Univerzita Karlova

Filozofická fakulta

Ústav anglického jazyka a didaktiky

Bakalářská práce

Vladimíra Krajcsovicsová

Keywords and Frequent Words of J.D. Salinger's *The Catcher in the Rye*

Klíčová a frekventovaná slova v románu J.D. Salingera *The Catcher in the Rye*

Praha 2017 Vedoucí práce: doc. PhDr. Markéta Malá, Ph.D.

Ráda bych poděkovala vedoucí své bakalářské práce, doc. PhDr. Markétě Malé, Ph.D., za její ochotu, trpělivost, cenné rady a vstřícnost.

Prohlašuji, že jsem bakalářskou práci vypracovala samostatně, že jsem řádně citovala všechny použité prameny a literaturu a že práce nebyla využita v rámci jiného vysokoškolského studia či k získání jiného nebo stejného titulu.

I declare that the following BA thesis is my own work for which I used only the secondary literature that is listed in the resources. This thesis was not used as a part of any other university study, nor was it used to gain a different university degree.

V Praze dne 1. 8. 2017 ……..…..………..…..…………

Abstract

This BA thesis aims to perform a corpus-stylistic analysis of J. D. Salinger's novel *The Catcher in the Rye*. The starting point for this analysis is a list of frequent and key words of Salinger's novel which are generated on the basis of comparison of frequency information in two corpora. The reference corpus consists of five novels published between 1996 and 2014 which share some fundamental similarities with Salinger's novel (i.e. the same target audience, informal language, first person narration).

The theoretical part focuses predominantly on the relevant research in the area of corpus stylistics and at the same time, it provides definitions for the basic terms which are applied in the practical part. The methodology then introduces the texts which are employed for the analysis, as well as the software used, along with its main functions. In the analytical part, top hundred keywords are sorted into three groups (proper names, grammatical and lexical words) and they are subject to further examination, focusing predominantly on their collocations and n-grams.

This analysis uncovered not only the features of the idiolect of the main hero of Salinger's novel, but also some basic characteristics of teenage language in use. At the same time, this research suggests that some of these characteristics changed over the course of the last sixty years.

Key words: keywords, frequency, corpus linguistics, stylistics, informal language

Abstrakt

Bakalářská práce si klade za cíl podat korpusově-stylistickou analýzu románu J. D. Salingera *The Catcher in the Rye* (*Kdo chytá v žitě*). Analýza se opírá o seznam frekventovaných a klíčových slov Salingerova románu, která jsou vygenerována na základě porovnání frekvencí slov v cílovém a referenčním korpusu. Referenční korpus tvoří pět knih, které vyšly mezi lety 1996 a 2014, a které jsou určeny stejné věkové skupině a sdílí se Saligerovým románem jisté charakteristcké prvky (zejm. neformální jazyk a vyprávění v první osobě).

Teoretická část popisuje především hlavní přínosy elektronické analýzy (literárních) textů a dále definuje nejdůležitější pojmy, které budou dále využívány v praktické části. Metodologická část pak uvádí konkrétní texty, které tvoří použité korpusy, a hlavní funkce softwaru, který byl pro analýzu využit. V praktické části bylo prvních sto klíčových slov rozděleno do tří skupin (vlastní jména, slova gramatická a lexikální) a tato slova byla následně podrobena dalšímu zkoumání, přičemž důraz byl kladen především na jejich kolokace a n-gramy.

Analýza identifikovala jak charakteristické rysy idiolektu hrdiny Salingerova románu, tak i rysy jazyka teenagerů obecně. Práce zároveň nažnačuje, že se některé tyto rysy během posledních šedesáti let změnily.

Klíčová slova: klíčová slova, frekvence, korpusová lingvistika, stylistika, neformální jazyk

Table of Contents

List of Tables and Figures

List of Abbreviations and Symbols

POS – Parts of Speech
KWIC – Key Word in Context
BNC – British National Corpus
Catcher – The Catcher in the Rye

1 Introduction

This BA thesis is going to perform a corpus-stylistic analysis of J. D. Salinger's novel *The Catcher in the Rye*, which was first published in 1951. The main motivation as to why explore this text is the fact that it seems to be very promising: the language of the novel is particularly marked, as the author uses various means to imitate teenage speech of his time. By examining the text we may find and name some typical tendencies of teenage informal language in general and in addition, we may uncover some specific language habits which are characteristic only of the narrator's personal idiolect. Moreover, the text is more than sixty years old and this may make our analysis even more intriguing, as we can also observe possible changes in teenage vocabulary and comment on the extent to which the language features changed.

