FILOZOFICKÁ FAKULTA
UNIVERZITY KARLOVY
V PRAZE

Ústav anglického jazyka a didaktiky

Department of the English Language and ELT Methodology

Název práce: Individuální textový profil: korpusově založený výzkum idiolektu
Title of thesis: The individual textual profile: a corpus-based study of idiolect

DIPLOMOVÁ PRÁCE / MA THESIS

Bc. Marek Leško

Praha, Červenec 2012 **Vedoucí diplomové práce: PhDr. Markéta Malá, Ph.D.**
Prague, July 2012 **Supervised by: PhDr. Markéta Malá, Ph.D.**

„Prohlašuji, že jsem diplmovou práci vypracoval samostatně, že jsem uvedl všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu. Souhlasím se zapůjčením diplomové práce ke studijním účelům.“

“I declare that the following MA thesis is my own work for which I used only the sources and literature mentioned and that the thesis was not used in the course of other studies or to earn the same or other degree. I have no objections to the MA thesis being borrowed and used for study purposes.”

Místo, datum a podpis (place, date and signature): ....

Acknowledgment

I would like to thank PhDr. Markéta Malá, Ph.D., for her consistently sound advice, diligent supervision and a great amount of patience during all stages of the research and writing.

Abstract - EN

The thesis analyzes the idiolect of the then-presidential candidate Barack Obama in the specific speech situation of the televized debates on the background of the utterances of other candidates in the years 2000, 2004 and 2008. The analysis uses corpus-driven methods to compare the Obama corpus with the reference corpus. The comparison is largely based on the analysis of keywords and their use in context, supplemented by the discussion of their collocations and associated clusters. The results of the analysis, i.e. the principal features Obama's idiolect, are presented in a structured summary, divided into specific areas.

Keywords: Idiolect, individual textual profile, Obama, keywords, corpus, corpus-driven

Abstrakt - CZ

Diplomová práce analyzuje idiolekt Baracka Obamy jako prezidenstkého kandidáta ve specifické řečové situaci předvolebních televizních debat ve srovnání s jinými kandidáty v letech 2000, 2004 a 2008. Práce srovnává Obamův korpus s referenčím korpusem s využitím metod korpusové lingvistiky a toto srovnání je založeno zejména na analýze klíčových slov a jejich užití v daném kontextu. Práce zkoumá i kolokace klíčových slov a shluky často se opakujících slov. Výsledky analýzy, tedy hlavní rysy Obamova idiolektu, se prezentují ve strukturovaném přehledu rozděleném do několika oblastí.

Klíčová slova: Idiolekt, individuální textový profil, Obama, klíčová slova, korpusy

Contents

1. Introduction ... 6
1.1. Motivation for the Study of Idiolects ... 6
1.2. Motivation for the Corpus-Driven Approach ... 7
2. The Research Goals ... 9
3. The Methodology – a Theoretical Background ... 10
3.1. Keywords ... 11
3.1.1. Preliminary Definition – Narrowing the Scope ... 11
3.1.2. Historical Perspective ... 12
3.1.3. Classification of Keywords ... 12
3.1.4. Identifying Keywords ... 13
3.1.5. Statistical Measures ... 13
3.1.6. The Reference Corpus ... 17
3.1.7. Lemmatization and Tagging ... 19
3.1.8. Methods of Further Analysis ... 20
4. Summary of the Research Data ... 22
5. The analysis of Keywords in the Obama corpus ... 23
5.1.1. Proper Nouns ... 26
5.1.2. Lexical keywords ... 28
5.1.3. Functional style indicators ... 48
5.1.4. Keywords after Lemmatization ... 63
5.1.5. Parts-of-Speech Analysis ... 66
6. Summary of Findings ... 67
6.1. Summary of Findings: Methodology ... 67
6.2. Summary of Findings: Obama's Individual Profile ... 68
7. Summary in Czech (Shrnutí práce v češtině) ... 73
7.1. Cíl výzkumu a základní východiska ... 73
7.2. Metodologie ... 73
7.3. Výsledky ... 76

1. Introduction

1.1. Motivation for the Study of Idiolects

It is widely agreed that every individual uses a unique variety of language called an idiolect (Mollin 2009: 369). Because all language varieties are ultimately based on the varieties of the individual speakers within the group, some authors claim that "idiolects are the only kind of language that we can collect data on." (Haugen 1972: 415) Nonetheless, identifying and describing an idiolect on the basis of textual analysis is a more specific and perhaps more difficult task than a conventional stylistic analysis: whereas the latter attempts to identify the features of a particular linguistic variety, be it a register, dialect or other kind, usually on the basis of a considerable number of representative texts, the former has to successfully pinpoint traits pertaining to a single speaker or writer, which may greatly vary according to the medium, purpose and general situation. The current task is to describe this process and demonstrate the use of corpus-driven methods for the purpose of constructing an individual textual profile.

The following example might illustrate the impact of idiolect analysis outside of academic research. In 1996, an anonymous novel called *Primary Colors* was published in the United States, presenting a fictionalized account of what was widely considered thinly-disguised Clinton's presidential campaign in 1992. Due to the politically sensitive nature of the novel, the speculations over its authorship became quite popular and *Washington Post* even published a list of twenty five people who allegedly might have written the book. After months of discussions and denials, the chief suspect, Bob Klein, Newsweek columnist and a political reporter, finally admitted that he was indeed the author.[1] Even though the final proof was rather extralinguistic in nature (the graphological analysis of handwritten manuscript notes), besides the narrowing factor that the author must have had a lot of inside information, textual analysis or "literary forensics"[2], as one of the analysts called it, was the only initial lead.

[1] http://www.telegraph.co.uk/culture/books/8240675/The-impact-of-Primary-Colors.html

[2] http://articles.cnn.com/2000-12-06/entertainment/foster.anonymous_1_funeral-elegy-shakespeare-sleuth?_s=PM:books

1.2. Motivation for the Corpus-Driven Approach

The term corpus-driven approach implies the use of “corpora of texts”, in other words “genuine examples of language in use” (Scott, Tribble 2006: 3) as opposed to the alternative of intuitive analysis. However, to distinguish the corpus-driven approach from any other type of research that involves the use of corpora, it is necessary to refer to two distinctive features. First, the corpus-driven research is characterized by employing both qualitative and quantitative methods, unlike mostly qualitative research, which is more appropriately called corpus-informed. Second, the researcher generally approaches the subject with fewer preconceptions based on either specific theoretical frameworks or subjective linguistic intuitions (Lee 2008: 88). This chapter will briefly describe the history of the use of corpora and the advantages of this approach.

Despite the fact that corpus linguistics is a fairly recent subfield, the idea of a corpus is not completely new: For instance, in the 19-th century, Oxford English Dictionary was compiled with the help of a large number of slips containing samples of authentic language collected by its editors. Nonetheless, it was only in the last 50 years that the technological progress in both hardware and software enabled a large-scale automated text processing, paving the way to new techniques and approaches (Scott, Tribble 2006: 4). Some authors view these newly-opened possibilities as a significant impulse to the development of linguistics (Scott, Tribble 2006: 4) and the corpus-based research has become an attractive and rapidly-advancing subfield (Mollin 2009: 368). From a wider perspective, the development of corpus linguistics demonstrates a general tendency that can be characterized by an ever increasing interest in language in use as opposed to language as an abstract set of rules (Scott, Tribble 2006: 6, 7).

The advantage of a corpus-based approach over an intuitive analysis is the possibility to objectively quantify particular features by statistical means and provide robust evidence for our claims concerning language in use. The statistical approach can also reveal regularities that may otherwise elude the researcher’s attention (Hunston 2008: 271), or provide support for some intuitively known linguistic tendencies, such as the now notorious example of the negative semantic prosody of the verb *cause* (Stefanowitsch and Gries 2003). In the recent decades, corpora have been extensively used to characterize various registers and genres from academic English to the language of politics and poetry, but also specific literary works.

Corpus-driven methods are also employed for diverse pedagogical purposes ranging from improving the linguistic self-awareness of scholars to helping college students with their grasp of academic English.[3]

Finally, it is also worth mentioning that advances in computer technology revolutionized not only the processing of data, but also its presentation. Anecdotal evidence suggests various kinds of "infographics", visual representations of data, have become very fashionable (see, for instance, Daily Infographics[4] or David McCandless: The beauty of data visualization on TED Talks[5]). Important political speeches or debates (such as the ones that the current work aims to analyze) are also often represented by the so-called "word clouds" (see for instance Tag Clouds Evolve: Understanding Tag Clouds[6]), in which every word is scaled according to the frequency of its use. As an illustration, this is a word cloud representing what Barack Obama said during the three televised presidential campaign debates in 2008:[7]

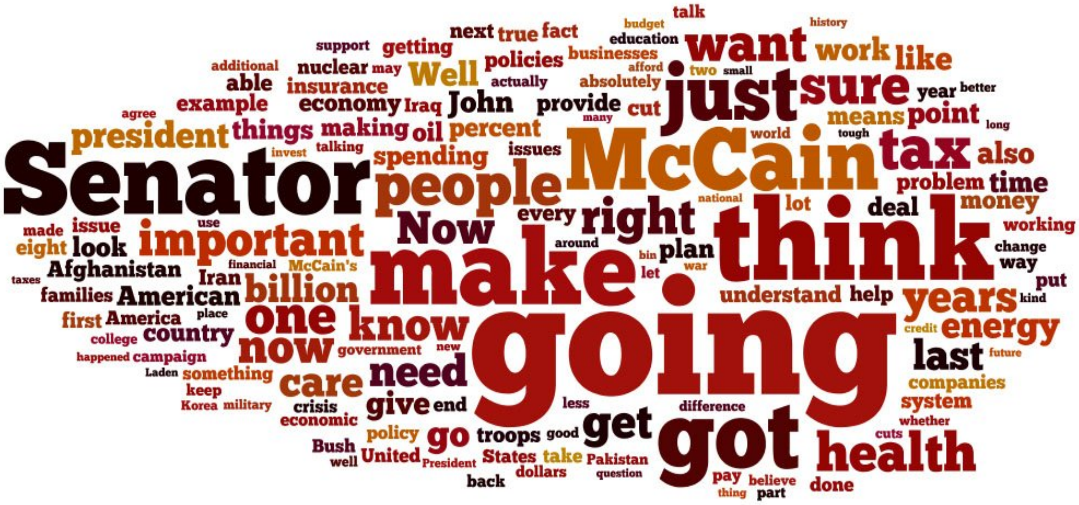

Figure 1: An illustration of a word cloud

From the perspective of corpus linguistics, word clouds are not the best tool: While it is possible to immediately recognize the most frequently repeated words, they offer very little beyond the visual representation and do not facilitate further qualitative research. For instance, the verb *going* is immediately visible as one of the largest items in the picture, but

[3] Cf. the references in e.g. O'Keefe, A. and M. McCarthy (2010) The Routledge Handbook of Corpus Linguistics.

[4] http://dailyinfographic.com/

[5] http://www.ted.com/talks/david_mccandless_the_beauty_of_data_visualization.html

[6] http://www.joelamantia.com/ideas/tag-clouds-evolve-understanding-tag-clouds

[7] This particular word cloud was generated by a free online generator called Wordle (http://www.wordle.net). Apparently, the articles, most prepositions and some auxiliary verbs are automatically excluded as trivial.

based solely on the word cloud, there is no way to know if its frequency is stylistically significant in comparison to other similar texts. Moreover, it is impossible to analyze the contexts in which the verb was used or determine whether the speaker used the verb in one of its lexical meanings or as an auxiliary verb in the construction *to be going to.* Despite these shortcomings and the naive methods behind their construction, word clouds demonstrate the increasing popularity of automated processing of texts and corpus-driven approach in the broadest sense of the term. It will be seen that some of the largest items in the illustrative word cloud above, in fact, do correspond to the keywords identified by more advanced and established statistical methods.

2. The Research Goals

The main aim of the present study is to construct an individual textual profile of a chosen person, in other words, to determine the principal features of that person's idiolect. This will be achieved by building and analyzing a corpus of texts produced by the subject. The person chosen for the discussion of the individual textual profile is President Obama in a singular speech situation, the official debates of the presidential candidates during the campaign in 2008. Because the speech situation is very specific and it should not be expected that all linguistic characteristics of Barack Obama as a presidential candidate are also manifest in other registers and situations, the term idiolect will be used in a narrow sense, i.e. Obama's individual profile in a series of three televised debates.

A politician was chosen primarily because of the necessity to acquire a large sample of texts with the implicit or explicit approval of the test subject. Unlike the texts produced by private persons, those of public figures are recorded, frequently transcribed and open to scrutiny, allowing a convenient and unproblematic access to data. Nonetheless, it is also necessary to mention some disadvantages of this approach: beside the issue of faithfulness of transcriptions, the speeches of public figures are often prepared in advance by a different author or authors. Mollin, who chose Tony Blair as the test subject, also noted these difficulties and admitted that "both preformulated speech and writing may indeed have been drafted by staff rather than by Tony Blair himself." (Mollin 2009: 371) Consequently, the individual characteristics of Barack Obama as a presidential candidate in televised debates may be removed from his day-to-day interactions to a considerable extent. Nonetheless, this does not invalidate the research because the focus is on the linguistic features as they appeared in the debates. In other words, the present work analyzes Obama's individual variety of the genre of a political debate. Finally, it should be emphasized that the motivation for this work is linguistic, not political, and references to President's Obama political personage will be limited to a minimum. The work does not aim to evaluate the rhetorical or argumentative persuasiveness of any candidates, let alone their actual proposals, promises and policies.

Before introducing the methodology behind the research, a few points should be made. Unlike corpus-driven research in general, the area of idiolects has been rather neglected (Mollin 2009: 368). Forensic linguistics is a notable exception, but even experts in this discipline admit that the methodology needs to be further developed (Mollin 2009: 369). However, there are several valuable studies focusing on identifying the linguistic features of individual speakers. The motivation varies: whereas some of them are mostly teaching-oriented, such as David Coniam's study on analyzing one's own academic prose (Coniam 2004), some are focused on developing adequate methodology, such as Mollin (2009), who attempted to characterize Tony Blair's idiolect by identifying his characteristic collocations, or Culpeper (2009), who demonstrated the use of keywords in the analysis of the language of the individual characters in *Romeo and Juliet*. Other articles, such as Semino and Swindlehurst (1996), are more focused on the literary analysis of a chosen work of art. Despite all these advances, some methodological questions, such as the choice and size of the reference corpus, are still far from decided (Scott, Tribble 2006: 64). For this reason, this thesis will also consider and evaluate some methodological decisions, such as the choice between lemmatized and unlemmatized corpora, and a separate chapter will be dedicated to the summary of findings pertaining to methodology.

3. The Methodology – a Theoretical Background

This chapter describes the methodological basis for this work. Because the analysis of Obama's idiolect relies on keywords, the first part is devoted to the definition of keywords along with a brief description of how they can be identified, classified and how they can serve the current purposes. Because a reference corpus is needed to perform a keyword analysis, some issues concerning its construction are discussed, followed by a justification of the final choice. Because compiling the list of keywords on the basis of lemmatized and unlemmatized corpora yields different results, this choice and its consequences need to be addressed before the proper analysis, as well as some technical details concerning tagging. Because the construction of the keyword list is only a starting point, a separate chapter is also dedicated to the methods of further analysis. Finally, at the end of **Chapter 3.1.6.**, a summary of the research data and the main parameters can be found to facilitate orientation.

3.1. Keywords

Since keywords are widely considered to be solid indicators of aboutness and style (Scott 2010: 43), the keyword list will serve as a central point and a basis for further discussion. Due to the complexity of the term keyword and a substantial amount of literature written on the subject, a separate chapter is dedicated to the topic of keyness, how a keyword list is obtained and what information can be obtained from it.

3.1.1. Preliminary Definition – Narrowing the Scope

The term *keyword* appears intuitive, but it is used in several different senses, which are only loosely related and to an extent incompatible (Stubbs 2010: 21). The first sense derives from cultural studies, where they represent the "focal point around which the entire cultural domains are organized" (Wierzbicka 1997:156). The second sense comes from quantitative corpus linguistics, where it refers to a word that is statistically prominent within a text in contrast to other texts (Stubbs 2010: 22). The third sense concerns clusters (phrases and schemas) that reveal culturally significant units of meaning (Stubbs 2010: 28). In the present work, the term *keyword* will be used only in the second, quantitative sense, due to the corpus-driven approach that has been chosen for this research. This entails a focus on the textual perspective, language as it is used in a selected discourse, as opposed to psychological, cultural or abstract grammatical aspects (Scott, Tribble 2006: 8). Before discussing specific methods of constructing the keyword list, a few notes should be made on the broad consequences of this choice. The obvious advantage of the quantitative approach is a level of objectivity that would otherwise be impossible due to an innate human predisposition to perceive texts in meaningful patterns, which could manifest itself as a bias towards perceptually salient keywords (Scott 2010: 45). The "blindness" of the automated processing is best viewed as a double-edged sword: on the one hand, it cannot replace qualitative analysis, but on the other hand, it can provide an objective basis for further discussion.

3.1.2. Historical Perspective

Keyword analysis is a relatively new method, due to difficulties with processing large volumes of text that were only resolved after technology with a sufficient computational speed and text-processing software became widely available (Scott, Tribble, 2006: 9). The first keyword analyses were performed more than 20 years ago, but the approach has become more widely used, which can be demonstrated on the array of articles dealing with keywords in romantic fiction, newspapers, gay and lesbian texts or spoken and written discourse (Culpeper 2009: 30). Besides keywords, Culpeper argues for analyzing key parts of speech and semantic domains in addition, because they help to reduce the number of categories that need to be investigated and group similar items that might be otherwise overlooked (ibid.).

3.1.3. Classification of Keywords

Before the proper analysis of keywords identified in the study corpus, it might be useful to review what kinds of keywords are likely to be found. Three classes are often identified: **(1) proper nouns**; **(2) lexical keywords** or **indicators of content**, often referred to as **keywords of "aboutness"** and **(3) grammatical words** that characterize a particular stylistic profile. (Glossary: 97)[8] However, this distinction is not without its problems. For instance, some researchers suggest that some discourse markers, such as vocatives, cannot be classified under (2) or (3), as they are peripheral to the syntax, and as such should perhaps represent a separate category (Culpeper 2009: 39). However, as will be seen in the later chapters, the greatest drawback of this classification is that it implicitly presumes that lexical keywords indicate the topic (or in the broader sense characterize the "aboutness" of a text). Because it will be demonstrated on particular lexical keywords that this is not always the case, the classification will be slightly adjusted. While Culpeper suggests that in order to solve this problem, the grammatical keywords should be a wider category that also includes lexical words which do not seem to characterize the aboutness of a text (Culpeper 2009: 39), the present work will use a different adjustment in order to keep the distinction between the lexical and functional keywords: the class of **lexical keywords** will be subdivided into the proper **indicators of content ("aboutness" keywords)** and **lexical indicators of style.**

[8] Glossary will from now on refer to Baker, Hardie and McEnery (2006)

Even though the first class of keywords, proper nouns, seem the least interesting to discuss, they should not be excluded from the analysis. For instance, Culpeper's analysis of keywords in Shakespeare's *Romeo and Juliet* revealed that while Romeo is frequently addressed by his name by other characters, he rarely refers to himself in this way, which arguably makes him "the fulcrum of the play." (Culpeper 2009: 38) The analysis of keywords in the Obama corpus will follow the distinction into the three classes as outlined in the previous paragraph.

