French/ Italian	6.67%	6.67%	24.00%	8.57%	8.33%	11.48%	24.44%	32.14%
English	2.67%	4.44%	0.00%	12.86%	3.39%	13.11%	10.00%	21.43%
Greek	0.00%	2.22%	0.00%	1.43%	0.00%	22.95%	2.22%	3.57%
Total lemmas	75	45	50	70	59	61	92	28