The method which will be used in this research is keyword analysis, which will be described in larger detail in Chapter 2.5. The main reason why this specific method was chosen is that it can reveal text-specific words in a very short amount of time. The software employed in this analysis, AntConc, can then be used in order to examine how keywords tend to behave, which words they attract, if they have positive or negative connotations and so on. Probably the biggest advantage of this method is that it can help us see recurrent language patterns which would be harder to notice only by intuitive reading.

More advantages, but also disadvantages, of this kind of approach are noted in the following chapter, along with definitions of the basic terms which we will be working with. The methodology then describes the software used for the extraction of the data and it will also include the parameters of the target and reference corpora. In the analytical part, the frequent and key words (proper names, grammatical and lexical words) will be examined.

2 Theoretical Background

2.1 Electronic Text Analysis

Corpus linguistics is an increasingly popular discipline which deals with the analysis of naturally occurring language on the basis of computerized corpora with the help of the computer[1]. The discipline has been expanding rapidly over the past few years, perhaps due to the development of information technologies and due to a growing interest of linguists in this type of approach (Adolphs, 2006: 1).

A corpus, which will serve as the basis for the analysis of the thesis, could be defined as a "collection of texts which has been put together for linguistic research with the aim of making statements about a particular language variety" (Biber et al., 1998: 4). In other words, corpus data are commonly used for language description, which is a process which aims to develop a deeper understanding of language in use[2]. One of the biggest advantages of corpus linguistics is that it is capable of revealing recurrent patterns in language use which "lie outside unaided human perception" and which "no amount of introspection or manual analysis could discover" (Stubbs., 2007: 131).

Nowadays, there is a rather large number of disciplines which employ some observations gathered through electronic text analysis. The group includes disciplines such as ELT, forensic linguistics, studies of language variation, sociolinguistics and most importantly to our purposes, corpus stylistics (Adolphs, 2006: 11).

2.2 Electronic Analysis of Literary Texts

2.2.1 Stylistics and Corpus Stylistics

Traditional stylistics, or "linguistic study of style" (Leech et al., 1981: 11) has been used in a great number of studies which focus at how particular aesthetic effects are achieved through language. Stylistics combines two different approaches: study of language on the one hand, and

[1] Cf. Nesselhauf, Nadja. *Corpus Linguistics: A Practical Introduction*. Available at http://www.as.uni-heidelberg.de/personen/Nesselhauf/files/Corpus%20Linguistics%20Practical%20Introduction.pdf (Accessed 9 April 2017).

[2] Sinclair distinguishes between two functions of electronic text analysis: language description and language application (Sinclair 2004 cited in Adolphs 2006: 2). Unlike language description, language application aims to achieve results which are relevant also in the non-linguistic community (an example of language application could be a production of a translating machine or a spell checker). In this thesis, only the notion of language description is therefore relevant.

study of literature on the other. The position of stylistics is therefore quite vulnerable as it may come under attack from both sides: literary critics may find it too systematic, while linguists may find stylistic analyses not systematic enough, as they incorporate too much interpretation (Mahlberg, 2007: 220). On the other hand, Fisher-Starcke (2010: 1) sees that the two disciplines complement each other and stylistics therefore has a great potential. In fact, stylistic analyses of literary texts are quite common and although poetry has been described from a linguistic point of view much more frequently, there have also been numerous studies which deal with literary fiction (Leech et al., 1981: x). As an example we could mention a study by Halliday (1971) who analyzed the occurrence of transitive and intransitive verbs in Golding's *The Inheritors*.

Another statistical approach to the analysis of literary style aims at identifying and describing authorial style and authorship. This approach, in fact, has also quite a rich tradition and Holmes (1989) even suggests that the beginnings of statistical stylistics (also called stylometry) could reach as far back as to 1850's, when Augustus de Morgan suggested that word length may be an indicator of authorship (Holmes, 1989: 112). The statistical methods which are used nowadays are of course more elaborate, as they work with numerical probabilities and focus on highly frequent words, range of vocabulary, sentence length or frequency of certain conjunctions (Leech et al., 1981: 12).

However, a number of studies started to combine pure statistical data with methods which have been developed in the area of corpus linguistics, such as analyses of multi-word sequences or typical collocations of words, in order to analyze and interpret a work of literature. This method which combines corpus linguistic analytic techniques with literary stylistic analysis is called corpus stylistics. It employs descriptive tools to identify repeated patterns and tendencies in language, but it still leaves room for individual qualities of the given text and thereby links it with literary interpretation (Mahlberg, 2007: 219). Corpus stylistics commonly serves as a complementary approach which is used alongside more traditional techniques of interpretation, but it can also function as an independent approach to text analysis (Adolphs, 2006: 64).