3.1.4. Identifying Keywords

As has been mentioned, keywords are statistically significant words within a chosen text or text collection, the assumption being that if proper methods are employed, keywords provide solid information on the texts' aboutness and style. The essential feature of keywords is that they are not obtained from an isolated corpus, but rather by comparing the study corpus with a reference corpus. The reference corpus serves as a basis for expectations of the study text; in effect, it creates a certain linguistic norm from which the studied samples can deviate. Therefore, keywords are more precisely defined as lexical or grammatical units that occur at a significantly different rate in the study corpus than in the reference corpus. (Scott, Tribble 2006: 59) It should be noted that neither the reference corpus nor the specific statistical measures of this deviation are given and their choice deserves a thorough discussion.

3.1.5. Statistical Measures

The simplest measure of statistical prominence is the raw number of occurrences in the study corpus compared to the number of occurrences in the target corpus. In the keyword analysis, however, this measure is clearly not a sufficient one. To illustrate this point, Barack Obama used the word *president* 55 times during the television debates before the elections in 2008. Assuming that the reference corpus comprises all utterances of the other candidates, the word occurs 556 times in this selection of texts. This would indicate that Mr. Obama is much more reluctant to refer to his future office than the other candidates, but this comparison is invalid due to different sizes of the corpora. It is possible to compensate for the difference by using normalized frequencies, i.e. the number of occurrences per million words, but this does not solve the core issue of how to interpret the two frequencies of occurrence in the target and the background corpus. A single variable that would adequately express the measure the statistical significance of a word is clearly needed to create a keyword list.

Therefore, several ratios and variables have been proposed as the optimal measure of statistical significance, such as **chi-squared**, **log-likelihood** or **Fisher's Exact Test** (Glossary: 31). Software products intended for corpus-based analysis, such as **WordSmith Tools**, have also adapted to the situation and offer a substantial statistical output. Culpeper (2009) used log-likelihood as the primary test, but confirmed his results by a chi-squared test. Because log-likelihood is based on the chi-squared and can be considered a refinement, the chi-squared test will be discussed first. It should be noted that these two statistics are widely seen as canonic for keyword analysis (Glossary: 97) and under normal circumstances they should produce almost identical results (Oakes 1998: 38). For these reasons, this thesis will also focus on chi-squared test (or chi-squared value) and log-likelihood as the statistical measures of significance, the first of which will be discussed in greater detail to illustrate the calculation.

Chi-squared[9] Test

This test is based on a statistical approach called hypothesis testing, the essence of which is to formulate a null hypothesis H_0, calculate the probability that a tested outcome would occur if H_0 were true and if the probability is too low, typically below 5% or 1%, the null hypothesis is rejected as improbable. (Manning and Schutze 1999: 163) In the case of the chi-squared test, the null hypothesis is that the difference between the observed frequencies of occurrence (i.e. frequencies in the target corpus) and the expected frequencies (i.e. frequencies in the reference corpus) is completely random. (Manning and Schutze 1999: 169). The test is performed by calculating the value of chi-squared for a 2x2 table, which lists the number of occurrences of a tested word and all other words in the target and the reference corpus. **Table 1** illustrates the method using the word *McCain* as an example.

	Number of occurrences	
Word	**Target corpus**	**Reference corpus**
McCain	99	110
[All Other words]	22 151	146 515

Table 1: An illustration of a 2x2 table used for calculating the value of chi-squared

The test of chi-squared can be applied to a table of any size, but the reduction to a 2x2 form is used for two reasons. First, the goal is to compute a measure of statistical significance for each individual word rather than to compute the statistical significance of the overall difference between the target and the reference corpus, which would be the result if all the

[9] In some sources, such as Manning and Schutze (1999), it is referred to as "chi-square test". For the purposes of simplicity and consistence, this thesis will always use the form "chi-squared".

words except for the tested one were not aggregated as in **Table 1** but listed on separate lines. Second, the calculation of chi-squared is much simpler for 2x2 tables (Manning and Schutze 1999: 169):

$$\chi^2 = \frac{N(O_{11}O_{22} - O_{12}O_{21})^2}{(O_{11} + O_{12})(O_{11} + O_{21})(O_{12} + O_{22})(O_{21} + O_{22})}$$

where *N* is the total number of words in both corpora and O_{ij} represents the number in line *i* and column *j* in the table. In the chosen example, the value of chi-squared for the word McCain in the corpora described in **Table 1** is 213.87.

For every table size and every chosen probability level *p*, there is a critical value of chi-squared. If the calculated value is higher than the critical, it is safe to reject the null hypothesis within the margin of error expressed by the probability level (ibid.).[10] For the purposes of this test, rejecting the null hypothesis means that the discrepancy between the observed and the expected frequency of occurrence is statistically significant. Following the example of *McCain*, the value of 213.87 is approximately 32 times greater than 6.63, the value corresponding to 1% probability level. Therefore, it is possible to conclude with a more than 99% certainty that *McCain* is a keyword in the target corpus.[11] It should be noted that the calculation can be reversed: for every calculated value of chi-squared, there is a probability level *p,* which can be interpreted as the statistical probability that the discrepancy in occurrences was a coincidence. Therefore, to rank the words according to their value of chi-square in the descending order is the same as to rank them from those most likely to be keywords to the least likely. In other words, the value of chi-squared can be used as the measure of statistical significance of keywords.

[10] The critical values can be derived mathematically, but this area of statistics goes well beyond the needs of this work. The critical values are fairly accessible on the internet and the appendixes of statistical handbooks, such as Manning and Schutze (1999).

[11] While the example is at this point intended only as an illustration, the data come from the actual corpora analyzed in this work. This demonstration also serves as a test if the software used for the analysis of keywords, WordSmith Tools, calculates the value of chi-squared in the same way. Except for a small discrepancy (less than 1% in the value of chi-squared), which can be attributed to a so-called “Yates Correlation” according to the electronic manual, WordSmith Tools produces the same results as the manual calculation.

The disadvantage of chi-squared is that it is has been observed to be unreliable with very low frequencies, in which case the log-likelihood test is preferred (Glossary: 31). Finally, as will be seen in later chapters, WordSmith Tools can automatically identify negative keywords[12] and assign them negative values as a matter of convenience despite the calculation method, which can clearly provide only zero and positive values.

Log-likelihood

Log-likelihood statistic, sometimes referred to as G-square or G score was developed by Ted Dunning in the early 1990's and strongly resembles chi-squared, on which it is based. (Dunning 1993: 61 – 74). It can be calculated as $G^2 = 2\sum_{i=1}^{2} O_i(log_e O_i - log_e E_i)$, where E_i is the expected occurrence of a given word, which is calculated on the basis of the respective corpora sizes (Glossary: 110). Similarly to chi-squared, the value of log-likelihood is calculated for individual words with the help of a 2x2 table and the higher the value, the less likely it is that the difference between the observed and expected rate of occurrence is a mere coincidence. The calculation of log-likelihood was also emulated manually and the results were identical to those obtained from WordSmith Tools. There seems to be consensus among researches that log-likelihood provides very reliable results as a measure of statistical significance (Rayson: 2003), perhaps because it provides better results with low-frequency words (Glossary 31).

Probability Level (*p*)

Even though it is customary and acceptable in social sciences to work with the level of probability of 5%, the cut-off point is sometimes set as low as 0.0001% (Culpeper 2009: 36) to obtain fewer keywords with greater certainty. It is important to note that an arbitrary cut-off point is necessary because unless a word has exactly the same relative frequency in the study corpus and the reference corpus, it will display some degree of keyness. Without setting a cut-off point, to make a keyword list would be the same as to sort all the words from the corpora according to the chosen measure of keyness in the descending order. Therefore, setting limit on the value of p is in effect the same as considering only the first n keywords from the list ranked according to their value of log-likelihood or chi-squared. An alternative approach is to review a larger number of ranked keywords and manually select relevant items

[12] Negative keywords are units that occur with a significantly lower frequency than expected on the basis of the reference corpus.

for further discussion. Finally, a frequency limit for individual words is sometimes set in order to exclude low-frequency words that are less representative of the sample and also tend to be clustered in a small chunk of the text. At the very least, it is considered a good practice to list the raw frequencies of the keywords as well (Culpeper 2009: 40).

Conclusions on Statistical Measures

In the present work, log-likelihood was used as the primary measure of statistical significance because it is generally favored by the current research (Culpeper 2009: 36) and as has been mentioned, it is more appropriate than chi-squared when used on low-frequency words. The probability level of 0.001% (i.e., $p = 0.00001$) was used, which limited the number of number of keywords to 37, a reasonable number for a detailed qualitative analysis. The threshold of at least five occurrences was considered to exclude very rare, unrepresentative units,[13] but all the 37 keywords were used five times or more, which rendered this criterion unnecessary.

3.1.6. The Reference Corpus

The list of keywords is not a product of an isolated analysis of the chosen text. Rather, it is based on the contrast between a chosen text and the reference corpus. The choice of the reference corpus is, therefore, essential, but as Culpeper notes, there is no magic formula for this decision (Culpeper 2009: 31). In general, a reference corpus represents "the general nature of the language through a wide-sampling corpus design" or, as in the present case, "the basis of comparison ... drawn from a wider range of ... sources" (Glossary: 137) Scott and Tribble (2006: 58) state that it "should be an appropriate sample (preferably many thousands of words long and possibly much more) of the language which text we are studying (the 'node text') is written in."

It follows from the definition of the reference corpus and the methodology of finding keywords that the choice of the reference corpus has to be based on the research objectives. For instance, the Corpus of Contemporary American English (COCA)[14] would be a legitimate choice if the research goal was to compare the linguistic characteristics of Obama's debating with the general spoken and written American English. It should be expected that this comparison would yield very different results from the contrastive analysis

[13] Scott, Tribble (2006) suggest the limit of at least two or three occurrences.

[14] http://corpus.byu.edu/coca/

of Mr. Obama against general American English, American political discourse or the speech of other presidential candidates.

Because the current research aim is to study idiolects, the target corpus and the reference corpus should be contextually as close as possible. The reason lies in the very definition of an idiolect, a linguistic variety of an individual speaker: The greater the contextual distance between the target and the reference corpus, the more likely it is that the purportedly individual characteristics of a chosen speaker are in fact attributable to a register or genre. In other words, "the closer the relationship between the target corpus and the reference corpus, the more likely the resultant keywords will reflect something specific to the target corpus." (Culpeper 2009: 35) Therefore, to maximize the contextual closeness of the corpora and thus the relevance of keyword analysis for determining Mr. Obama's idiolect, the reference corpus comprises only utterances of other candidates in an identical speech situation, i.e. the presidential campaign debates from the same decade (elections in 2000, 2004 and 2008).

The main advantage of this approach is the above-described contextual closeness and consistency of register. The main disadvantage is, of course, the limited size of the corpus, which in itself decreases its representativeness and may introduce undesirable statistical anomalies. If an alternative approach was taken, the reference corpus could be extended by adding samples from a similar genre, such as older campaign debates, State of the Union Addresses, or other political debates and interviews. However, the question of what constitutes a register or speech-situation similarity is open to debate and the issue of balancing the corpus can be particularly complex. It should be noted, however, than Mollin (2009) shows that it is still possible to abandon all consideration of genre closeness and use a general reference corpus, such as BNC, even for the study of an idiolect.

To conclude, a consistent small corpus was preferred to a larger but more diverse set of texts. This approach is supported by, for example, Xiao and McEnery (2005: 71), who suggest that "the size of the reference corpus is not very important in making a keyword list" because they obtained almost identical results using two reference corpora of significantly different scopes.

3.1.7. Lemmatization and Tagging

Lemmatization is a form of automatic annotation which reduces the words in a corpus to their lexemes, i.e. canonical forms. (Glossary: 104, 105). This allows the researcher to evaluate the frequency of use and distribution of a lexeme without having to manually search for all possible forms. (ibid.) Because the keyword lists based on lemmatized and unlemmatized corpora may differ (and it will be shown that in the present case they in fact do), the choice between lemmatized and unlemmatized corpora represents a methodological dilemma which requires a review of advantages and disadvantages of each approach.

The main reason for choosing unlemmatized corpora is that contractions and inflected forms are also significant from the perspectives of a stylistic analysis and idiolect. For instance, the form *we've*, which obviously cannot be considered a single word, is nonetheless a valid item for analysis. Its use is significant from the perspective of discourse formality or it may at least represent a linguistic habit. However, it must be admitted that there are potential problems with this approach: for instance, if *we've* were always contracted in the Obama corpus but always transcribed as *we have* in the reference corpus, *we've* would be misleadingly identified as a positive and *we* and *have* as negative keywords, even if the normalized frequencies of *we've* and *we have* were the same.[15] This may be undesirable if the focus is on the use of the pronoun and the verb themselves regardless of contractions. Fortunately, because the target and reference corpus are from the same source and the transcription appears consistent in transcribing contracted forms as such, the use of contractions can be legitimately analyzed as a part of the studied idiolect. There is, however, a more serious problem with using unlemmatized corpora as the basis of analysis. This issue can be illustrated on the hypothetical example of the pronoun *he*, which can occur in the forms *he, his* and *him.* If the keyword analysis treats the three forms as independent items, their somewhat higher rate of occurrence might appear as statistically insignificant. However, if the corpora were lemmatized and the three forms evaluated as a single keyword, the pronoun would appear much more prominent. Consequently, while the keyword analysis is primarily performed on unlemmatized corpora because the inflected and contracted forms are

[15] For the sake of the argument, it is assumed that the words *we* and *have* occur only in the construction *we have,* which is, of course, not the case. The unrealistic assumption was used for the illustration of relationship between positive and negative keywords. In reality, the fact that *we* and *have* occur independently in various other constructions might be the reason why they did not show up among negative keywords.

significant for the idiolect analysis, a keyword list compiled on the basis of lemmatized corpora is included as well and some important differences are commented upon there.

A part-of-speech tagging is a method of automated annotation, which assigns each word some grammatical categories (Glossary: 128). The POS tagging is a useful enhancement of a corpus as it provides additional information and search options, however, even the best tagging algorithms invariably produce mistakes. The current work uses a freely available TreeTagger,[16] which uses the Penn-Treebank tagset.[17] It should be noted that most of the work was performed on unlemmatized and untagged versions of the corpora and these methods of automated processing should be viewed as supplementary in the current framework.

3.1.8. Methods of Further Analysis

Even though the list of keywords has in itself a certain value, much more can be learned by a closer look at the identified units. A detailed analysis can reveal a specific context, distribution, semantic prosody or collocations of the chosen keyword. Moreover, it may show how keywords are linked together through a topic, context or a recurring phrase. Such an analysis is even more important when the list of keywords is based on non-lemmatized corpora due to the inherent fusion of lexical and grammatical preferences, such as in the case of *here's* or *policies.* Moreover, phenomena such as polysemy or homonymy (see the discussion of *going* or *make*), which are abundant in English and which might distort the keyword list, also necessitate a detailed analysis. In the present work, three basic methods have been used for further investigation of the identified keywords: **concordance, collocations** and **clusters.** While other techniques have been used as well, these three have been consistently employed as the basic tools and merit a brief introduction.

1) Concordance

"A concordance is a list of all of the occurrences of a particular search term in a corpus, presented within the context in which they occur – usually a few words to the left and right of the search term." (Glossary: 42) Because concordance shows all the uses of a chosen

[16] Available at http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

[17] The description of the tagset is available at http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/Penn-Treebank-Tagset.pdf

keyword in context, it is the most thorough method of analysis. In the present work, concordance was used to verify and correct the results of other methods. For instance, due to a sequence of false starts *just – just – just a quick follow-up,* which occurred twice in the Obama corpus, the last part *a quick follow-up* is mistakenly identified as a cluster that frequently occurs after *just.* Moreover, looking at the concordance lines is very useful when there are only a few uses of the keyword because it is possible to quickly identify all the uses and contexts of the item.

2) Collocations

Collocation is defined as "the tendency of words to be biased in the way they co-occur. … It can be considered as the tendency of two words to co-occur, or as the tendency of one word to attract another." (Hunston 2002: 68) It can be also understood in terms of statistically significant association of words (Glossary: 36). Even though linguists have developed a number of techniques to measure the frequency and exclusivity ("the strength") of the association (Glossary: 37), the present work identifies the collocations using the simplest measure of relative frequency of co-occurrence. Even though more sophisticated methods exist, such as mutual information, Z score or log-likelihood, there are two practical reasons for this choice. First, the goal here is not to measure the strength of collocations, only to quickly identify the words with which the chosen keyword co-occurs most often to discover frequent associations. Second, the analysis of collocations is limited by the small size of the corpora. For instance, if the combination of a keyword plus its most frequent collocations (in terms of absolute numbers) only occurs a few times in the entire corpus, using more sophisticated methods would often result in collocations that are statistically valid but only occur once or twice in the entire corpus, which is clearly not sufficient for any meaningful conclusions on the idiolect of Barack Obama. It should also be added that only lexical collocations are usually listed in the tables of collocations. The reason is that significant grammatical associations, i.e. colligations (Glossary: 36), are better captured as clusters because the combination of a keyword and an article or a preposition may in itself be very common and not particularly interesting from the perspective of an idiolect or stylistic analysis. For instance, the first two collocations of *time* in the Obama corpus according to the relative frequencies are the articles *the* and *a,* which is of little relevance for the idiolect analysis because these results should be expected on the basis of the English grammar. However, using the example of *time*, the cluster *for the first time (5)* is much more distinctive and therefore, imporant for the idiolect analysis. In the present work, the collocations were searched for in the interval four words to the left to four words to the right of the search term in order to include non-immediate or discontinuous associations.

3) Clusters

Clusters, also known as lexical bundles or n-grams, are defined as recurring sequences of words (Biber et al. 1999: 989). The minimum length is usually considered three words and the sequences have to be continuous, but not necessarily structurally complete (Biber et al. 1999, 989 - 990). "Recurring" refers to a minimum frequency of occurrence per million words in a given corpus and this limit is usually set to 10 – 40 (ibid.). A more detailed theoretical overview of clusters goes beyond the scope of this work, but it has been demonstrated in many studies that clusters can be used to characterize various genres and registers (see Biber and Barbieri 2007 or Hyland 2008). In the present work, clusters will be used in combination with keywords by looking at the sequences of at least three words associated with the particular keyword. The term "associated" means that the keyword does not necessarily have to be a part of the cluster, but has to occur within the distance of three words. It is hoped that this analysis will provide not only additional details about the keywords, but also clusters that will prove unique to or at least characteristic of Barack Obama.

4. Summary of the Research Data

Before the keyword list and qualitative analysis itself, this short chapter is dedicated to a summary of the data used in this research. The corpus that will serve as the basis for the current analysis of Mr. Obama's idiolect comprises all his utterances during the three presidential debates in 2008 broadcasted on the national television in the United States, obtained from the official source, Commission on Presidential Debates.[18] The utterances of his opponent, Senator McCain, the moderator, the public and the notes found in the official transcript (e.g., *[silence]* or *[laughter]*) have been deleted, so that this corpus consists of Barack Obama's speech only.

[18] The transcripts can be found on the address http://www.debates.org/index.php?page=debate-transcripts

The background corpus comprises the utterances of John McCain in the three televised debates where Barack Obama was also present as his opponent. Additionally, it includes the utterances of all candidates for the office of president and vice-president in all televised debates in the election years 2000, 2004 and 2008[19]: Bush vs. Gore, Bush vs. Kerry, Lieberman vs. Cheney, Cheney vs. Edwards and Palin vs. Biden. The questions from the moderator and the audience as well as the transcript notes have also been edited out.

The table below summarizes the research data and parameters used in this work:

	Target Corpus	**Reference Corpus**
Number of sources (debates)	3	12
Number of words (tokens)	22 013	145 266
Number of unique words (types)	2 539	6 924

Keyword analysis – primarily performed on unlemmatized corpora	
Measure used	log-likelihood
Probability level	0.001%
Keywords at the prob. level	37

Table 2: The summary of research data

5. The analysis of Keywords in the Obama corpus

At the probability level of 0.001%, 37 keywords have been identified using log-likelihood as the measure of keyness, which seems a reasonable number for further discussion and analysis.[20] All the identified keywords occurred five times or more, which is sufficient to consider them valid indicators rather than anomalies (see Chapter 3.1.5.). The complete list follows, sorted according to the value of keyness.

The column labels refer to:

N	– the rank of the keyword
Keyword	– the keyword identified in the text
# in OC	– the number of occurrences in the Obama corpus
% in OC	– the frequency of occurrence in the Obama corpus as a percentage

[19] There are three debates for the presidential candidates and one debate for their so-called running mates (the candidates for the office of vice-president).

[20] The list of keywords using chi-squared as the measure of keyness yielded virtually the same results. As log-likelihood is the preferred measure when the frequencies are rather low, the following analysis considers only the keyword list based on log-likelihood.

# in RC – the number of occurrences in the reference corpus

% in RC – the frequency of occurrence in the reference corpus as a percentage

Keyness – the strength of statistical significance expressed as the log-likelihood value

N	Keyword	# in OC	% in OC	# in RC	% in RC	Keyness
1	MCCAIN	99	0.45	110	0.08	143.75
2	SENATOR	115	0.52	212	0.14	102.49
3	THAT	688	3.1	3,211	2.19	64.77
4	GOING	181	0.81	603	0.41	57.46
5	GOT	108	0.49	307	0.21	49.03
6	WE'VE	94	0.42	251	0.17	48.13
7	MAKE	119	0.54	367	0.25	45.41
8	ARE	205	0.92	776	0.53	45.13
9	TRUE	25	0.11	24	0.02	40.3
10	EMPLOYER	14	0.06	4	0	38.86
11	POLICIES	25	0.11	26	0.02	38.09
12	POINT	35	0.16	56	0.04	36.55
13	PROVIDE	27	0.12	33	0.02	36.28
14	MCCAIN'S	18	0.08	13	0	34.53
15	ENERGY	44	0.2	95	0.06	31.79
16	SOME	81	0.36	247	0.17	31.67
17	IS	357	1.61	1,698	1.16	29.88
18	WE	458	2.06	2,271	1.55	29.79
19	FINANCIAL	13	0.06	7	0	28.82
20	DEAL	29	0.13	52	0.04	26.66
21	LAST	50	0.23	132	0.09	26.11
22	NOBODY	13	0.06	9	0	25.51
23	POTENTIALLY	6	0.03	0	0	24.34
24	SURE	63	0.28	194	0.13	24.08
25	CRISIS	20	0.09	28	0.02	23.84
26	PAKISTAN	17	0.08	20	0.01	23.56
27	IMPORTANT	56	0.25	168	0.11	22.68
28	ADDITIONAL	16	0.07	20	0.01	21.09
29	JUST	93	0.42	350	0.24	20.78
30	OIL	31	0.14	71	0.05	20.52
31	ADVISERS	5	0.02	0	0	20.28
32	MUDDLE	5	0.02	0	0	20.28
33	HERE'S	17	0.08	24	0.02	20.1
34	SO	117	0.53	476	0.32	19.98
35	IRAN	25	0.11	51	0.03	19.53
36	I	376	1.69	3,140	2.14	-20.13
37	THE	890	4.01	7,620	5.2	-60.57

Table 3: Keywords in the Obama corpus

Several points can be made before the classification and analysis. All three classes discussed in Chapter 3.1.5. (**proper nouns**, **indicators of content** and **grammatical keywords**) that characterize a textual profile are represented in the list. There are two negative keywords. Negative keywords, i.e. words that are avoided by the speaker from the perspective of what might be expected based on the reference corpus, are generally less numerous than the positive counterparts. Culpeper explains this tendency by the fact that it is easier to exceed the norm, particularly with a small target corpus, than to go far below the expectations. (Culpeper 2009: 38) In other words, all frequent units are used by necessity at the expense of some others, but those other words are likely to be evenly distributed.

Three of the identified keywords (*muddle*, *advisers* and *potentially*) have zero occurrences in the reference corpus, which may suggest a disparity between the preferences of Barack Obama and the other candidates. However, because their frequency in the Obama corpus is also rather low (5 – 6 occurrences), they did not rank very high on the keyword list.

Because the keyword list was compiled on the basis of unlemmatized corpora, two "keywords" in fact represent contracted forms (key-phrases): *we've* and *here's.* The lack of lemmatization also caused the proper name *McCain* to appear both in the basic and the possessive form. Moreover, *policies* and *advisors* were identified as keywords only in plural. The discussion of these non-canonic forms is included in the analysis of individual keywords on the list.

The following table shows the keywords divided into three basic categories, **proper nouns**, **lexical keywords** (i.e., **lexical indicators of content** and **style**) and **grammatical keywords (i.e., functional style indicators),** ordered according to keyness as expressed by log-likelihood value. Negative keywords are marked with (-). It should be noted that this classification is based on the predominant use of the words in the Obama corpus. For instance, *going* can function as a participle or gerund form of a lexical verb to go, but as shall be seen, in the majority of instances it was used as an auxiliary verb in the future construction *to be going to*.

Proper nouns	**Lexical keywords** (lexical indicators of content and style)	**Grammatical keywords** (functional style indicators)
McCain, Senator, McCain's, Pakistan, Iran	*make, true, employer, policies, point, provide, energy, financial, deal, last, potentially, sure, crisis, important, additional, oil, just, advisers, muddle*	*that, going, got, we've, are, some, is, we, nobody, here's, so, (-)I, (-)the*

Table 4: Keywords divided into three basic classes

The keyword list confirms Culpeper's claim that a qualitative analysis is needed before any meaningful conclusions can be reached. For instance, the verbs *make* is so versatile that a concordance is needed in order to reveal in what constructions (causative, phrasal, idiomatic or simple lexical) it appears so often. The same applies for verbs that can be used as lexical or auxiliary, such as *going*, or words which can be used as nouns or verbs, such as *deal.* The qualitative analysis will follow for each major category of keywords, which should lead to a basic notion of the individual stylistic profile of Barack Obama as a candidate in television debates.

5.1.1. Proper Nouns

The categorization of the keyword *Senator* as a proper noun requires some justification. It is certainly not a proper noun as such, but the concordances reveal that in all 115 occurrences, it is used as a title in the phrase *Senator McCain* or *Senator McCain's*,[21] which makes it a bound part of a proper name in the Obama corpus. *McCain* and the corresponding possessive form *McCain's* are the most frequent proper noun keywords. This finding can appear quite significant, however, it can be demonstrated that referring to the opponent by the proper name is a part of the etiquette. In this debate format, the candidates refer to each other in third person even in situations that would in an ordinary conversation elicit a response in the second person: *Senator McCain continues to repeat this; this is a major difference between Senator McCain and myself (OC); I don't believe Senator Obama has the knowledge; the*

[21] 99 occurrences of *McCain* and 18 occurrences of *McCain's* add to 117, which is still two more than 115 uses of the title *Senator.* The reason is that in two instances, Barack Obama refers to his opponent as *John McCain,* not *Senator McCain.*

proposal that we have from Al Gore, basically, doesn't do that (RC).[22] On the other hand, this is not an absolute rule and candidates sometimes refer to each other directly: *It's not like you want to close the loopholes; And you know, Senator McCain, I think the "Straight Talk Express" lost a wheel on that one (OC); The first time I ever met you was when you walked on the stage tonight; And Senator, frankly, you have a record in the Senate that's not very distinguished. (RC)* It can also be shown that in the course of the debates, Senator McCain speaks about Barack Obama even more than the other way around (161 to 117). To conclude, the proper noun keywords *(Senator) McCain* and its possessive form are characteristic of the debate format rather than the idiolect of Barack Obama, even though a direct address is apparently permissible in the campaign debates.

Because these proper nouns are fairly numerous, it is reasonable to look at their collocations. The following table lists the most frequent general lexical collocations of the name Obama in the Reference Corpus and McCain in Obama corpus:

***McCain* in the Obama corpus**	***Obama* in the Reference Corpus**
Senator (97), mentioned (6), agree (3) voted (3), talks (3), keeps (3), opposed (2), says (2), wants (2), suggested (1)	*Senator (107), voted (12), said (11), wants (8), says (4), talks (3), mandates (2), opposes (2), supported (1)*

Table 5: Lexical collocations of *Obama, McCain*

These collocations are remarkably similar, which suggests that the candidates use the opponent's name to attribute to them certain intentions, statements or actions as a general strategy. However, there are two notable differences.

The first consists in the grammatical structure *McCain keeps on saying/making/using,* which is only used by Barack Obama, presumably to portray the claims of his opponents in the negative light as repetitive. However, in the six clauses in the Obama corpus with the predication *[to keep on] + ing form,* John McCain is the subject only in half of the instances, specifically in those where the verb is in the third person singular. In the remaining three instances, the subject is *we,* and while the semantic preference varies, the structure retains a negative semantic prosody due to associations with unsustainability or ineffectiveness, as in *we can't keep on borrowing from the Chinese (OC).* While the phrase *[to keep on] + ing form* was used only six times (three times as *keeps on* and three times as *keep on*) by Barack

[22] The abbreviation OC stands for the Obama corpus and RC stands for the Reference corpus. When these abbreviations are in parentheses, they mark the source of the keyword, collocation, cluster or other quotation.

Obama, there is not a single instance of it in the Reference Corpus, which is approximately seven times as large. This suggests an idiosyncratic preference for the structure *[to keep on] + ing form* as opposed to a synonymous structure *[to keep] + ing form* (Longman Dictionary of Contemporary English: keep on), which can be found frequently in the Reference Corpus but not the Obama corpus.

Second, Barack Obama expresses explicit agreement with his opponent, using phrases such *Senator/John McCain and I (do) agree.* This tendency can be confirmed by looking at the most frequent clusters associated with the keyword *McCain,* which include *Senator McCain and I* (thirteen uses, nine explicit agreements and four instances of distancing oneself or disagreement). While there are eight instances of *Obama and I* in the Reference corpus, this phrase was uttered by Obama's running mate six times out of eight. Based on these data, a provisional conclusion might be reached that Barack Obama uses a more inclusive language than his opponent and tends to emphasize agreement.

As for Pakistan and Iran, they are not necessarily the countries that Barack Obama referred to the most (of countries of interest,[23] Iraq has 22 occurrences, Afghanistan 26 and United States 21 in the Obama corpus), but these two have been identified as keywords because the other candidates speak about them disproportionately less. While this does reveal something about his plans concerning foreign policy, it is also clear that the use of proper nouns is extremely dependent on the topic *(Iran, Pakistan)* and the particular immediate situation *(Senator McCain)*. While they are unlikely candidates for permanent idiosyncratic features of any chosen speaker, they may serve as starting points for the identification of more significant patterns, such as those mentioned above *(McCain keeps on, Senator McCain and I*).

5.1.2. Lexical keywords

Even though the **indicators of content**, or **aboutness keywords**, have been introduced as a single category, it is arguably more appropriate to distinguish between **lexical topic indicators** and **lexical indicators of style**. The former group includes words that indicate specific subjects: in the present case, campaign topics that were accentuated more by Mr.

[23] Because it would be impractical to verify the occurrences of a more than a hundred state entities in the Obama corpus and the reference corpus, a few countries that are the most likely to be discussed with regard to the current American foreign policy were selected on the basis of common knowledge.

Obama than by the other candidates, such as *employer, energy,* or *oil.* Because these items represent the key topics of the analyzed speaker, the label "aboutness keywords" seems quite appropriate. The latter category consists of general lexical items that are not suggestive of any particular political, social or economic topic, such as *make, true,* or *point.* This demonstrates that it would be misleading to identify lexical keywords with indicators of content and that the traditional distinction between open-class keywords pointing to aboutness as opposed to closed-class keywords revealing something about style is very rough at best.[24] Therefore, lexical keywords in the Obama corpus will be treated as two separate categories instead, only the former of which can be properly called **aboutness keywords:**

Lexical Topic Indicators	**Lexical Indicators of Style**
energy, financial, crisis, oil, employer	*make, true, policies, point, provide, deal, last, potentially, sure, important, additional, advisers, just, muddle*

Table 6: Aboutness keywords in the Obama corpus

Lexical Topic Indicators

Employer

The word *employer* occurs in the context of taxes and health insurance in all 14 instances. Since *tax* and *taxes* occur 72 times in the Obama corpus, and *health* is referred to 63 times, and yet none of these items is a keyword in the Obama corpus, it seems that Obama does not emphasize these topics more than the other candidates do but merely emphasizes the aspect of employers as those responsible for paying taxes and providing health care. Moreover, 50 per cent of the occurrences are bound in the phrase *employer-based (health care system/plan),* which is used only once in the reference corpus. For this reason, Obama's frequent use of the noun *employer* could be regarded as linked to the phrase *employer-based health care system/plan.* It is also clear that it is this plan that should be properly regarded as a topic rather than the employers themselves, which demonstrates the need of qualitative analysis of individual keywords.

[24] This distinction is discussed in Culpeper (2009: 39) or Scott, Tribble (2008: 143).

Energy, Oil

The appearance of *energy* and *oil* among keywords is in accordance with the general perception that the sources of energy were an important campaign topic in the 2008 elections, particularly for Barack Obama. However, the comparison of the most important immediate lexical collocations for the keyword *energy* in the OC and RC shows that the word occurred in very similar contexts:

Obama corpus	**Reference Corpus**
alternative (6), policy (5), independence (4), consumption (4), solar (3)	*Independence (12), independent (4), policy (8), plan (6), domestic (5), alternative (5), resources (3),*

Table 7: Lexical collocations of *energy*

Besides the quantitative differences caused by the different sizes of the corpora, the other candidates speak of both *policies* and *plans* with regard to *energy,* whereas Obama prefers the former term. More importantly, Obama speaks about *energy consumption*, in all instances in the context of its reduction as a long-term goal (e.g. *if we can reduce energy consumption)*, which is not a topic in the reference corpus. It should also be noted that the qualification *alternative* is much more prominent in the Obama corpus, which is further supported by *solar* as a specific alternative source of energy. The most frequent associated clusters in the Obama corpus are *through alternative energy (3), we've got to (3), if we can (3)* and *our energy consumption (3),* whereas the most prominent clusters found in the Reference corpus are *sources of energy (8), an energy policy (4)* and *funding alternative energy (4).* The cluster analysis partly confirms the previous findings that there is less emphasis on the alternative energy and no mention of reducing consumption in the reference corpus.

Obama corpus	**Reference Corpus**
companies (10), company (2), world's (7), reserves (2), eastern (2)	*companies (21), home heating (6), big (6), foreign (6), barrel/barrels (6)*

Table 8: Lexical collocations of *oil*

The analysis of collocations for *oil* presents a similar picture, as both Obama and the other candidates most often speak of *oil companies.* However, the collocations *reserves* and *world's* suggest that Obama is more concerned about the future supply of oil at the current rate of consumption. This is particularly evident from the recurring structure *we use/we have*

N% of world's oil, which is present in all concordance lines where the collocation *world's* is also present. Unlike the other candidates, Obama did not use the word *barrel(s),* which the other candidates mention in connection to the quantity or prices of oil. Finally, it is worth mentioning that some instances of the use of *oil* in the reference corpus can be attributed to a less prominent and more specific topic occurring only in the Reference corpus, namely *home heating oil.* The most frequent clusters associated with *oil* in the Obama corpus are *of the world's (8), percent of the (8)* and *the world's oil (7),* which also suggests that Obama is more concerned about the supplies and consumption of oil than the other candidates.

To conclude, the analysis of the keywords *energy* and *oil* shows that while the topics are present throughout the debates, the quantitative analysis revealed that they were a more important topic for Obama than the other candidates. Moreover, the qualitative analysis of collocations points to a slightly different focus on the part of Obama's campaign, namely the consumption and reserves of energy resources.

Financial, crisis

The adjective *financial* occurs in the phrase *financial crisis* in six out of thirteen instances in the Obama corpus. However, all the remaining uses are also related to the global economic recession starting in 2008: there are two occurrences of *financial package, regulations for/deregulations of the financial system, (regulatory system for ...) financial markets, financial rescue plan.*[25]

There are only seven instances of *financial* in the Reference corpus and only one of them refers to the global recession starting in 2008. The remaining occurrences suggest other topics *(false financial information, financial centres, financial interest)* or other crises *(Asian financial crisis, international financial crises that come up*[26]*).* These data demonstrate that the adjective *financial* is a keyword in the Obama corpus that can be linked to the broader topic of the recent financial crisis.

[25] Because *financial package* refers to government stimulus proposed to counter the recession, it can reasonably be considered a topic related to the financial crisis. *Regulations* and *deregulations* are also discussed with the initial causes of the credit crunch in 2006 and subsequent financial crisis.

[26] The last example is from a debate between Bush and Gore in 2000, therefore, it definitely concerns other topic than the 2008 global recession.

In general, because *financial* is an adjectival modifier associated with the topic of *(financial) crisis,* it would be helpful to verify if the most straightforward reference to this topic, *crisis,* is a keyword or not in the Obama corpus. It has really been identified as a keyword (rank 25) and even though it was preceded by *financial* only in six instances out of twenty, it can be linked to the recent financial crisis in as many as seventeen instances.[27] The remaining three uses concern the Russo-Georgian War of 2008, *the climate crisis* and *the health care crisis.* Out of 28 uses of *crisis* in the reference corpus, fifteen concern the recent financial crisis; five concern the energy *(energy crisis, home heating crisis)*; three concern international or military crises (*Africa crisis response team*, *Cuban missile crisis*); two concern other financial crises *(Asian financial crisis)*; three are related to other or general crises *(a crisis of the middle class, I believe we have a crisis here at home)*.

Crisis type	**Obama**	**Reference**
Global recession	85%	54%
International/military	5%	11%
Health care	5%	0%
Environment/climate	5%	0%
Energy	0%	18%
Other financial	0%	7%
Other/general	0%	11%

Table 9: Summary of crisis types in OC and RC

Because the debates in 2008 comprise only 24% of the reference corpus, but 54% of the occurrences of *crisis* come from this segment, it can be argued that the situation is generally perceived as more acute. Moreover, the analysis of associated clusters reveals that Obama often stressed the severity of the crisis by stating that it is *the worst financial crisis (4) since the Great Depression (5),* while no significant clusters connected to *crisis* were identified in the Reference corpus, probably because the topics were more diverse than in the Obama corpus.

To conclude, the adjective *financial* and the noun *crisis,* which were used most often to refer to the topic of global economic recession starting in 2008, were identified as keywords in the Obama corpus. However, it was shown that the topic of economic recession was prominent for both candidates in the 2008 debates and the noun *crisis* is no longer identified as a keyword when the Obama corpus is compared only to the McCain corpus, as the difference

[27] For instance, *"a middle-class tax cut for people... if they're experiencing a crisis"* was considered related to the financial crisis, as it concerns economic difficulties caused by the recession. Only unrelated crises, e.g. international, environmental, political , moral and such, were excluded.

between the relative frequencies (0.09% in the Obama corpus and 0.05% in the McCain corpus) is not significant on the chosen probability level. From the perspective of the idiolect analysis, there is a relatively firm link in Obama's speech between the noun *crisis* and the adjective *financial:* six out of thirteen uses (46%) of *financial* are followed by *crisis* and six out of twenty uses (30%) of *crisis* are preceded by *financial.* This dependency was not observed in the reference corpus, but the significance of this finding is limited by the fact that the topic only occurs in 2008.