2.2.2 Recent Studies

In vast majority of cases, our approach to literary work is, to some extent, influenced by already existing interpretations or previous discussions of the given text. Employing corpus stylistics in these cases can be very useful, as the analysis may generate new insights on the work, or provide evidence for themes already identified by literary critics (Adolphs, 2006: 65).

An example of such an approach could be a study carried by Stubbs (2005) where he focused at Joseph Conrad's novella *Heart of Darkness*. In the study, Stubbs proved that language data, like frequencies and recurrent phraseology, can indeed provide more detailed basis for widely accepted interpretations, but at the same time it can also identify significant linguistic features which literary critics have failed to notice. He works with the theme of vagueness, which has been already discussed in literary circles and he relates it to linguistic features of the text. Stubbs admits that critics have recognized few content words which contribute to the lack of clarity, such as *vague*, *indistinct* or *fog* (Stubbs, 2005:10). However, he notes that grammatical words denoting uncertainty are very frequent as well, even though they have not been noticed or stressed in previous discussions. These are usually "*some-*" compounds (*something*, *somehow* etc.) and expressions like *kind of*, *sort of* and *like* (as a preposition). The significance of these expressions is then confirmed by the fact that their frequency is higher in the novella than in both the written part of the BNC and in a corpus of literary fiction (ibid).

Another interesting corpus-stylistic analysis has been carried out by Fischer-Starcke (2009) who worked with keywords and frequent phrases of Jane Austen's *Pride and Prejudice*. In the article, Fischer-Starke examined how recurrent language patterns shape textual meanings and similarly to Stubbs, she was able to uncover some meanings which were not discussed by literary critics.

Electronic text analysis can also manifest in what ways the language of a certain character contributes to their portrayal. Culpeper (2009) examined the speech of six main protagonists of *Romeo and Juliet* and his conclusions provided deeper understanding of these characters and of their depiction. For example, he discovered that Juliet's keywords (*if*, *yet*, *or*, *would*,...) do not tend to express facts but rather wishes or possibilities, which Culpeper sees as an evidence of the anxieties she experiences in the play (Culpeper, 2009: 25).

Apart from studying only individual words, we can also look at the so-called 'clusters', or "repeated sequences of words" (Mahlberg, 2007b: 1). An example of such an approach could be a study performed by Michaela Mahlberg, who looked at key clusters in Dickens' novels. Unlike conventional study of clusters, which aims to make generalizations about language use, Mahlberg's method focuses on clusters which are specific to individual texts. She argues that clusters can be interpreted as pointers to local textual functions and that they can be employed as a useful tool for literary analysis (ibid).

2.2.3 Intra-textual Analysis

In relation to electronic exploration of literary texts, we can distinguish between two main approaches, depending on whether they rely on intra-textual or on inter-textual analysis (Adolphs, 2006: 65).

As indicated by its name, “intra-textual analysis” examines only one particular text or a text collection. There are various ways how we can approach the text. It is possible to perform an analysis which draws on themes which have been already identified by literary critics, as was for example the case of Stubbs's study described above. On the other hand, the analysis does not have to rely on the researcher's previous knowledge of the text, and a starting point for further analysis can be made by generating frequency lists or by examining collocates of individual words.

2.2.4 Inter-textual Analysis

Inter-textuality in its broad sense means simply that there are links or relations between various texts, which are sometimes conscious, i.e. quotations or cases of plagiarism, but it also includes other instances which are much less noticeable. The whole concept of inter-textuality is very subjective and the extent to which texts allude to the previous works is to great extent arguable – (Teubert, 2007: 78). In this sense, electronic text analysis may be helpful in that we may search for specific words and phrases in one text and compare their occurrence with that in another text in order to quickly reveal some links.

However, Adolphs (2006: 66) defines the concept of intertextual analysis as the “comparison of individual lexical items and phrases in literary texts with those that occur in other corpora with the aim of analyzing deviations and their status as literary effects”. In this context, a reference corpus is needed, as it serves as a norm to which the text is compared to.

2.2.5 Advantages of Electronic Text Analysis

Traditional language research tends to use native speaker intuition as the basis for linguistic theories. However, such an approach could introduce a high degree of bias and the conclusions of such studies may have been achieved entirely by introspective judgment of the particular researcher. Moreover, the intuition of a native speaker may be an unreliable source for making judgments about language in use (Adolphs, 2006: 7).