Advisers

The word *advisers* is used in four out of five instances to refer to the claims made by the staff and political supporters of John McCain *(this is one of your own advisers; I'm using the same words that your advisers use)*. While there are no instances of the word form *advisers* in the reference corpus, there are two occurrences of the singular form with a different spelling, *advisor.* Nonetheless, the contrast in the plural form was significant enough for the noun to become a keyword. Because the noun occurred only five times and in a very specific context, it would be best regarded as one of the less important topic indicators.

Indicators of Topic: Summary of Findings

It has been shown that topic indicators can show the relative importance of chosen topics for the analyzed candidates. However, similarly to proper names, they are very dependent on the topic and unlikely to persist as idiosyncratic speech habits. This dependency can be also demonstrated on the extreme sensitivity of the analysis to any diachronic issues, as the keywords *financial crisis* seem to characterize the 2008 debates rather than Barack Obama as an individual. Even though Mollin, who studied individual stylistic profiling on the example of former British Prime Minister with BNC as the reference corpus, suggests that even among relatively topic- and context-independent items there might be some confusion between the individual traits and the development of language as such[28] (Mollin 2009: 372), the problem is much more severe in the case of aboutness keywords due to the rapid development in the most discussed political topics. The possibilities of adjusting for this issue in the present work

[28] "It is ... possible that some collocations attributed to Tony Blair ... could have become more common in recent years." (Mollin 2009: 372)

are limited, as John McCain is the only recorded speaker from the same period as Barack Obama.

However, the idiolect analysis can still benefit from the discussion of topic indicators, as they can reveal the specific ways the chosen subjects refer to the most prominent topics. For instance, it was shown that Barack Obama prefers to explicitly pre-modify *crisis* as *financial* and that he notably uses the rather technical phrase *employer-based health care system (plan)*. In the case of frequent keywords, their collocations can provide further information on the context in which the keyword is mentioned as well as the perspective of the speaker. For example, not only was it demonstrated that Barack Obama focuses on energy more than the other candidates, it was also shown that he emphasizes the aspect of sustainability through references to consumption and remaining reserves. From the methodological perspective, it is important to note that even though the keywords as such are found by statistical methods, it is impossible to proceed mechanically, especially when the researcher has to identify the broader topic, analyze possible synonyms or lexical alternatives and evaluate the context in which the keywords are used.

Lexical indicators of style

Make, sure, point

Because *make* is very frequent in all common varieties of English, it forms the basis of many phrasal and prepositional verbs and Merriam-Webster lists as many as 25 separate meanings (Merriam-Webster's 11-th Collegiate Dictionary), it would be confusing to analyze this verb as a single item. Therefore, **Table 10** lists the number of uses of the verb *make* according to the most common usage types. It should be noted that this typology is not systematic: *to make a point* was distinguished from the rest of the "abstract performative" category due to its prominence. Moreover, it mixes syntactical-semantic categories such as "causative" with particular phrases. Nonetheless, this *ad hoc* distinction based on the Obama corpus is sufficient to demonstrate how the verb was used.

It should also be noted that the category *to make a point* includes all predications, in which the noun *point* occurs as the abstract eventive object, such as *the last point I want to make* or *let just make a couple of points (OC).* Similarly, the first category *to make sure* also includes the variants with adverbial modification, such as *were I not absolutely sure (RC).*

Type of use	Examples from OC	OC	RC	OC	RC
		#	#	%	%
to make sure/certain/clear	*make sure that we're helping*	44	140	37%	38%
to make a point	*let me just make a couple of points*	22	2	18%	1%
abstract eventive object	*make an apology/decision/sacrifice/choice*	19	108	16%	30%
causative	*make our businesses ... better off*	16	86	13%	23%
to earn	*make less than $250 000*	9	18	8%	5%
to manage/ succeed/survive	*where you could make it if you tried*	6	7	5%	2%
concrete resultant object	*plants to make these highly fuel-efficient cars*	2	2	2%	1%
idiomatic/phrasal/other	*struggling to make ends meet, make up for*	1	4	1%	1%
Total		**119**	**367**	**100%**	**100%**

Table 10: Make in the Obama corpus and the Reference corpus

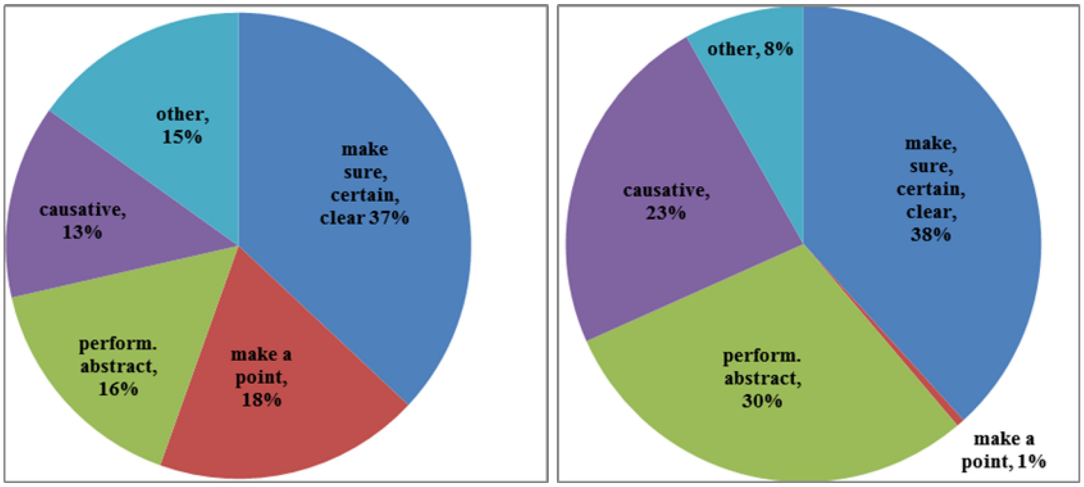

Figure 11: *Make* in the Obama corpus (left) and the Reference corpus (right) – simplified for the sake of clarity (the last four categories from Table 10 were merged into one a single category "other").

The first type includes three independent phrases, *make sure/certain/clear.* However, *make sure* makes up a vast majority of this type in both corpora. There is only one occurrence of *make certain* and one occurrence of *make clear* in the Obama corpus. In the reference corpus, there are ten instances of *make certain* and two instances of *make clear.* Even though the percentage of this type in the Obama corpus and the reference corpus are very similar, it should be pointed out that both elements, the verb *make* and the adjective *sure*, are marked as keywords in the Obama corpus due to a significantly high rate of use. 42 out 63 (67%) uses of *sure* occur in the phrase *make sure,* which demonstrates that this phrase is chiefly responsible for the keyness of *sure.*[29] Based on this evidence, it seems appropriate to treat the phrase *make sure* as the key lexical indicator of style rather than as two separate keywords.

[29] In the Reference corpus, this percentage is 62%.

Its relative frequency in the Obama corpus is 1891 occurrences per million words as opposed to the 819 expected uses based on the reference corpus. If treated as a single unit, its keyness would be 28.38,[30] which would be sufficient at the chosen probability level. Because this phrase is quite context-independent, its occurrence among keywords may suggest a permanent speech habit. While the phrase *to make sure* can be used in the agentive sense (to take necessary steps to achieve a goal) or in the cognitive sense (to verify conclusively that something is true) (Longman Dictionary of Contemporary English: sure), only the agentive sense is used in the Obama corpus: *we've got to make sure that we're helping homeowners, make sure that college is affordable for every young person in America* or *to make sure that your child has health care (OC).* However, the comparison with the reference corpus is problematic because while the agentive sense appears predominant there as well, some cases are ambiguous, e.g. *And we've just got to make sure, before somebody thinks they're buying a product, that it works (RC).*

The use of *to make a point* appears even more significant in the Obama corpus, as there are only two occurrences of this phrase in the reference corpus. Because 22 out of 35 (63%) of the occurrences of *point* occur in this phrase – analogically to the pair *sure*/*make sure* – it is arguably the cluster *to make a point* that should be considered the key indicator of style rather than the individual components. However, it needs to be emphasized that all forms of this predication were included in this category. The "canonical" form *make a point* occurs only once in the Obama corpus and is completely absent from the reference corpus. In fact, the noun *point* in singular is preceded by the indefinite article only in one instance in Obama corpus, which suggests that in both corpora, the noun *point* is only used in specific references. When one looks at the most frequent clusters associated with the keyword *point,* whereas Obama often uses the phrases *point I want to make (9)* and *last point I want to make (6),* the most common clusters in the reference corpus are *the point is (28), point is that (11)* and *but the point* (9). These phrases are almost completely avoided by Barack Obama, as there is only a single instance of *the important point is that.* From the perspective of discourse analysis, the noun *point* serves as an organizer (Biber 2006: 142): it prepares the audience for a distinct topic, argument or thought. When accompanied with a qualifier or a numeral, it also contributes to the overall structure of the speech by providing some information on how the new element fits into the line of argumentation *(the second point I*

[30] Calculated according to the algorithm explained in the previous chapters using log-likelihood as the measure of statistical significance.

want to make is, the last point I want to make). On the most general level, the higher occurrence rate of *point* points to a greater degree of explicitness concerning speech organization. Finally, the construction *the point I want to make* is not only an organizing element, but it also expresses the speaker's volition. This adds a subjective element to the discourse organization, which contrasts with the impersonal construction *the point is,* characteristic of the reference corpus.

Regarding the use of *make* with an abstract object and causative constructions, it may seem that there is a notable difference in their use between the Obama corpus and the reference corpus (see percentages in **Table 10** above). In reality, the number of occurrences of *make* + abstract eventive object in the Obama corpus per million words (854) is similar to the rate in the reference corpus (736). The same can be said for causative constructions: 719 per million words in the Obama corpus, 587 per million words in the reference corpus.

To conclude, the data suggest that the keyness of words *make, sure* and *point* can be linked to a significantly higher rate of use of two particular phrases, *to make sure* and *to make a point.* Because they are not linked to a specific topic and it was demonstrated that they are used by Barack Obama significantly more than would be expected on the basis of the reference corpus, they can be considered a part of his idiolect as lexical indicators of style.

Policies

It should be noted at the very start that only the plural form of the noun was identified as a keyword. The singular form, *policy,* occurred 21 times in the Obama corpus (944 p.m.w.) and 89 times in the reference corpus (607 p.m.w.), which is not a significant difference on the chosen probability level. While no significant associated clusters were identified in either corpus, the collocations show important difference in the noun's use.

Obama corpus	**Reference corpus**
failed (6), tax (5), economic (5), George (4), Bush/Bush's (4), eight (4), last (3), years (3)	economy (3), people (3), president's (3), Senator (2), insurance (2), grow (2), jobs (2)

Table 12: Lexical collocations of *policies*

All but two collocations in the Obama corpus *(tax, economic)* clearly refer to the previous administration, sometimes in various combinations *(failed economic policies promoted by George Bush; after eight years of failed policies),* which suggests that this word form was used by Barack Obama with a negative semantic prosody. To verify this claim, all concordance lines in the Obama corpus and reference corpus were analyzed individually in terms of positive, neutral or negative evaluation of the policies that the speakers refer to.[31]

	Obama corpus	**Reference corpus**
Positive	16%	46%
Neutral	12%	15%
Negative	72%	38%

Table 13: The connotation of *policies*

Even though the difference is not as significant as it would appear on the basis of collocations, there is still a marked tendency towards negative semantic prosody of the noun *policies* in the Obama's corpus, which should be considered a part of the idiosyncratic lexical preference.

<u>*True*</u>

12 of 25 (48%) uses of this adjective occur with a negative particle[32] in the Obama corpus, making *that's not true (6)* the only significant associated cluster. In the reference corpus, 16 occurrences out of 24 (67%) are used negatively and from the positive uses, one was used as a pre-modifying qualifier *(tried and true Republican response).* There was only one significant associated cluster, *it's not true (5).*

Obama Corpus	Reference corpus
think (4), absolutely (4), John (4), look (3), Jim (3), sorry (2)	*said (6), just (4), build (3), Hussein (2), Saddam (2)*

Table 14: Lexical collocations of *true* in OC and RC

[31] The assessment was based on the overall context of the utterances, not merely preceding adjectives. For instance, *these are the policies I have fought for* was categorized as a positive connotation, *unless we understand the rest of our tax policies* was considered neutral and *this is the kind of policies that ultimately end up undermining our ability to fight the war on terrorism* is an example of a negative context.

[32] Based on pragmatic criteria, i.e. *"that's hardly true"* or *"only a fool could think this is true*" (hypothetical examples) could be regarded as negative uses of *true*. However, no such examples were identified in the Obama corpus.

It appears that there is a mixed tendency in the Obama corpus: on the one hand, there is reinforcing by *absolutely (that's absolutely not true)*, but on the other hand, there is also hedging by *I think (I think it's true)*. *John* and *Jim* may suggest a habit of addressing the opponent and the moderator by their first names. The imperative *look* as a pragmatic device for forming a bond with the audience and drawing their attention also suggests a lower degree of formality. However, because these features are not further supported by keyword analysis *(absolutely, think, John, Jim, look* are rather rare in the Obama corpus and none of them are keywords), the only reliable conclusions seem to be that Obama was more explicit in evaluating the truth value of various discussed claims, less inclined to deny the validity of something through negative particles and that he often uses a demonstrative pronoun *(that)* to refer to the refuted statements.

<u>Deal</u>

All except one occurrence (97%) of the keyword *deal* are bound with the preposition *deal with,* the one exception being *a good deal for American people.* In the reference corpus, this percentage was only (69%). This suggests that the prepositional verb *to deal with* is largely responsible for the keyness of *deal.* In all 28 instances of *deal with* in the Obama corpus, the subject was animate: the inclusive *we (we've got to deal with Pakistan),* general human agent *(you don't deal with Russia based on staring into his eyes and seeing his soul)* or human agents expressed in a metaphorical way *(just one company trying to deal with that).* This is notable because all the instances represent only one sense of the prepositional verb, "to take ... action, especially to solve a problem" (Longman Dictionary of Contemporary English). The sense of "to concern, be about", as in "these ideas are dealt with in Chapter Four" (ibid.) or any other senses do not occur in the Obama corpus. However, the subjects in the reference corpus are also animate (using the broader notion explained above) in all instances and barring two exceptions, only the sense of "to solve a problem" is used. The two exceptions are *I know how these people think. I deal with them all the time* and *I know how to deal with our friends,* which arguably meant to express "to do business or to interact with" (ibid.). There seems to be very little difference between the Obama corpus and reference corpus in the way the verb is used, which could be explained by the specific genre of political debates that focus on solving problems. It also seems that there is a large lexical variability in the use of the prepositional verb in both corpora, as no associated clusters were identified.[33] The lexical collocations in the Obama corpus are indicative of some campaign topics: *Russia (3), energy (2), social security (2), troops (2), Afghanistan (2), Pakistan (2).* In the reference corpus, most of the collocations do not characterize specific issues: *think (3), time (3), administration (3), people (2), things (2).* However, due to the variability with which *deal with* is used, all the collocations show a rather low frequency of co-occurrence. Regarding grammatical associations, there is a tendency to use inclusive subject *we* in both corpora, but the tendency is more significant in the Obama corpus (nine uses of *we* and five uses of *we've* as opposed to five uses of *we* in the reference corpus, which is, however, seven times as large). Because *we've* was also identified as a key contracted phrase, this topic will be addressed in more detail among functional keywords.

Provide

From the syntactic point of view, the verb *provide* is used in a similar way in the Obama corpus and the reference corpus. **Table 13** shows the use of *provide* in all valency structures identified in the corpora.

Valency structure	**Examples (OC)**	**OC**	**RC**	**OC %**	**RC %**
provide something to someone	*provide education to kids*	12	9	44%	27%
provide something for someone	*provide health care for those*	4	9	15%	27%
provide somebody something	*provide you the option*	3	3	11%	9%
provide somebody with something	*provide them with resources*	1	2	4%	6%
monotransitive/implicit recipient	*provide moral support*	7	10	26%	30%
Total		27	33	100%	100%

Table 15: Syntactic structures of *provide* in the Obama corpus and the reference corpus

The largest difference between the Obama corpus and the reference corpus with regard to valency structures seems to be in the ditransitive uses with a prepositional object, "to provide something to someone". At first, it may seem that Barack Obama prefers only this structure, but all the occurring categories are clearly overrepresented in the Obama corpus when the respective sizes of the corpora are taken into account, which suggests an overall lexical preference for this verb.

[33] The default frequency limit is five occurrences of the cluster. However, even when this limit was decreased to three occurrences, no clusters were identified in either corpus.

Obama corpus	**Reference corpus**
health care (7), tax (5) [cut(3)/credit(2)/relief(1)][34]*,* *want (4), moral support (3)*	*health care (4), leadership (3), support (2)*

Table 16: Lexical collocations of *provide* in the Obama corpus and the reference corpus

In both corpora the verb *provide* is most often linked to health care.[35] However, Obama also speaks about providing tax cuts[36] and moral support, which are topics that do occur in the reference corpus (53 uses of tax cuts, two occurrences of moral support), but they are not linked to the verb *provide.* While the overall preference for the verb does not suggest any specific topic, its higher rate of use, especially in connection to the collocations, may be relevant to the perception of Obama's policies on the liberal-conservative axis.

Last

Due to a high rate of use the keyword *last* in recurring contexts in both corpora, it is possible to perform a detailed analysis of the associated clusters. **Table 15** shows the occurrence of *last* in clusters and how several related smaller clusters (three words) aggregate into larger clusters (five to six words). The larger clusters are particularly notable because they are known to be quite rare (Biber et al. 1999: 993) and in consequence, their frequent use is very significant from the perspective of the idiolect analysis. The clusters found in both corpora are marked green to facilitate the comparison.

[34] These collocations were grouped as very close synonyms. Even though none of them exceeds the limit of five uses, there are six instances when they are considered together.

[35] Obama prefers the construction *provide something to somebody* when he talks about health care and moral support, but *provide something for somebody* when he talks about tax cuts. However, there are only very few examples of the form *provide something for somebody* to form definite conclusions and there is one exception to the rule – see examples in the table. This tendency was not observed in the reference corpus.

[36] Tax cut, tax credit and tax relief can be considered synonymous.

Obama corpus		**Reference corpus**	
Small clusters	**Large clusters**	**Small clusters**	**Large clusters**
over the last (24) *last eight years (19)* *the last eight (19)*	*over the last eight years (14)*	*in the last (35)* *last eight years (23)* *the last eight (23)* *for the last (18),* *over the last (12)* *last four years (9)* *the last few (7)*	*in the last eight years (8)* *over the last eight years (5)* *in the last four years (3)* *over the last four years (3)*
last point I (7) *I want to (6)* *point I want (6)* *one last point (5)*	*last point I want to make (6)*		
		as a last resort (9)	

Table 17: The cluster analysis of *last.*

In the Obama corpus, there are two main cluster groups. The center of the first is the notion of a previous time period, specifically previous eight years. Even though the period of eight years was clearly determined by the American electoral cycles,[37] Obama shows a very strong preference for the prepositional construction *over the last,* which is represented far less in the reference corpus (1080 per million words in OC, 82 per million words in RC). Even the five-word cluster *over the last eight years* has a normalized frequency of 630 occurrences per million words, which is an extreme number considering its length.[38] It should be noted, however, that this reference is firmly linked to the situation of the 2008 election, in which Obama attempted to distance himself from two Republican administrations,[39] and thus would be best regarded as an indicator of topic. This would also explain the absence of the cluster *(in the) last four years.* However, Obama also completely avoided the preposition *in* in the same type of structure *(in the last * years)* and there is only one instance of *for the last (eight years)*, which are fairly frequent clusters associated with *last* in the reference corpus. Because

[37] When Barack Obama speaks about the last eight years, he refers to two administrations of George W. Bush, who was elected in 2000 and re-elected in 2004.

[38] Four-word clusters are usually defined by a frequency limit of 10 - 40 occurrences per million words (Biber and Barbieri 2007: 267) To further illustrate this rate, the entire phrase *over the last eight years* is approximately as frequent in the Obama corpus as the common possessive pronoun *my.*

[39] It is generally known that an appeal to change was an essential part for Obama's campaign. See, for instance, David Meerman Scott: Ten Marketing Lessons from the Barack Obama Presidential Campaign (http://www.webinknow.com/2008/11/ten-marketing-lessons-from-the-barack-obama-presidential-campaign.html)

the phrases *over the last N years* and *in the last N years* can be considered synonymous, it appears that Obama has a strong preference for the former. It should also be noted that the comparison of the relative frequencies of the structures *over the last* in the Obama corpus and *in the last* in the reference corpus shows that it is not merely a lexico-grammatical preference, but indicates a much stronger focus on the previous administrations on the part of Barack Obama (1079 instances p.mw. in the Obama corpus, 239 instances p.m.w. in the reference corpus).

The other group of clusters associated with *last* is centered on the noun *point,* which has already been discussed as a keyword in the Obama corpus. The smaller clusters often aggregate into a six-word bundle *last point I want to make,* which is also significant because only a single instance of *last point* can be found in the reference corpus. Unlike the bundles referring to the previous administrations, the clusters including *point* can be regarded as lexical indicators of style and discourse organizers. Their use also shows a greater explicitness in the structure of speech, which has already been discussed under the keyword *point.*

Finally, the phrase *the last resort* is completely missing in the Obama corpus, but as it was only used by John Kerry, it is not relevant for the idiolect analysis of Barack Obama.

<u>Potentially</u>

Potentially might suggest a tendency towards hedging, but it should be kept in mind that there are only six instances of this disjunct in the entire Obama corpus. It seems notable, however, that its use is in 50 per cent of instances non-standard: *they sent nuclear secrets, potentially, to countries like Syria* and *he would not meet potentially with the prime minister* use a non-standard word order and in the latter case, the adverb appears redundant if not meaningless, which is also true for the third construction: *$700 billion, potentially, is a lot of money.* The remaining occurrences correspond to the standard use as a content disjunct *(countries ... potentially have an interest, that could potentially happen).* It should be noted, however, that syntactic and lexical mistakes as well as redundancies are generally common in spoken language. The comparison of Obama's use of the adverb with the reference corpus was impossible as there are no instances of this adverb there. Its occurrence among keywords may suggest a slight idiosyncratic lexical preference and when the mistakes are taken into account, it is possible that Obama uses the disjunct as an idiosyncratic filler, albeit a very infrequent one (270 occurrences per million words).

Important

Owing to the relatively frequent use of the adjective in both corpora, it is possible to perform the analysis of associated clusters and collocations. Because there is a large group of smaller clusters in the Obama corpus that does not quite connect into a larger bundle but always conforms to the same syntactic structure, the underlying pattern was described instead of a larger bundle. This structure, *I think it's/it is important (for NOUN/PRONOUN) to VERB*[40] was identified as the basis for many frequent structures in the reference corpus as well. However, the structure *important for us*, which functions as an adjunct of respect or the subject of an infinitival clause is a prominent feature is the Obama corpus (nine occurrences), and there are only three such instances in the reference corpus. On the other hand, the pattern *it/is a very important thing* is completely missing in the Obama corpus, even though it was used by several speakers in the reference corpus.

Obama corpus		**Reference corpus**	
Clusters	**Underlying pattern**	**Clusters**[41]	**Underlying pattern**
for us to (11) *I think it's (10)* *important for us(9)* *think it's important (7)* *it is important (7)* *it's important for (6)*	*I think it's/it is important (for [NOUN/PRONOUN PHRASE]) to [VERB] (6)*	*I think it's (19),* *it's important to (19),* *it's very important (14),* *think it's important (13),* *think it's very (10),* *it's important (9),*	*I think it's/it is important (for [NOUN/PRONOUN PHRASE]) to [VERB] (8)*
		a very important (14) *is a very (9)* *this is a (8)*	*it/this is a very important [NOUN] (10)*

Table 18: The clusters and underlying patterns associated with *important*

[40] Parentheses mark a non-obligatory part of the pattern.

[41] Only the first ten most frequent clusters were chosen as it would be impractical to list all 21. The remaining clusters also fit into the identified underlying structures.

Obama corpus	**Reference corpus**
understand (7), recognize (3), point (2), issue (2)	*president (9), America (6), issue (6), question (5), make (5), understand (4)*

Table 19: Lexical collocations of *important* (does not include items found in the clusters – see Table 16)

The analysis of collocations shows that Obama often used the structure *it's important (for someone) to understand (that),* such as in *It is important for us to understand that the way we are perceived in the world is going to make a difference.* However, in one instance he also used the construction *It's important that we understand (they're not the old Soviet Union).* While the collocation *understand* can also be found in the reference corpus, it is far less prominent. Moreover, while *recognize* is not a full synonym of *understand,* it falls within the same semantic category, the verbs of cognition, and performs the same pragmatic function: *important is that we recognize that to solve the key problems* or *It's important that we recognize there are going to be some areas of common interest.* The third collocate, *point*, which has already been discussed, occurs in the noun phrase *the important point*. The set of collocations in the reference corpus, the most prominent of which are nouns, unlike in the Obama corpus, shows a focus on the office of the president, either through a *for*-prepositonal phrase (four instances, e.g. *and I think it's important for the president to set a tone),* the president as a cognitive subject (three instances, such as *president felt that it was important*) or other connection (two instances: *what's really important, Charlie, is the president is just trying...* and *ladies and gentlemen, important to understand[sic], the president and his friends try to make a big deal out of it*. Very similar constructions can be identified for the proper noun America.

From the pragmatic perspective, the use of *important,* particularly in combination with *point* or *issue* suggest that Obama prefers a greater degree of explicitness with regard to discourse organization, as he directly emphasizes the prominence of the some elements.

<u>Additional</u>

Out of sixteen occurrences of *additional,* ten are connected to *tax cuts/breaks* through constructions such as *add an additional tax cut over the loopholes* or *additional four billion in tax breaks.* In seven cases, the adjective is followed by a numeral. The remaining instances refer to the military *(additional troops, additional brigades),* general finance *(additional funding)* and oil *(get some additional oil).* In the reference corpus, only three out of twenty occurrences can be linked to tax cuts, six to the military and three to general finance. The rest can be found in miscellaneous constructions such as *additional choices, additional bill* or *additional protections.* In only one instance, the adjective was followed by a numeral. Overall, it seems that Obama's lexical preference for the adjective *additional* is accompanied by a tendency to provide specific numbers (44%).

Just

The relatively frequent use of the adverb *just* might be a good candidate for a permanent speech habit because it is completely independent of topic and usable in many different – though admittedly mostly neutral to informal – registers. In Barack Obama's speech, the adverb often fulfills a pragmatic discourse function as an emphasizer (CGEL: 447) *(I just fundamentally disagree; look, that's just not true)*, but also as a diminisher (CGEL: 598) *(let me just make a closing point, let me just correct the record here)*. Because its pragmatic function is to minimize the intrusion (the time that the speaker asks for), it can be considered a means of enhancing politeness. It is also used in the sense of *only*, mostly in the negative form *not just (not just more troops, not just when there's a crisis).* Due to its relatively high frequency in both corpora, it is possible to look at clusters associated with this adverb.[42]

Obama corpus	**Reference corpus**
let me just (7)	*let me just (28)*
just want to (5)	*just want to (5)*
	the president just (15)
	me just say (12)
	not just a (9)
	president just said (8)
	it's not just (8)
	we just have (7)
	he just said (7)
	the vice president [just] (7)
	just a few (7)

Table 20: Clusters associated with *just*

While Barack Obama as well as the other speakers uses the construction *let me just (say, correct, repeat, make a point),* two phrases prominent in the reference corpus, *the president just (said, talked about, didn't level with you)* and *the vice president,* are completely missing

[42] There was no instance of *just* used as an adjective in the Obama corpus.

in the Obama corpus. However, because these constructions were only used by Senator Kerry, their absence in the Obama corpus does not appear very significant.

Obama corpus	**Reference corpus**
make (9), let (8), want (7), quick (7), point (6), true (5), correct (5), important (4), people (3), know (2)	*president (32), let (32), said (27), say (25), think (15), know (13), people (10)*

Table 21: Lexical collocations of *just*

The lexical collocations of *just* show that in the in the Obama corpus, the adverb is often connected to other keywords such as *make, (quick) point, true* and *important.* It is also clear that some of the collocations can be attributed to the bundles *let me just* and *just want to.* It is also notable that the phrase *make a (quick) point* is used as an alternative verbum dicendi in the Obama corpus instead of the more direct *say/said* often occurring with *just* in the reference corpus. From the pragmatic perspective, the cluster *let me just make a (quick) point* and similar constructions can be considered an element enhancing politeness because it explicitly asks for a time to respond while it attempts to downplay the intrusion *(just, quick).*

The noun *president* remains the most significant difference because it does not occur as a collocation of *just* in the Obama corpus. Unlike the cluster *president just said,* the noun *president* was used by at least four speakers in the (-5,+5) vicinity of *just.* From the perspective of the idiolect analysis, it appears that Barack Obama used the adverb with a relatively high degree of variability, especially in comparison with keywords such as *employer,* which occurred in a specific phrase in 50% of the instances and in all cases could be linked to a specific topic. Therefore, it seems that the keyness of *just* cannot be attributed to a specific phrase, bundle or context and it can serve as a good example of an idiosyncratic preference that is likely to occur in other speech situations. Nonetheless, the bundles associated with the adverb are statistically significant and can also be considered a part of Barack Obama's idiolect, even though they do appear also in the reference corpus (*let me just* 314 p.m.w. in OC, 191 p.w.m. in RC; *just want to* 225 p.m.w. in OC, 34 p.m.w. in RC).

Muddle

Muddle is among the lexical keywords despite only five uses because there was no occurrence of this word in the reference corpus. In all five instances, it was a part of the phrasal verb *(to) muddle through (we can muddle through Afghanistan, you don't muddle through stamping out the Taliban).* It is marked as a Briticism in the Longman Dictionary of Contemporary English, which can explain its absence in the reference corpus. While five samples are clearly insufficient for a definitive conclusion, this keyword might point towards an idiosyncratic preference for a chiefly British phrasal verb.

Lexical Indicators of Style: Summary of Findings

Unlike topic indicators, lexical indicators of style often provide insights that could not be predicted on the basis of a different type of analysis. For instance, some of the identified key topics could have been determined by comparing what Obama's campaign focused on with the campaigns of other candidates, but a relatively frequent use of particular words, often with a low perceptual salience (such as *just*), is arguably difficult to notice without the help of a keyword analysis. Some lexical preferences are linked to larger units – clusters or lexical bundles, such as *over the last (eight years).* Because the lexical indicators of style are in some ways similar to functional style indicators (lower perceptual salience, not indicative of particular topics, often linked to other elements), it is expected that the analysis of functional style indicators might yield similar results.

5.1.3. Functional style indicators

At first glance, it might appear that grammar keywords (functional style indicators) are less informative than their lexical counterparts because they are very common and by definition do not represent autonomous units of meaning. However, their relatively high rate of occurrence is, in fact, an advantage for the corpus-driven approach because larger samples are less prone to statistical anomalies. Moreover, Hunston (2008: 273) demonstrates that function words can be "crucial to textual meaning" due to their role in grammatical patterns associated with particular semantic sequences. In accordance with this observation, the analysis of function keywords in the Obama corpus will focus on clusters and underlying patterns rather than individual concordance lines, which is also more practical considering their number.

That

Because *that* is one of the most common words in English and it can perform several grammatical functions, the analysis will proceed in two steps using the tagged version of the corpora. The first step will be the breakdown of its use according to the word class of the preceding word. The second step will concern the word classes of the keyword itself. The rationale for this approach is that the preceding unit contributes significant information on the combinatory properties of *that,* but at the same time, this information does not unambiguously determine the word class of *that* in the particular context. For instance, "VERB that" is typical of declarative clauses such as *Senator McCain said that* but also copular constructions such as *doesn't have to be that way*.

Structure	**OC**	**OC – p.m.w.**	**RC**	**RC – p.m.w.**
NOUN that	244	11 084	1105	7 607
VERB that	194	8 813	1060	7 297
ADJECTIVE that	77	3 498	140	964
CO. CONJUNCTION that	70	3 180	230	1 583
SENTENCE END that	67	3 044	344	2 368
ADVERB that	40	1 817	149	1 026
COMMA that	37	1 681	196	1 349
SUB. CONJUNCTION that	32	1 454	318	2 189
Others	30	1 363	227	1 563
TOTAL	**791**	**35 933**	**3 769**	**25 946**

Table 22: The breakdown of *that* according to the word class of the preceding word

Surprisingly, *that* is used with higher frequency per million words in the Obama corpus in all types of constructions except for "[subordinating conjunction] + that" and the category of minor[43] miscellaneous uses, such as "[predeterminer] + that" or "[colon] that". The difference in this category is at any rate insignificant. Therefore, it appears that the aggregate effect is responsible for the keyness of *that* rather than a single type of construction. That being said, there are several specific structures that appear typical of Barack Obama.

The category "adjective + that" includes the following notable structures: *make sure that* (49% of the uses of *[adjective] that*) and evaluative comments such as *it is (absolutely) true*

[43] Fewer than five occurrences in the Obama corpus per word class.

that (8%) and *it is (very) important that (8%)*. The fact that 65 per cent of the combination "adjective + that" includes other keywords or key phrases *(McCain, true, important, make sure)* shows that the analysis of function words can not only lead to identifying lexical keywords but also help in finding links between various key units.

The combination "coordinative conjunction + that" comprises *and that (94%)* and *but that (6%)* in the Obama corpus. While the relative representation is the same in the reference corpus, the significant difference in the frequency per million words suggests that Barack Obama has a significant preference for the combination *and that*, e.g. *And that means, yes, increasing domestic production.*

The combination "[sentence end] + that" demonstrates an increased tendency to use *that* as a relative pronoun in the initial position in the Obama corpus. While the initial *that* could be also used as a subordinating conjunction, such as in *That he lied was naughty (hypothetical example),* no such instances have been identified in the Obama corpus[44] and only eleven cases were found in the reference corpus. In those instances, *that* was only used in the initial position in a specific structure used for emphasis, "[main clause] *that* [subordinate clause]. *That* [a second subordinate clause on the same level as the first]", which could be exemplified by *to have an asset that you can call your own. That you can pass from one generation to the next (RC).* It is, therefore, possible to conclude that there is a slight preference for relative *that* in the initial position in the Obama corpus, but this tendency is not very significant when it is analyzed separately (log-likelihood 14 according to manual calculation).

The combination "adverb + that" consists of the structure *so that* (52,5%) and various minor combinations, such as *then that (2)* or *just that (2)*. However, *so that (25)* is not statistically significant in the Obama corpus, as it is also quite frequent (60) in the reference corpus and the calculated log-likelihood is 14. As for the other combinations, the difference in use measured in the number of occurrences per million words is either not significant, such as in the case of "[verb] + *that*", or the precedent varies to such degree that the discrepancy is hard to describe beyond the numbers listed in the table (see "[noun] + that").

[44] Even though some uses have been categorized as subordinate conjunctions in the Obama corpus by the tagger, a closer review showed that those were in fact tagging errors caused by a non-standard word order (*THAT I don't think is an example of "speaking softly" - OC)* and software deficiencies in general.

The second step of the analysis concerns the functional word classes of the keyword itself. The table below shows a breakdown of *that* according to its word class in the Obama corpus and the reference corpus.

Word class	**OC**	**OC %**	**OC – p.m.w.**	**RC**	**RC %**	**RC – p.m.w.**
Subordinating conjunction	432	55%	19 416	1664	44%	11 349
Relative pronoun	178	23%	8 000	904	24%	6 165
Determiner or demonstrative pron.	174	22%	7 820	1177	31%	8 027
Adverb	3	0%	135	24	1%	164
Total	787	100%	35371	3769	100%	25705

Table 23: The functional word classes of *that*

It is evident that only *that* in the function of a subordinating conjunction constitutes a statistical difference between the Obama corpus and the reference corpus. Because *that* is a non-obligatory element in nominal *that*-clauses which function as a direct object or a complement as well as when the subject of *that* clause is extraposed (CGEL 1049), it could be speculated that Obama prefers the more formal variant (ibid.) without omission in these sentence structures. The alternative explanation would be that Obama used more *that*-clauses than the other candidates. Because it would be impractical to verify this hypothesis on thousands of instances, three most frequent verbs introducing *that*-clauses were chosen from the Obama corpus and compared to the reference corpus. The term explicitness refers to the percentage of instances where the verb is immediately followed by *that.*

Nominal declarative structure	**Explicitness - OC**	**Explicitness - RC**
make sure (that)	90%	38%
think (that)	22%	14%
means (that)	41%	19%

Table 24: Explicitness of *that* in the Obama corpus and the reference corpus. Explicitness refers to the percentage of instances in which *that is* not omitted.

These verbs are much more frequently used without the omission of *that* in the Obama corpus and in two out of three cases, the difference is quite significant. It seems reasonable to conclude that the lower degree of omission of *that* after verbs introducing nominal content clauses in the Obama corpus is the significant feature largely responsible for the keyness of *that*. This analysis also explains why almost all combinations "[word class] + that" listed in **Table 22** are more frequent in the Obama corpus, as most of them, particularly the three most common ones (a noun, a verb, an adjective + that) can introduce declarative content clauses.

Going

164 out of 181 instances (91%) occur in the construction *to be going to*,[45] which is a semi-auxiliary construction used to express "future fulfillment of the present" (CGEL: 214), either is the sense of probable outcome or a present intention. This percentage is similar in the reference corpus (84%). Consequently, it is reasonable to treat the auxiliary construction rather than *going* itself as the key unit and look at the associated clusters.

Obama corpus	**Reference corpus**
going to have to (35)	I'm not going to (28)
we're going to have (22)	going to have to (28)
we are going to (15)	and we're going to (24)
not going to be (11)	we're going to have (18)
be able to (10)	we're not going to (17)
we're not going to (9)	going to be a (15)
going to be able (9)	and I'm going to (15)
we're going to do (6)	it's going to be (11)
	not going to be (10)
	we're going to do (9)

Table 25: Associated clusters of *going to* in the Obama corpus and the reference corpus

While there is some overlap, particularly in the expressions of obligation *(going to have to)*, it appears that there is a stronger tendency to use first person singular in the reference corpus. To verify this hypothesis, a list of nominal collocates was drawn in both corpora for a further analysis of the subjects in the future auxiliary construction. Of course, left collocations in the distance of four words of less of the phrase do not necessarily have to be the subject of the clause; however, it appears a reasonable approximation, particularly when some of the collocations include the contraction of the auxiliary *be*. To facilitate orientation in the list, the collocations were sorted according to an ad-hoc measure of typicality, which was calculated as the ratio between the frequency p.m.w. in the Obama corpus and the frequency p.m.w. in the reference corpus in per cent adjusted for the different rate of occurrence of *going to* itself. Thus adjustment means that the higher relative frequency of *going to* in the Obama corpus is

[45] The cases such as *I'm going to Paris (hypothetical example)* are also included here, even though they do not constitute the same syntactic structure. This simplification was needed due to the number of occurrences, particularly in the reference corpus.

not reflected in the calculation. For instance, the structure *we're (*) (*) (*) going to*[46] has a degree of typicality of 281%, which means that *we* occurred with *going to* 2.81 times more often in the Obama corpus than in the reference corpus when the different sizes of the corpora are taken into account.[47] However, the combination of *we* + *going to* in reality appears 9.33 more often in the Obama corpus.