In this case, electronic text analysis could be more useful, as it works with naturally occurring discourse and the results can be easily verified, since other researchers may replicate the steps. In addition, there are certain aspects of language (i. e. word frequency, co-occurrence of words etc.) which are not open to intuitive inspection, but are easily recognized by a concordance software and it may therefore provide new and surprising facts about language use. Another advantage of employing software packages is that they allow us to manipulate language data in different ways to suit a specific research purpose. In this way, it is possible to work with a large body of text from which we can generate exact empirical data in very short amount of time (ibid.).

Probably the biggest advantage of this kind of approach, especially in relation to literary analysis, is that the collected data may reveal recurrent patterns in language which would be hard to detect and describe by intuitive analysis. These include typical phrases and clusters, but the analysis can also show if a word carries positive or negative connotations and it can reveal which semantic concepts surround individual words (ibid.: 8). Moreover, exploring a work of literature from a linguistic point of view can provide new insights or perspectives on the text or provide factual evidence for already existing interpretations.

2.2.6 Possible Problems and Limitations

In comparison with more traditional approaches, electronic text analysis does rely much less on intuitive interpretation; yet it would be inaccurate to claim that intuitive aspect is not present at all. This is visible especially at the beginning of a study when a researcher decides what aspect of language he or she will explore and which queries they will address. Because of the presence of these prior subjective decisions, quantitative methods of text analysis have found some objectors, as they find the approach to be very selective.

The criticism of corpus stylistics seems to refer to issues which concern stylistics in general: Stubbs (2005) mentions, for example, some critical remarks noted by Stanley Fish (1996)[3], who claims that stylistics depends heavily on selective attention to data and that researchers either "select a few linguistic features, which [they] know how to describe, and ignore the rest" or that they will "select features which [they] already know are important, describe them, and then claim they are important" (Stubbs, 2005: 6). Nevertheless, this effect can be minimized if the

[3] Fish, S. E. (1996) 'What is Stylistics and Why are They Saying Such Terrible Things About It?', in J. J. Weber (ed.) *The Stylistics Reader*, pp. 94–116. London: Arnold.

researcher works only with frequency lists and keywords, because those are objective features which makes the subjective selection of data almost impossible (Fischer-Starcke, 2009: 4). Similarly, Stubbs stresses that the observer "must not influence what is observed" and that data and analysis have to be independent (Stubbs, 2007: 130).

At the end of the study, however, the conclusions are always the work of a linguist who interprets his or her results, while another linguist could interpret the data differently. However, even though these decisions involve subjective interpretation, they are based on observable and replicable data (ibid.: 170).

At the same time, it is important to realize that a corpus linguist is always restricted to features which the software can find (Stubbs, 2005: 6). For example, corpora which are tagged for parts of speech (POS) are often easier to work with, because we can study the sentence structure and look at the occurrence of specific parts of speech. It also distinguishes between words which have gone through conversion, like *use*, which can stand for both a noun and a verb. However, not all corpora are POS-tagged and the work with such a corpus is a little limited, at least in comparison with a tagged one. On the other hand, some homographs, i.e. words which have the same spelling but different meanings, may belong to the same word class, as is the case with *letter* or *bank* and therefore the semantic difference would not be detected even by a POS tagged corpus. For these reasons, a manual examination of similar data is required.

2.3 Basic Information about a Text

There is currently a number of software packages which allow the analysis of language data. Apart from performing more complicated tasks, such as extracting multi-word units or generating key words, which shall be discussed later on, it can also provide some very basic information about the text. This information includes average sentence length, word length or word count (Adolphs, 2006: 39). Some of the information is expressed in terms of ratios, out of which the most common one is the so-called "type-token" ratio. "Token" refers to the total number of running words in the text, whereas "type" refers to the number of different words (ibid.). This ratio therefore provides a basic understanding of the lexical variation, as texts with low type-token ratios are likely to be more complex. However, if we wish to make elaborate statements of the given text, more tools have to be employed, but these observations may be used as a good starting point for further analysis (ibid.).

Another common function of these software packages is the possibility to display the data in different ways, one of which is the KWIC concordance (Key Word In Context) (Stubbs, 2007: 129). The KWIC concordance is very convenient, especially because it can find all occurrences of the word in a large corpus and show its right and left contexts, which allows the human analyst to see recurring patterns of language (ibid.).