Subject/Contracted form	OC #	OC – p.m.w.	RC	RC – p.m.w.	Degree of typicality
we	27	1213	19	130	281%
it	15	674	19	130	156%
this	5	225	8	55	124%
it's	14	629	23	157	120%
that	32	1438	60	409	105%
we're	60	2697	133	907	89%
I	9	404	22	150	81%
that's	7	315	21	143	66%
you	7	315	21	143	66%
they're	5	225	15	102	66%
he	4	180	19	130	42%
you're	5	225	32	218	31%
he's	3	135	21	143	28%
people	1	45	18	123	11%

Typical of Obama – Untypical of Obama

Table 26: Subjects of *going to*.[48]

It is clear that Obama prefers to use the inclusive *we* when he talks about his plans or the future as such. It should also be noted that the use of the non-contracted form is much more significant in comparison with the other candidates. In general, Obama prefers impersonal pronouns such *this, that* and *it* to second and third person personal pronouns *you, he* and *they* found in the reference corpus. The use of *people* in the future constructions is also quite rare in the Obama corpus. Quite significantly, there is a strong avoidance of the contracted form *I'm* and a slight avoidance of *I,* which again suggests a preference for a more inclusive language. It should also be noted that there are only two contracted forms among the preferred subjects+verb combinations in the Obama corpus, but four contracted forms are typical of reference corpus, which indicates increased formality on the part of Barack Obama.

[46] *(*)* denotes a non-obligatory element (word).

[47] A more precise measure of significance could have been also used, but the present goal is only to facilitate orientation in the presumed subjects of the construction.

[48] The subject *thing* has been omitted from the table as unrepresentative, as there was only one instance in each corpus.

We've got (to)

57 per cent of all instances of *got* occur in the cluster *we've got to,* as opposed to only 19% in the reference corpus. Moreover, *we've* is a left collocate[49] of got in 74% of the instances and it seems reasonable to assume that it is the subject of the clause in a majority of those cases. This ratio is only 37% in the reference corpus. Considering these data, it appears appropriate to consider the cluster *we've got to* as the key unit rather than the individual words. This key phrase would have the log-likelihood value of 90, thus showing a very high degree of statistical significance. The structure expresses a shared obligation of a group in which the speaker is included, which supports the claim that Obama has a tendency to employ inclusive language. However, the contracted form goes against the observed preference for a more formal language, particularly when the informality of the semi-modal *have got to [base form of the verb]* is taken into account. Because there are several ways of expressing a joint obligation, ranging from *we must/have to/need* to the more formal *it is necessary (for us),* none of which have been identified as keywords,[50] this structure can be considered a significant lexico-grammatical idiosyncrasy on Obama's part.

We, are

In the case of very frequent keywords such as *we,* the cluster analysis is the most suitable tool to discover significant patterns. Because *we* very often occurs as a subject of a clause, which narrows the scope of recurring structures, and many clusters have been identified, it is possible to extend this analysis by looking at the semantic sequences. This technique, in detail described by Hunston (2008), is based on a semantic classification of structures preceding or following the discussed keywords. The obvious advantage of this approach over a simple list of clusters is that when the speaker uses synonymous or variant expressions, semantic analysis can reveal patterns that do not appear significant at first glance. However, it should be noted beforehand that some of the categories overlap, such as common obligation and future or obligation and hedging. Due to a high number of the identified three-word clusters and a large degree of overlap among them, the table only lists four-word bundles.

[49] As has been stated in the chapters dedicated to methodology, the range is set to four words unless stated otherwise.

[50] With a possible exception of nearly synonymous *(it is) important (for us).*

Semantic class	**Clusters – OC**	**Clusters – RC**
Common obligation	that we have to (13) we have to do (7) and we have to (5) going to have to (7) think we have to (5)	we need to do (25) think we ought to (23) we need to have (21) we have got to (18) we have to do (17) that we ought to (15) we have to be (15) that we ought to (15) that we have to (14) we need to (14) I think we need (14)
Common future	we are going to (15) we are going to have (7)	are we going to (8) we will come home (5)
Certainty with regard to achieving goals[51]	make sure that we (10) to make sure that (we) (7) sure that we are (6) making sure that we (5)	to make sure that (20) make sure that we (16) everything we can to (14) to make sure we (13) do everything we can (13)
Preceded by a cognitive verb or hedging	I think that we (7) I think we have (6)	I think that we (18) I think we have (16) I don't think we (10) I believe that we (8)

Table 27: Semantic sequences associated with *we*.

Even though there are differences in wording (Obama, for instance, avoids *ought to*), the most frequent bundles in the Obama corpus and the reference corpus fall into the same semantic categories. This suggests that the more frequent reference to the inclusive first person plural in various bundles and semantic contexts in the Obama corpus is responsible for the keyness of *we* rather than a particular structure. The keyness of *are* is closely connected to the keyness of *we* as they co-occur in 32 per cent of instances in the Obama corpus and it is, of course, the only possible form of the verb *to be* in the first person plural in the present

[51] Even though the phrase to *make sure* can be interpreted in two distinct ways, i.e. (1) to make necessary steps to achieve a goal and (2) to conclusively verify the truth of a proposition, the previous analysis of these keywords has shown that Obama only uses the phrase in the former sense.

tenses in the indicative mood. It also seems that Obama is less likely to contract *we are* into *we're:* only 60% of instances are contracted in comparison to 83% in the reference corpus, which contributed to the keyness of *we* and *are* as separate keywords. From the perspective of discourse analysis, *we* in Obama's speech invariably[52] refers to the collective identity of the nation, but the reference is sometimes metaphorical *(we have lost 4000 lives, we have to fix our health care system, we are spending $300 billion on tax cuts).* The use of inclusive language might have served as a way to form a closer bond with the audience, as *we* implicitly suggests common interests and attitudes *(I'm convinced we can do it, we cannot tolerate a nuclear Iran).* However, there are also two specific functions. First, by avoiding the distinction between the first and second person, expressions of obligation and necessity are more polite because they no longer seem like a directive *(we have to change our policies, we have to have president who...).* Second, by using the first person plural, Obama avoids direct accusations of his predecessors or other agents: *we took our eye off the ball, we hadn't caught Bin Laden, we did not use our military wisely in Iraq.* To conclude, it appears that Barack Obama tends to use inclusive language in both positive and negative contexts.

<u>Some</u>

Some can be used as a pronoun, determiner or adverb, which creates the same difficulties as *that.* The analysis of clusters shows that the clusters found in the Obama corpus *(some of the (17), there are some (7), to make some (5), some of these (5))* are also among the most frequent clusters of the reference corpus. However, the structure *some of the [noun]* occurs 764 times per million words in the Obama corpus, but only 191 in the reference corpus. The ratio is also quite significant for the existential construction *there are some* (360 instances p.m.w. in the Obama corpus, 41 instances p.m.w. in the reference corpus), but the number of these phrases in both corpora is too low for us to make any conclusions (6 in OC, 8 in RC). It is also notable that whereas *things (7)* is the most frequent lexical collocation of *some* in the Obama corpus, it is far less prominent in the reference corpus (9). In constructions such as *some areas of common interest, safer in some ways, in some cases* or *stop some of the abuses (OC),* this keyword is used as a tool for hedging, i.e. expressing uncertainty (CGEL: 1089), or limiting the extent of the speaker's claims. *Some* is also used as an adverb in the reference

[52] For practical reasons, the test was performed on 50 random samples.

corpus (*some $900 billion, some 9 000 parents, some 29 separate tax credits*), but there are no such instances in the Obama corpus.

Is

This form was expected to appear very frequently and with a great variability in both corpora. It is, therefore, surprising that it was identified as a keyword in the Obama corpus. As the first step, it might be useful to consider the structures in which *is* can occur. The cluster and collocation analysis as well as a review of concordance lines were used to identify recurring structures that might be responsible for the increased frequency of *is* in the Obama corpus. Subsequently, those structures were analyzed separately to identify statistical discrepancies between the two corpora. Three main types of significant recurring sequences were identified: pseudo-clefts *(what I've called for is tax cuts),* combination of *is + [ADJECTIVE]* or more specifically *it is + [ADJECTIVE] (it is constitutional), this is* and *it is* in general*,* which can be divided into anticipatory clauses *(it is important for us to understand)* and other uses *(it is breaking family budgets)*. It should be noted that this list is not intended as a classification because the categories often overlap (e.g., *it is important for us to understand* is an example of *is + [ADJECTIVE], it is + [ADJECTIVE]* and an anticipatory clause). The collocation analysis was used to estimate the number of pseudo-cleft clauses, as it seems reasonable to assume that *what* occurring six words or less before *is* marks a pseudo-cleft clause. Admittedly, this is method is quite rough, as some pseudo-cleft clauses longer than six words will be omitted and some other structures, such as questions and reported questions, will be mistakenly included. However, it has been observed that questions are generally rare in this type of debates due to a lower degree of interactivity of the debates, particularly in comparison to informal conversation. Because the other structures are continuous and have only one form, they were searched for directly in the tagged versions of the corpora. The categorization of *it is* into anticipatory and other uses was performed manually, which was possible due to relatively lower frequencies of occurrence.

	Obama corpus – p.m.w.	**Reference corpus – p.m.w.**
Pseudo-cleft	2 157	887
is [ADJECTIVE]	1 888	996
it is [ADJECTIVE]	539	116
it is	1 303	798
it is (anticipatory)	890	307
it is (other)	404	491
this is	1 978	832

Table 28: *Is* in recurring structures in both corpora. The left brace marks categories that may overlap. The number of the structures *it is [ADJECTIVE]* is included in the total number of *is [ADJECTIVE]*. The category *it is* can be divided into *it is (anticipatory)* and *it is (other)*, which are distinct and complementary.

The results indicate that Obama tends to use pseudo-clefts more often that the other candidates. The structure *is [ADJECTIVE]* was also more common in the Obama corpus and it could be linked to the lexical keyword *important,* which was identified as a frequent collocation of *is (33)*. *It is* could be linked to the key structure *(it is) going to (35),* however, it seems that the anticipatory constructions are far more prominent. The last significant structure, *this is,* is variously used with noun phrases *(this is an example),* adjectives *(this is undeniable)* and verb phrases *(this is going to be an important issue)* with different modifications or negation. While no particular phrase stands out, it seems that Barack Obama generally prefers to use the phrase *this is* to refer to the co-text (previous parts of discourse), which contributes to the overall discourse organization, similarly to elements such as *(last) point.*

<u>Nobody</u>

Barack Obama used the negative pronoun *nobody* in 6 out of 13 instances to indirectly distance himself from certain claims or institutions: *nobody is talking about losing the war, nobody talked about attacking Pakistan* or *John, nobody is denying that $18 billion is important.* In accordance with this function, verbum dicendi was used in all such instances. The secondary function of *nobody* is to form a bond with the audience by pointing out the common attitudes and one instance of such a use was identified: *Look, nobody likes taxes. I would prefer that none of us had to pay taxes, including myself.* The use of nobody is significant in connection with the negative keyword *I,* which suggests that Barack Obama prefers indirect ways of referring to himself or his views.

Here's

The contracted form *here's* is used to introduce a new aspect of a situation (*Here's the problem*) or to draw attention to an important or concluding point (*But here's what I do know; So here's what my plan does; Here's what I would say*). *Here's* also often (42%) followed by a nominal relative clause, such as *here's what I do know* or *here's what I would do.* Even though there are only ten instances of *here's what* in each of the corpora, the discrepancy in frequencies per million words (449 and 68, respectively) is quite significant. As there are 215 instances of *here* in the Obama corpus and yet it was not identified as a keyword, it appears that only the contracted form, often followed by a nominal relative clause, is idiosyncratic on Obama's part. This feature of Obama's idiolect again contributes to the explicit discourse organization.

So

So was also used by Obama to mark a shift in perspective to a new or concluding point (*so let's talk about this; so my attitude is*), but it is also used as a conjunction of cause, reason and purpose. It should be noted that a relatively higher rate of *so* fits the linguistic profile in which sources, inferences and the opinion on validity are often expressed explicitly, thus contributing to discourse organization, which was discussed under the several other keywords, such as *make, point* or *last* or *important.*

As for the collocation and clusters, *so* manifests a great degree of variability in both corpora, which is why these types of analysis did not prove particularly illuminating. However, it was observed that Obama often uses *so* in the initial position in a sentence, such as in *So if we are going to...* or *So let's get the record straight.* The comparison of the corpora shows that Obama uses the initial *so* at the rate of 2337 occurrences per million words (44% of all uses of *so*) in contrast to only 928 in the reference corpus (29%). It is also important to note that *so* ceases to be a keyword if the use in the initial position is not included in the calculation, as the calculated log-likelihood of non-initial *so* is only 2.45, which is far below the selected probability level. It is, therefore, possible to conclude that only the initial *so* is significant in the Obama corpus.

Moreover, it appears that there is a connection between the keywords *so* and *here's,* as there are five instances of a larger cluster *So here's what,* which in all cases introduces a plan or policy *(So here's what my plan does, So here's what we have to do).* There is only a single instance of *So here's what* in the reference corpus.

Negative keyword: I

Two negative keywords have been identified in the Obama corpus, *I* and *the.* This should not, however, create an impression that Barack Obama avoids these function words altogether: *I* occurs 376 times in the corpus and there are 890 instances of *the.* However, they are used with a lower frequency than it would be expected on the basis of the reference corpus. The lower rate of occurrence of *I* might be connected to two features that have been identified as typical of Barack Obama: an inclusive language and indirectness. The first factor is connected to keywords *we, we've* and *are,* which suggest that Obama prefers the first person plural to singular, including himself in the wide audience of the American nation. The avoidance of *I* strongly supports this hypothesis because it shows that the first person plural is used at the expense of the singular form. The analysis of clusters shows that while the use of *I* is on the whole very similar in both corpora, Obama significantly avoids the structure *I don't think [nominal declarative clause],* which occurs at the rate of 225 instances p.m.w. as opposed to 573 p.m.w. in the reference corpus. The second factor, indirectness, is connected to the keywords *nobody* and *is,* which suggests that Obama prefers third person or negative pronouns to express disagreement. Because there are only thirteen uses of *nobody,* the evidence for this explanation is weaker than the evidence for the inclusive *we.* However, the use of *nobody* can be, again, linked to Obama's avoidance of direct disagreement through phrases such as *I don't think.*

Negative keyword: the

The avoidance of *the* appears particularly difficult to analyze due to its extreme frequency of use and variability of contexts, which are both linked to its obligatory nature in many syntactic situations and semantic contexts. As a preliminary step, it will be assumed that there are two possible reasons for the statistical discrepancy between the Obama corpus and the reference corpus. First, it may be the case that Obama simply avoids certain specific phrases in which the definite article is always obligatory. For instance, if the data suggest that Obama avoids the proper name *the United States* and prefers the inclusive *we,* it might contribute to the explanation. However, it seems unlikely that there are a few specific phrases which would cause the difference in the range of hundreds of occurrences. Therefore, it is more plausible to hypothesize that there is an underlying difference in the use of structures connected to the concept of definiteness. For instance, the data might reveal that Obama is less likely to refer to the previous co-text *(a better way to approach the issue - RC)* or less likely to modify nouns in ways which require the use of definite article (hypothetical examples: *the most important thing, the very top - RC).*

With regard to specific phrases, the most frequent clusters[53] including *the* in the reference corpus are *the United States, the American people, one of the, in the world, the vice president* and *president of the* and *the federal government.* Unfortunately, **Table 29** demonstrates that these phrases cannot explain the statistical discrepancy in the use of *the:* in some cases, the frequencies were similar or even greater in the Obama corpus and all these phrases are responsible for less that 10% of all occurrences of *the* in either corpus.

Phrase	OC	OC p.m.w.	RC	RC p.m.w.
the United States	24	1 080	267	1 824
the American people	23	1 035	141	963
one of the	21	945	120	820
in the world	7	315	92	629
the vice president	0	0	85	581
the federal government	2	90	50	342
president of the	8	360	56	383

Table 29: The most frequent clusters including *the*

Therefore, it is necessary to investigate the other possibility, i.e. underlying differences linked to the syntactic and semantic patterns. To analyze these patterns, it was necessary to use the tagged versions of the corpora and look at the distribution of *the* among various syntactic patterns. The choice of the syntactic categories used in this breakdown was limited by the tags used by the software. For instance, TreeTagger does not account for ad hoc conversion of nouns and marks all structures with nominal premodification, such as *the health care* or *the auto industry (OC),* as *[determiner] [noun] [noun],* which inflates the category of *the [noun]* if the syntactic perspective is preferred. The numbers in the table were adjusted for this deficiency, which means that the first category of *the [general noun]* does not include the

[53] An arbitrary limit of 50 occurrences in the reference corpus was selected for this type of analysis. The relatively high number can be justified by the current method, which is to focus on a few specific but very frequent phrases that could explain the difference between the Obama corpus and the reference corpus.

cases where that noun was used as a pre-modifier. The category “other” is included mainly to demonstrate that there is no significant hidden pattern responsible for the discrepancy, as these unanalyzed miscellaneous constructions[54] comprise only less than 4% of all occurrences in both corpora.

Pattern	**OC**	**OC p.m.w.**	**RC**	**RC p.m.w.**	**Log-likelihood**
the [common noun]	452	20 337	4214	28 792	-54.23
the [proper noun]	74	3 329	912	6 231	-32.20
the [nominal premodification] [noun]	36	1 620	355	2 426	-5.95
the [positive adjective] [noun]	198	8 908	1285	8 780	0.04
the [positive adjective] [other than noun]	58	2 610	347	2 371	0.45
the [comparative adj.]	2	90	12	82	0.12
the [superlative adj.]	29	1 305	230	1 571	-0.94
other	35	1 575	259	1 770	-0.43
total	884	39 773	7614	52 023	-60.57

Table 30: The distribution of *the* among syntactic patterns

It appears that there are only two patterns where the difference between the Obama corpus and the reference corpus is significant on the chosen probability level: *the [common noun]* and *the [proper noun].*

The relative lack of proper nouns in the Obama corpus can be partly explained by the use of the inclusive *we* instead of proper nouns such as *the United States* or *the American people.* It is also possible to speculate that in the debates before 2008, some topics commonly referred to by proper nouns were more prominent than in the debates where Barack Obama was present. The examples from the reference corpus include *the AIDS epidemic, the Congress authorization* or *the Patriot Act.*

As for the combination *the [common noun],* possible reasons for its avoidance include that Obama referred to the previous co-text by the definite article to a lesser degree than the other candidates, avoided some deictic references that are frequently employed by the other candidates or prefers indefinite reference when both options are possible, such as in *and that’s a problem* (OC) instead of *and that’s the problem* – (hypothetical). However, there are no data to support this conclusion, as the indefinite article was not identified as a keyword

[54] To provide an example, the verbal adjectives, such as *the underlying problem (OC),* were tagged as –ing forms of verbs.

and the full-scale analysis of features such as references to co-text or deixis would require a substantial effort perhaps meriting a separate work.

Functional Indicators of Style: Summary of Findings

It has been demonstrated that the discussion of function keywords depends on keyword, cluster or syntactic analysis for two reasons. First, because some grammar words occur with a very high frequency in virtually all genres and registers, it is impractical to go through concordance lines directly. Second, function words are very often the focal points of significant clusters, syntactic patterns or semantic sequences. It has been shown that keywords that convey very little autonomous meaning can serve as the starting points for a deeper analysis revealing not only syntactic idiosyncrasies but also pragmatic features, such as indirectness, formality or inclusiveness.

5.1.4. Keywords after Lemmatization

As has been mentioned in the chapters dedicated to methodology, the analysis of Obama's idiolect has been based on the unlemmatized version of the corpora. The main advantage of this approach is that it revealed significant contracted forms such as *we've* and *here's* as well as a key plural form *advisers.* However, it is still useful to consider the keyword list based on lemmata, which might provide new information. From the methodological perspective, it is interesting to explore the effects of lemmatization in order to evaluate which approach is more suitable for keyword analysis. The following table demonstrates that lemmatization has not had a major effect on the keyword list.

N	Key word	Freq.	OC %	RC. Freq.	RC. %	Keyness
1	MCCAIN	117	0.51	123	0.08	176.67
2	SENATOR	115	0.5	228	0.15	93.14
3	THAT	791	3.43	3,769	2.49	65.49
4	WE	643	2.79	3,034	2	55.99
5	TRUE	25	0.11	24	0.02	40.2
6	GO	244	1.06	1,008	0.66	39.02
7	NOBODY	17	0.07	9	0	37.9
8	PROVIDE	37	0.16	62	0.04	36.63
9	MAKE	166	0.72	630	0.42	35.85

10	HAVE	614	2.67	3,123	2.06	32.98
11	SOME	81	0.35	247	0.16	31.46
12	ENERGY	44	0.19	96	0.06	31.17
13	POINT	38	0.16	79	0.05	28.82
14	FINANCIAL	13	0.06	7	0	28.76
15	EMPLOYER	15	0.07	11	0	28.47
16	POLICY	46	0.2	115	0.08	26.3
17	LAST	50	0.22	132	0.09	25.97
18	POTENTIALLY	6	0.03	0	0	24.31
19	ADVISER	6	0.03	0	0	24.31
20	GET	188	0.82	824	0.54	23.38
21	DEAL	33	0.14	72	0.05	23.37
22	SURE	63	0.27	198	0.13	22.82
23	IRAN	27	0.12	52	0.03	22.66
24	PAKISTAN	17	0.07	21	0.01	22.57
25	IMPORTANT	56	0.24	168	0.11	22.53
26	ADDITIONAL	16	0.07	20	0.01	21.03
27	CRISIS	20	0.09	32	0.02	20.81
28	JUST	93	0.4	350	0.23	20.59
29	OIL	31	0.13	71	0.05	20.42
30	STRAIN	5	0.02	0	0	20.26
31	MUDDLE	5	0.02	0	0	20.26
32	SO	117	0.51	476	0.31	19.76
33	AMERICA	24	0.1	364	0.24	-20.16
34	HE	84	0.36	943	0.62	-25.65
35	WILL	73	0.32	891	0.59	-30.83
36	I	452	1.96	3,883	2.56	-31.48
37	THE	890	3.86	7,620	5.02	-61.77

Table 31: Keywords based on the lemmatized corpora.

There are several differences between the two keyword lists. First, lemmatization eliminated the contracted forms *we've* and *here's*. However, whereas the personal pronoun *we* remained on the list and moved upwards as a result, *here* disappeared from the list, which suggests that only the combination *here's* was significant in Barack Obama's speech. The forms *is* and *are* disappeared as well, since the base form of the verb *to be* was not identified as a keyword.

From the new additions, the noun *strain* appears only five times in the entire sample, which suggests a very slight lexical preference. However, *America, he* and *will* were identified as negative keywords. Because the most common alternative reference by a proper noun *the United States* is also used with a lower relative frequency in the Obama corpus than in the reference corpus (944 and 1398 occurrences p.m.w., respectively), it can be concluded that Obama prefers the inclusive pronoun *we* to the direct reference by proper nouns when he speaks about his country. With regard to the keyword *he,* it was used to refer to Joe Biden, Barack Obama's running mate, in 11% of all instances in the Obama corpus. In 59%, it was used to refer to the opposing candidate and in 40%, the pronoun referred to other people, mostly G. W. Bush. Because *he* clearly refers to opponents in the majority of instances, this negative keyword suggests that Obama avoided direct criticism and blame. Moreover, this observation can be connected to the keywords *nobody, we* and *going (to)*, as Obama used *nobody* to distance himself from certain claims *(nobody is denying),* the first person plural to present past failures as a collective phenomenon *(we hadn't finished the job in Afghanistan)* and *going to* suggests a focus on the future rather than the past, which also correlates with the avoidance of *he.* Finally, the negative keyness of *will* can be linked to the positive keyness of *going to*: Obama presents the future using the modal auxiliary, which can be perceived as a lexico-grammatical preference, but it also suggests that Obama presents the future as a predictable outcome of current tendencies.

To conclude, lemmatization did not significantly influence the keyword list. Because the keyness of *we've* could have been discussed on the basis of the lemmatized corpora as well due to the presence of *we,* the only significant differences were the keywords *here's, is* and *are* that only appeared using the unlemmatized corpora and *strain, America(-), will(-)* and *he(-),* which only appeared using the lemmatized corpora. Because the previous discussion showed that all these keywords contribute to Obama's linguistic profile, it is reasonable to conclude that both methods should be used as complementary.

5.1.5. Parts-of-Speech Analysis

The final step of the analysis of Obama's linguistic profile utilizes TreeTagger to convert both corpora to lists of parts of speech that replace individual words. For instance, the first sentence of the Obama corpus, *Well, Allan, thank you very much for the question,* is converted to a code RB, NP, VVP PP RB RB IN DT NN SENT. While one the one hand, it is no longer possible to analyze individual concordance lines, on the other hand, it is possible to look at the distribution of the parts of speech and compare the two corpora using log-likelihood as the measure of significance.

N	Key word	OC #	OC %	RC #	RC. %	Keyness
1	*that* as a subordinating conjunction	434	1.77	1 664	1.03	92.44
2	-ing form of verbs	607	2.48	2 707	1.67	71.76
3	*'ve* or *have*	337	1.37	1 619	1	26.68
4	*'s* or *is*	624	2.54	3 393	2.09	19.55
5	*lexical verbs – past tense*	269	1.1	2 447	1.51	-27.38
6	*determiner*	1 942	7.92	14 608	9.02	-32.66
7	full stop	1 175	4.79	9 449	5.83	-44.94

Table 32: Key parts of speech in the Obama corpus.

It is remarkable that almost all these discrepancies have already been identified in some form in the previous chapters. *That, we've* and *is* have been discussed as positive keywords in the Obama corpus. The keyness of *–ing* forms can be largely attributed to the key auxiliary *to be going to,* which comprises 30% of this class. The relative lack of verbs in past tense appears as a new observation which, however, correlates very well with the focus on the future discussed under the auxiliary construction *to be going to.* The lower rate of occurrence of determiners is clearly attributable to the negative keyness of *the*. The only substantially new information that the parts-of-speech analysis provides is that full stop is a negative key element in the Obama corpus, which means that Barack Obama on the average uses longer sentences. This difference can be calculated as 18.94 words per sentence in the Obama corpus in comparison to 15.51 in the reference corpus. To conclude, this supplementary method is useful when the aim is to verify the findings obtained from the keyword analysis and to compare the length of sentences.

6. Summary of Findings

The summary the findings will be divided for into two parts the practical purposes. The first part will discuss the conclusions on the methodology used in this paper. Because the idiolect analysis of spoken English performed by corpus-driven methods represents a relatively new research ground and the methodology is far from established, the **Chapter 6.1.** is dedicated to a summary of the most important findings pertaining to methods that were employed in this work with an attempt at their brief evaluation. Because the research goal was to characterize Obama's idiolect, **Chapter 6.2.** will present the most important features of his individual profile in a structured and compact way.

6.1. Summary of Findings: Methodology

From the methodological perspective, it has been shown that it is possible to work with relatively small and specialized corpora. While the size of the reference corpus, which was approximately seven times as large as the study corpus, did not appear an issue at any point, the size of the Obama corpus itself was to some degree deficient when the analysis focused on rare expressions. For instance, while the frequency limit for clusters is usually set to 10 – 40, a single occurrence of any given phrase corresponds to 45 uses per million words in the Obama corpus due to its small size. To avoid using only a very few instances to form definite conclusions, the absolute frequency limit for clusters was set to 5, which corresponds to 225 uses per million words, a rather large number. Moreover, the rare expressions also made the collocation analysis difficult and prompted the decision to work with relative frequency of co-occurrence as the statistical measure even though more representative methods exist. However, it is also clear that the results of the present work are not distorted by differences in genre and context. In those cases where it was possible to obtain a large amount of data, such as in the analysis of *make sure, that* or *going to,* the results unambiguously represent the chosen debating candidate's individual speech profile.

This work also demonstrates that keywords are often linked in syntactic structures *(make sure + that, we've + got, we + are),* but there are deeper connections, the interpretation of which goes beyond the syntactic level *(*e.g., *going to* + lack of verbs in past tense = focus on planned/predicted future). While such conclusions partly depend on the researcher's interpretation and some bias may be unavoidable, interconnected characteristics encompassing several layers of language (from lexical preferences or grammatical tendencies to pragmatics) are arguably one of the most valuable goals of corpus-driven research. This work demonstrates that to achieve this goal, it is best not to rely on a single method but attempt to look at all key units from various angles, such as collocation and cluster analysis, semantic sequences or parts of speech.

6.2. Summary of Findings: Obama's Individual Profile

The aim of this chapter is to provide a quick summary of the most important features of Barack Obama's idiolect, which will be divided into the following categories: **topics, discourse organizers, inclusive language, expression of future, expressions influencing the degree of formality, expressions influencing the degree of directness or assertiveness** and **other significant features.** Before proceeding with the summary, this classification requires some justification.

Topics

Topics or **indicators of topic** are a distinct category because from the perspective of individual profile, they are very unlikely to persist as permanent idiosyncratic features. However, the aim of this work was to describe Barack Obama as a presidential candidate in the televised debates and the topics that were identified as prominent in his speech represent a part of this profile.

Topic	**Keywords**	**Specific areas, perspectives or focus**
Employers	*employers*	Employer-based health care system/plan
Energy	*energy*	Alternative energy, energy consumption
Oil	*oil*	Global oil reserves, American usage (consumption) of oil
Financial crisis	*financial, crisis*	The recent global recession
McCain's advisers	*advisers*	McCain's advisers, their claims

Table 33: Topics

Discourse organizers

Discourse organizers, which are defined as expressions that introduce a new topic or perspective (Biber 2006: 142), are listed as a separate category because it has been observed that Obama has an increased tendency towards their use. The fact that five keywords of this type have been identified suggests a general tendency towards an explicit meta-textual organization on Obama's part, which is an important finding from the perspective of his individual textual profile.

Keywords	**Significant structures, clusters or examples**[55]	**Pragmatic function**
make, point	*make a point*	Signals a new argument
last, point	*last point*	Signals the final argument on the topic
just, point	*let me just (make a point)*	Besides introducing a new argument, it also explicitly asks for time while minimizing the intrusion (means of politeness)
	I just want to (make a point)	
here's	*(so) here's what*	Signals a shift to a new perspective or a point of importance
so	*so* in the initial position	Signals a shift to a new perspective or a concluding point

Table 34: Discourse organizers

Inclusive language

Inclusive language refers to the observation that Obama prefers to include himself in the audience and often uses the first person plural when he speaks about the American nation. It is important to note that nine keywords have been linked to this phenomenon, which makes it the most prominent systematic feature of Obama's idiolect.

Keywords	**Significant structures, clusters or examples**	**Type of inclusiveness / Comments**
important	*important for us (to)*	Inclusive *we* in the adjunct of respect or as the subject of an infinitival clause
going	*we are going to* *(-) I/I'm (*)*[56] *going to*	*We* is a frequent subject of *going to* *I* and *I'm* are avoided as subjects Increased politeness by stating obligations as shared
deal	*(we) (*) deal with*	Associated with the first person plural subject
we, are	non-contracted form *we are*	Increased politeness by stating obligations as shared

[55] Unless stated otherwise, all examples in all sections are from the Obama Corpus.

[56] (*) marks one or more non-obligatory elements.

we've	*we've got to*	Emphasis on shared obligation
(-)[57] *he*	*he doesn't display the qualities (RC)*	Avoidance of third person – correlates with the increased use of first person plural
(-) I	*I think it is manmade (RC)*	Avoidance of *I* – correlates with the increased use of first person plural
nobody	*nobody* + verba dicendi	Reference to shared attitudes
McCain	*McCain and I*	Emphasis on agreement

Table 35: Inclusive language

Expressions of future

It has also been discovered that Obama is more focused on the **planned or predicted future** than the other candidates. Besides the future auxiliary ***going to,*** there is additional evidence in the form of the negative key part of speech, verbs in past tense, and the negative keywords *will* and *he*. However, Obama also focuses on the previous administration using the cluster *over the last (eight years),* which contradicts the future orientation.

Keywords	**Significant structures, clusters or examples**	**Pragmatic function / Relation to future**
going	*we are going to*	Focus on the planned/predicted future
going	*going to have to*	Focus on predicted obligations
(-) *will*		Correlates with focus on the planned or predicted future
(-) verbs in the past tense		Correlates with the focus on the future
(-) *he*		Correlates with the focus on the future, as it is used to refer to the opponents
last	*over the last (eight years)*	Focus on previous administration

Table 36: Expressions of future

Expressions influencing the degree of formality

With regard to **formality,** it appears that Obama used a more formal language than the other candidates. However, there is also some contradictory evidence in the form of the rather informal significant cluster *we've got to,* which is, therefore, marked as gray in the table below.

[57] (-) marks a negative keyword.

Keywords	**Significant structures, clusters or examples**	**Pragmatic function / Relation to formality**
that	*that*-clauses	Fewer omissions of *that:* formal
we, are	non-contracted form *we are*	Contractions less frequent: formal
we've, got	*we've got to*	Informal semi-modal, contracted form

Table 37: Expressions influencing the degree of formality

Expressions influencing the degree of directness or assertiveness

The **degree of directness or assertiveness** is a broad category that includes mainly expressions related to the hedging or boosting. However, it has to be recognized that televised political debates are by purpose and custom quite confrontational. For this reason, this category includes all significant expressions that heighten or lower the degree of confrontation. The evidence gathered in this work suggests that Obama is significantly less direct and confrontational than the average candidate, even though there is one exception, again marked by gray background.

Keywords	**Significant structures, clusters or examples**	**Pragmatic function / Relation to directness**
we, are, we've	*we haven't caught Bin Laden, we haven't adequately funded veterans, etc.*	Avoids denunciation of individuals – presents past failures as collective
nobody	*nobody likes taxes*	Indirect expression of opinions
some	*some of the, there are some some things*	Hedging, limiting the scope of statements or vagueness in reference
last	*over the last (eight years)* *(-) in the last (eight years)*	Semi-direct criticism (does not name particular people, but invokes previous administration)
important	*I think it's important for X to Y*	Marks the claim as an opinion / hedging
potentially	*$18 billion is, potentially, a lot of money*	Hedging
(-) he	*he doesn't mention he voted against (RC)*	Indirectness, as *he* most commonly refers to the opponent
true	*that's absolutely true*	Concession or common ground
just	*let me just (make a point)*	Politeness (asks for a time to respond and attempts to minimize the intrusion)
=McCain	*Senator McCain and I*	Mostly expresses agreement
true	*that's not true*	Direct refutation

Table 38: Expressions influencing the degree of directness or assertivness

Other significant lexical or grammatical features

The final category summarizes other observations that did not fit into any other category and thus do not represent a significant systematic tendency. It should also be noted that there are some overlaps where there was more than one aspect to a keyword or a significant cluster, such as in the case of *we are going to,* which uses inclusive *we* and at the same time refers to planned or predicted future. Finally, the following summary does not include any statistical information, which makes it difficult to estimate the comparative significance of the individual phenomena, but it is hoped that the current format will facilitate orientation and provide a better overview.

Keyword	**Significant structures, clusters or examples**	**Comments on the semantics or pragmatic function**
provide	*provide [something] to [someone]*	Associated with the topic of health care
additional	*additional tax cuts/breaks*	Often followed by specific numbers
potentially	(see Chapter 5.1.2)	Non-standard use (word order, redundancy)
policies	*failed policies* *tax policies*	Negative semantic prosody
is	pseudo-clefts	Pseudo-clefts – emphasis by syntactic means
is, imporant	is [adjective], it is [adjective]	Predicative use of adjectives; connected to the keyword *important* – explicit evaluation
	it is (anticipatory)	Anticipatory it; connected to the keyword *important* – evaluative focus
make, sure	*make sure*	Appeal to a shared obligation
[full stop]	Sentence length	Obama uses longer sentences (18.94 in comparison to 15.51)

Table 39: Other features of Obama's idiolect

7. Summary in Czech (Shrnutí práce v češtině)

7.1. Cíl výzkumu a základní východiska

Diplomová práce se zaměřuje na výzkum idiolektu, tedy jazykové variety konkrétního mluvčího, s využitím metod korpusové lingvistiky. Za předmět zkoumání byl zvolen současný americký prezident Barack Obama v konkrétní řečové situaci předvolebních televizních debat v roce 2008. Práce není motivována politicky a veřejný činitel byl vybrán pouze pro dostupnost přepisu debat a implicitní souhlas s analýzou projevu. Obamův idiolekt je zkoumán na základě kontrastu s jeho tehdejším protikandidátem Johnem McCainem a všemi ostatními kandidáty na úřad prezidenta a viceprezidenta v letech 2000, 2004 a 2008. To, že se jazykové rysy Baracka Obamy porovnávají s ostatními kandidáty v identické a velmi specifické řečové situaci, zajišťuje, že výsledkem výzkumu je individuální jazykový profil Baracka Obamy, nikoliv charakteristika daného žánru, tedy předvolebních politických debat v USA nebo v širším smyslu současného amerického politického diskursu. Kromě sestavení individuálního řečového projevu si práce klade za cíl ověřit využitelnost současných metod korpusové lingvistiky pro studium idiolektu zejména s ohledem na omezenou velikost studovaného i referenčního korpusu skutečnost, že výzkumu idiolektu je v současné korpusové lingvistice věnována spíše menší pozornost.

7.2. Metodologie

Výzkum se opírá o metody korpusové lingvistiky ve smyslu anglického termínu *corpus-driven* (výzkum determinovaný korpusovými metodami)*,* který na rozdíl od obecnějšího termínu *corpus-based* (výzkum založený na korpusových metodách) specifikuje přístup, který je co nejméně zatížen teoretickým rámcem a subjektivním pohledem badatele. Dalším rysem tohoto přístupu je kombinace kvalitativních a kvantitativních metod.

Pro porovnání Obamova idiolektu s „průměrným“ účastníkem předvolebních debat byly sestaveny dva korpusy, které se dál budou označovat jako **Obamův korpus** a **referenční korpus**. Obamův korpus obsahuje všechny řečové projevy Baracka Obamy ze tří televizních debat s Johnem McCainem. Referenční korpus obsahuje všechny řečové projevy ostatních kandidátů z let 2000, 2004 a 2008 z celkem dvanácti debat. Ze všech přepisů byly odstraněny otázky publika i moderátora, stejně jako meta-textové poznámky typu „(sic)“, „(ticho)“ nebo

„(smích)“. Obamův korpus obsahuje 22 013 individuálních slov (tokenů), z toho 2 539 různých slov (typů). Referenční korpus obsahuje 145 266 tokenů a 6 924 typů.

Východiskem pro konstrukci Obamova idiolektu a jedním z hlavních kvantitativních rysů práce je seznam klíčových slov. Termín klíčová slova se používá ve specifickém smyslu statistické významnosti, která obecně v korpusové lingvistice může být měřena různými veličinami. V práci jsou blíže vysvětleny statistické metody **chi-squared** a **log-likelihood**[58]. Ve zkratce se tyto veličiny dají popsat jako míra statistické významnosti rozdílu relativních frekvencí užívání daného slova v cílovém (studijním) a referenčním korpusu. Ke konkrétním hodnotám těchto veličin se obecně dá přiřadit pravděpodobnost, která vyjádří, zda je rozdíl ve frekvenci jejich užívání pouze náhodný. V praxi se však běžně postupuje tak, že za klíčová slova se považují všechna ta slova, jejichž statistická významnost měřená jednou z těchto veličin převyšuje zvolenou úroveň pravděpodobnosti náhody, například 1%. Klíčová slova lze navíc seřadit podle jedné nebo druhé veličiny, protože určují míru „klíčovosti“. Pro konstrukci seznamu klíčových slov byla na základě zkušeností jiných badatelů a obecných doporučení zvolena statistická metoda log-likelihood jakožto veličina vhodnější pro menší korpusy. Metody chi-squared a log-likelihood však obecně produkují velmi podobné výsledky, k čemuž dospěla i tato práce.

Při zvolené úrovni pravděpodobnosti 0.001% bylo za pomocí korpusového softwarového nástroje WordSmith Tools nalezeno 37 klíčových slov, která se následně blíže analyzovala kombinací kvalitativních a kvantitativních metod. U každého klíčového slova byl zkoumán kontext pomocí tří technik: **kolokací**, **shluků**[59] a **konkordančních řádků.**

Kolokace obecně označuje tendenci slov vyskytovat se v okolí jiných slov s vyšší pravděpodobností, než by odpovídalo náhodě. Práce zkoumá kolokace jednotlivých klíčových slov pro objasnění jejich konkrétního užití a kontextu. Například nejfrekventovanější kolokace adjektiva *financial* (finanční) je v Obamově korpusu podstatné jméno *crisis* (krize) a síla asociace vypovídá o tom, že klíčové slovo *financial* Obama užívá

[58] Chi-square by se dalo přeložit jako druhá mocnina chí nebo uvádět řecným písmenem se symbolem mocniny χ^2. Log-likelihood by se dalo překládat jako logaritmická pravděpodobnost. S ohledem na to, že tyto veličiny jsou v korpusové lingvistice užívány jako obecně známé technické termíny, shrnutí se bude držet původních anglických názvů.

[59] Tímto *ad-hoc* termínem se ve shrnutí označují opakující se slovní spojení, která se v anglické literatuře běžně označují jako „lexical bundles“, „clusters“ nebo„N-grams“.

specificky v narážce na globální recesi. V práci se zkoumaly nejenom bezprostřední kolokace, ale všechny lexikální asociace, které se nacházely maximálně čtyři slova nalevo a napravo od zkoumaného výrazu.

Shluky označují často se opakující řetězce tří a více slov, které se nacházejí v okolí zkoumaného výrazu, nebo ho obsahují. Jako příklad lze uvést *make a point* (doslova „uvést argument“, ale v praxi se preferují idiomatičtější překlady), shluk spojený s klíčovými slovy *make* a *point*. Význam shluků spočívá kromě přesnější charakterizace užití klíčových slov i v tom, že opakující se slovní spojení jsou obecně méně frekventovaná než jednotlivá slova, a proto má jejich případné častější užívaní větší význam při popisu individuálního řečového profilu.

Kontext se dá zkoumat i přímo s pomocí konkordančních řádků, jejichž výpis je běžnou funkcí korpusového softwaru. Konkordanční řádky zobrazují všechna použití klíčového slova i s okolím. Badatel tak může posoudit kontext, v němž se slovo vyskytuje, mnohem přesněji než s využitím předcházejících metod. Lze také vyhodnotit opakující se sémantické vzorce nebo sémantickou prosodii výroků s klíčovými slovy navzdory variacím na lexikální nebo frazeologické rovině.

Primární seznam klíčových slov vychází z nelematizovaných korpusů (původních textů bez převodu všech jednotek na kanonický tvar). Toto rozhodnutí se opírá o skutečnost, že i stažené formy, množná čísla a nekanonické tvary osobních zájmen jsou součástí Obamova idiolektu. Bylo například zjištěno, že Barack Obama ve srovnání se zbylými kandidáty častěji užívá staženou formu *we've* (*máme* nebo část konstrukce s významem *musíme*), kterou při absenci lematizace korpusový software považoval za jedno klíčové slovo. Jistý prostor je však věnován i sekundárnímu seznamu klíčových slov, který vychází z lematizovaných korpusů, a srovnání výsledků obou postupů. Kromě automatické lematizace se v práci využilo i morfologického značkování, které ke každé jednotce přiřadilo slovní druh, což usnadňuje analýzu multifunkčních slov, jako je například *that.*

Samotná klíčová slova lze rozdělit do několika kategorií. Tradiční dělení rozlišuje mezi **vlastními jmény, lexikálními** a **gramatickými klíčovými slovy** a často jej provází předpoklad, že lexikální klíčová slova se vztahují k tématům promluvy (v angličtině se také nazývají *aboutness keywords,* tedy doslova klíčová slova určující, „o čem“ text je) a gramatická klíčová slova mají vliv na styl textu. Tento předpoklad někteří výzkumníci považují za neopodstatněný a neosvědčil se ani v této práci, protože některá plnovýznamová slova, jako je například *just, make* nebo *last,* se nevztahují ke konkrétnímu tématu a necharakterizují obsah textu, nýbrž spolu s neplnovýznamovými slovy charakterizují styl promluvy. Proto bylo v této práci zvoleno rozdělení tradiční kategorie lexikálních klíčových slov na **indikátory témat** a **lexikální indikátory stylu.**

7.3. Výsledky

Seznam klíčových slov lze nalézt v tabulce na začátku páté kapitoly a strukturovaný přehled základních rysů Obamova idiolektu v šesté kapitole. Jelikož práce odhalila v Obamově projevu několik systematických tendencí, uvede tato část shrnutí každé z nich i s příklady a odkazy na příslušná klíčová slova.

Co se týče **témat**, Barack Obama se ve srovnání s ostatními kandidáty víc věnoval zaměstnavatelům (*employer*), energii (*energy*), ropě (*oil*), finanční krizi (*financial crisis*) a McCainovým poradcům (*advisers*). Klíčové slovo zaměstnavatel je spjaté s tématem zdravotní péče poskytované zaměstnavatelem a je možné také poznamenat, že poradci republikánského kandidáta byli v Obamově korpuse zmíněni pouze pětkrát a jako klíčové slovo bylo toto podstatné jméno vyhodnoceno jen díky kontrastu s referenčním korpusem, kde se vůbec nenachází. Obecně se lze domnívat, že témata nepředstavují trvalý jazykový rys konkrétního řečníka a charakterizují pouze zaměření promluvy či diskuse.

U Obamy je patrná zvýšená tendence **k explicitní organizaci textu**. Přispívají k tomu fráze jako *make a point,* jako i spojení klíčového slova *point (bod, argument)* s klíčovými adjektivy *last (poslední)* a *just (jen).* Obama se často odvolává na určitou část diskurzu frázemi *let me just make a point (dovolte mi říct ještě jednu věc)* nebo *I just want to make a point (rád bych ještě něco řekl).* Tyto fráze lze také chápat jako zdvořilostní formule, kterými řečník přímo žádá o svolení moderátora. Do této kategorie může být zahrnuto také klíčové slovo *so* a staženina *here's,* které se často objevují spolu ve spojení *so here's what* ve funkci upozornění publika na následující důležitou myšlenku nebo shrnutí, a které se často nachází na začátku věty. Jako příklad lze uvést s*o here´s what my plan does – takže můj plán docílí tohle.*

Obama se také častěji stylizuje jakou součást širšího publika – diváků nebo občanů Spojených států. Tuto tendenci tato práce nazývá termínem **inkluzivní způsob promluvy.** Základní charakteristikou tohoto jevu je zvýšená tendence k užívání první osoby množného čísla *(we are – my jsme; we've – máme, musíme),* a naopak snížená frekvence první a třetí osoby jednotného čísla *(I – já; he – on).* Kromě těchto základních rétorických prostředků k navazování vztahu s publikem využívá Obama ke stejnému cíli i negativní osobní zájmeno *nobody (nikdo)*, pomocí něhož se nepřímo vymezuje proti jistým názorům *(nobody says that – to nikdo neříká),* nebo jím vyjadřuje sdílené postoje *(look, nobody likes taxes – podívejte se, daně nemá rád nikdo).* Za zmínku stojí i klíčová slova, která sama o sobě neimplikují sklony k inkluzi, ale kvalitativní analýza ukázala, že se pojí s inkluzivními strukturami. Například sloveso *deal* se často vyskytuje s podmětem *we, going* se objevuje ve frázi *we are going to* (vyjádření plánované budoucnosti) a *important* se vyskytuje s infinitivní vazbou *for us to ...* nebo modifikací *for us – důležité, abychom .../ pro nás.* Obama také často zdůrazňuje programový průnik nebo shodu se svým oponentem – *McCain and I agree (s McCainem se shodneme, že...).*

Dalším systematickým prvkem Obamova idiolektu je **důraz na předvídatelnou nebo plánovanou budoucnost**. To se projevuje zejména zvýšeným užíváním pomocného slovesa *going to* a sníženou frekvencí základního prostředku vyjádření prosté budoucnosti *will.* Za zmínku stojí i skutečnost, že Obama často vyjadřuje budoucí společné povinnosti *(we're going to have to – budeme muset),* na čemž se dá demonstrovat jisté prolínání základních rysů jeho idiolektu. Kromě toho obecně užívá méně sloves v minulém čase a vyhýbá se také osobnímu zájmenu *he,* které se v žánru televizních debat často vztahuje k předchůdcům nebo oponentovi. Protichůdným rysem je však důraz na předcházejících osm let, tedy na dvě volební období prezidenta Bushe, ke kterým se Obama často vrací slovy *over the last eight years – za posledních osm let.*

V televizních debatách se Obama vyznačuje **formálnějším projevem**, což lze demonstrovat na dvou jevech. Prvním je užívaní uvozujícího *that (že, který),* které lze v angličtině u některých typů vedlejších vět vypustit, ale Obama ho ve svém projevu ponechává. Druhým je již zmiňovaná tendence nestahovat *we are* do formy *we're.* Protichůdným rysem v této kategorii je časté užívaní neformálního semimodálního slovesa *got to,* které se objevuje zejména ve frázi *we've got to – musíme.*

Dalším rysem Obamova idiolektu je **nižší míra přímočarosti, konfrontačního postoje a asertivity.** Do této oblasti lze zařadit již zmiňovanou preferenci první osoby množného čísla, v tomto případe v kontextu přiznávání chyb nebo selhání. Obama například při diskusi o problémech Ameriky často místo kritiky konkrétních osob užívá zájmena, kupříkladu *we haven't caught Bin Laden – pořád jsme nechytili Bin Ládina,* nebo *we haven't adequately funded veterans – nepodařilo se nám přiměřeně zajistit válečné veterány.* Tento rys se projevuje i v užívání záporného zájmena *nobody,* které již bylo zmiňováno v kategorii inkluzivního projevu. Obama působí méně asertivním dojmem také kvůli explicitním vymezením tvrzení jako subjektivních názorů, omezením jejich platnosti nebo záměrnou neurčitostí, což odborná anglická literatura označuje jako *hedging.* Do této kategorie spadají klíčová slova *some (nějaký, nějací, někteří a pod.), potentially (potenciálně, možná), just* v konstrukci *let me just make a point – dovolte mi jen něco říct* a *important* v konstrukci *I think it's important – podle mně je důležité.* K tomuto rysu přispívá i vyhýbání se osobnímu zájmenu *he,* které se v žánru televizních debat užívá k označení oponenta nebo předchůdců, jakož i časté užívaní *true* na vyjádření souhlasu *(that's absolutely true – to je úplná pravda).* Obama navíc vyjadřuje přímý souhlas s oponentem *(McCain and I agree – s McCainem se shodneme na tom, že...)* Je však zároveň nutné zmínit, že Obama adverbium *true* využívá téměř ve stejné míře k přímému vyvracení různých tvrzení, což je protichůdná tendence.

Na závěr strukturovaného popisu Obamova idiolektu se uvádí **ostatní rysy**, které nepatří do žádné ze jmenovaných systematických tendencí. Patří sem klíčová slova *provide – poskytnout*, které se pojí se zdravotní péčí; *additional* v kontextu daňových výjimek; *policies,* které se užívá s negativní semantickou prozodií, a fráze *make sure* ve významu zajistit. Obama také častěji užívá anticipačního *it*, které se často pojí s klíčovým adjektivem *important: it's important that we – je důležité, abychom...* K dalším Obamovým rysům patří pseudo-vytýkací konstrukce a nestandardní, redundantní užívání slova *potentially.* Na závěr lze zmínit, že Obama mluví v delších větách než průměrný kandidát (18,94 slov na větu ve srovnání s 15,51 v referenčním korpusu).

Po metodologické stránce práce demonstruje, že je možné dospět k relevantním závěrům i na základě velmi malých korpusů. Ukázalo se však, že zejména Obamův korpus svým rozsahem znesnadňuje interpretaci relativní frekvence opakujících se slovních spojení, jelikož i jediné užití libovolné struktury odpovídá přibližně 45 výskytům na milion slov, což je obecně považováno za častý výskyt. Navíc nebylo možné při hledání kolokací využít pokročilejších měřítek jako relativní četnosti společného výskytu, jelikož log-likelihood a jiné veličiny mají tendenci při velmi malých počtech výskytů zkreslovat sílu vzájemné asociace. Navzdory těmto nedostatkům se ukázalo, že i na korpusech menší velikosti lze spolehlivě odhalit významné tendence individuálních řečníků, zejména v případě idiosynkratického opakování jistých frází, konkrétně třeba u Baracka Obamy výše zmíněného *make sure*. Práce také ukázala, že kvalitativní analýza s pomocí několika různých kvantitativních technik umožňuje zjistit nejenom vzájemné souvislosti nebo přímo závislosti klíčových slov (např. *make sure* + *that, we've* + *got, we* + *are*), ale též systematické rysy řečníků v dané situaci, jako je například zaměření na plánovanou budoucnost nebo nižší míra asertivity. Objevení systematických netriviálních tendencí individuálních řečníků s využitím korpusových metod lze považovat za důležitý přínos tohoto přístupu.

References:

Biber, D. (2006). *University Language. A corpus-based study of spoken and written registers.* Amsterdam and Philadelphia: John Benjamins Publishing Company.

Biber, D. et al. (1999). *Longmann Grammar of Spoken and Written English.* Harlow: Longman.

Biber, D., Barbieri, F. (2007). 'Lexical bundles in university spoken and written registers.' *English for Specific Purposes, Vol 26, Issue 3,* pp. 283-304 (2007)

Culpeper, J. (2009). 'Keyness - Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet.' *International Journal of Corpus Linguistics 14:1 (2009)*, 29–59.

Dunning, T. (1993). 'Accurate methods for the statistics of surprise and coincidence.' *Computational Linguistics 19.1 (Mar. 1993)*, 61-74.

Baker, P., Hardie, A., McEnery, T. (2006). *Glossary of Corpus Linguistics*. Edinburgh: Edinburgh University Press. (In the text referred to as Glossary)

Haugen, E. (1972) [1960]. "From idiolect to language". In *E. Scherabon Firchow, K. Grimstad, N. Hasselmo & W. A. O'Neil (Eds.), Studies by Einar Haugen. Presented on the Occasion of his 65th Birthday*. The Hague/Paris: Mouton, 415–421.

Hunston, S. (2002). *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press

Hunston, S. (2008). Starting with the small words – Patterns, lexis and semantic sequences. In *International Journal of Corpus Linguistics 13:3 (2008), 271–295.*

Lee, D. (2008). Corpora and discourse analysis. In *Advances in Discourse Studies.* Oxon: Routledge.

McEnery, T., Wilson, A. (2004). *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press.

Manning, Christopher D., Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* Boston: MIT Press

Mollin, S. (2009). Sandra Mollin: “I entirely understand” is a Blairism - The methodology of identifying idiolectal collocations. In *International Journal of Corpus Linguistics 14:3 (2009)*, 367–392.

Oakes, Michael P. (1998), Statistics for Corpus Linguistics. In *International Journal of Applied Linguistics Volume 10, Issue 2, pages 269–274, December 2000*

Quirk et al. (1985): Comprehensive Grammar of the English Langauge. New York: Longman.

Scott, M., Tribble, C. (2006). *Textual Patterns – Key words and corpus analysis in language education.* Amsterdam/Philadelphia: John Benjamins Publishing Company

Scott, M. (2010). Problems in investigating keyness, or clearing the undergrowth and marking out trails… In *Keyness in Texts,* ed. Marina Bondi and Mike Scott. Amsterdam/Philadelphia: John Benjamins Publishing Company

Semino, E. and Swindlehurst (1996) Metaphor and Mind Style in Ken Kesey's One Flew Over the Cuckoo's Nest. In *Style, 30, 1, 143-166.*

Stefanowitsch, A., Gries S. (2003). Collostructions: Investigating the Interaction of Words and Constructions. *In International Journal of Corpus Linguistics 8:2 (2003)*, 209-243.

Stubbs, M. (2010). Three concepts of keywords. In *Keyness in Texts,* ed. Marina Bondi and Mike Scott. Amsterdam/Philadelphia: John Benjamins Publishing Company

Wierzbicka, A. (1997). *Understanding cultures through their key words: English, Russian, Polish, German, and Japanese.* New York: Oxford University Press.

Xiao, Z., McEnery, T. (2005): Two Approaches to Genre Analysis. In *Journal of English Linguistics, Vol. 33 / No. 1, March 2005* 62-82

Sources

The Telegraph: The impact of Primary Colors
(http://www.telegraph.co.uk/culture/books/8240675/The-impact-of-Primary-Colors.html)

Review of Primary Colors by Anonymous (Joe Klein)
(http://www.bearcave.com/bookrev/primary_colors.htm)

Don Foster enlightens readers with 'Author Unknown'
(http://articles.cnn.com/2000-12-06/entertainment/foster.anonymous_1_funeral-elegy-shakespeare-sleuth?_s=PM:books)

New York Times: Columnist's Mea Culpa: I'm Anonymous
http://www.nytimes.com/1996/07/18/us/columnist-s-mea-culpa-i-m-anonymous.html?pagewanted=2&src=pm

Daily Infographic (http://dailyinfographic.com/)

David McCandless: The beauty of data visualization
(http://www.vizworld.com/2010/08/tedtalks-david-mccandless-beauty-data-visualization/)

Tag Clouds Evolve: Understanding Tag Clouds
(http://www.joelamantia.com/ideas/tag-clouds-evolve-understanding-tag-clouds)

Engineering Statistics Handbook: Critical Values of the Chi-Square Distribution
(http://www.itl.nist.gov/div898/handbook/eda/section3/eda3674.htm)

David Meerman Scott: Ten Marketing Lessons from the Barack Obama Presidential Campaign (http://www.webinknow.com/2008/11/ten-marketing-lessons-from-the-barack-obama-presidential-campaign.html)

Software, Tools and Dictionaries

WordSmith Tools 5.0.0.334 (Hic fecit Michael Scott)

TreeTagger - a language independent part-of-speech tagger (Developed by Helmut Schmid, Version for MS Windows)
(http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/)

Wordle (http://www.wordle.net/)

Log-likelyhood calculator (http://ucrel.lancs.ac.uk/llwizard.html)

Merriam-Webster’s 11-th Collegiate Dictionary, Version 3.0 (2003)

Longman Dictionary of Contemporary English, 5-th Version

Sources for the corpora

Debate Transcripts by the Commission on Presidential Debates (years 2000, 2004 and 2008)
(http://www.debates.org/index.php?page=debate-transcripts)