2.4 Frequency Lists

Frequency as such is extremely important in the area of corpus linguistics. It has been argued that there is a direct relationship between frequency of a linguistic feature and its significance in the corpus, in other words, items are frequent precisely because they are typical of the particular text (Fischer-Starcke, 2010: 15). At the same time, it is crucial for stylistic analysis as well, as frequency is an indicator of typicality of language use, whereas style is the typical language of the text, which makes frequent items particularly relevant when discussing literary style (ibid.: 16).

The extraction of frequency information of individual words is a very quick process, which is carried out solely by the software without human interference and is consequently very objective and unbiased. The most frequent items usually consist of grammatical words, such as determiners, prepositions, personal pronouns or auxiliaries, as they are very common in the English language (Adolphs, 2006: 41). In order to establish what items are truly typical of the given corpus (and not only of English in general), comparison of two corpora may be very helpful. This comparison enables us to identify keywords, i.e. words, which are "unusually frequent (or infrequent) in a text compared to the reference corpus" (Mahlberg, 2007: 223).

2.5 Keywords

The term 'keyword' could be defined as a "word which is statistically characteristic of a text" (Culpeper: 2009, 30) or as a "statistically relevant lexical item" (ibid.: 32). Unlike frequency lists, which are based on "absolute frequency", keywords are based on "relative frequency", which means that they are generated on the basis of the comparison of frequency information in two corpora (ibid.). In order to find keywords in a text, we have to compare the frequency list derived from the target corpus to the one derived from a larger reference corpus (Adolphs, 2006: 44). The data which appear with significantly higher frequency in one corpus when compared to the reference corpus are then “positive keywords”, whereas items which appear with significantly lower frequency are called “negative keywords” (ibid.).

To generate a list of keywords, we have to use the Keyword Tool of the concordance software which we are working with. Usually, there are two statistical methods which can be used for the calculation: a chi-square test of significance or Ted Dunning's Log-Likelihood analysis which can give a better estimate of keyness, especially when contrasting longer texts (Culpeper, 2009: 33).

2.5.1 Proper Nouns

Scott and Tribble (2006) (also Culpeper, 2009 and Mahlberg, 2007) distinguish between three types of keywords: proper nouns, content words and function words.

Proper nouns tend to appear on keyword lists more frequently than other parts of speech and in fact, nouns in general can make up to 70% of the keyword types (Scott and Tribble, 2006: 72). Proper nouns are this common perhaps because the names of the characters or places consist a great part of the fictional universe and repetitions are therefore quite natural.

Nevertheless, Fischer-Starcke (2010: 95) believes that proper nouns are not particularly relevant for identifying dominant topics of the text or for analysis of its structural organization. This is because proper nouns are “necessarily identified as keywords, because it is unlikely that names occur with equal frequencies in two sets of data”. Similarity, Scott dismisses proper nouns as mostly unimportant and adds that some proper nouns may be occasionally identified as key, even if they do not relate to the themes of the text: “a text about racing could wrongly identify as key, names of horses which are quite incidental to the story” (Scott, 1998: 71).

2.5.2 Lexical Words

By lexical (or content) words we generally understand items which belong to open word classes, such as nouns, adjectives and lexical verbs (Stubbs, 2007: 191). More precisely, content words could be defined as words referring to “a thing, quality, state or action” or as words which “have meaning when used alone” (Scott and Tribble, 2006: 96).

These are keywords which “human beings would recognize” and which give a good indication of the text's content, or 'aboutness' (Scott, 1998: 71). Culpeper exemplifies this by looking at Romeo's content keywords like *beauty* and *love*, which, indeed, could be intuitively recognized as important to the play by most readers (Culpeper, 2009: 38).

2.5.3 Grammatical Words

The category of grammatical (function) keywords consists of closed word classes like prepositions, conjunctions or auxiliaries. Since most of the frequent words in English are grammatical ones (Stubbs, 2007: 181), it is always important to find what the cause of their unusual frequency in the target corpus is, ideally by investigating the individual concordance lines (Scott, 1998: 71).

Unlike lexical words, grammatical words are usually not identified by the reader as key and while lexical keywords generally reveal patterns of 'aboutness', grammatical keywords reveal stylistic features of the text (ibid.).

Still, these keywords are important for literary analysis. For example, Culpeper (2009) looks at Juliet's function keywords like *if*, *would* or *be*, which most people are "unlikely to predict", yet which are relevant to the plot, because they reveal her tendency to use subjunctive mood and conditional clauses more frequently than others. This tendency could be explained by the fact that Juliet is "in a state of anxiety for much of the play" (Culpeper, 2009: 38-39). For this reason, it does not seem wise to exclude function words from stylistic analysis, as it would result only in a partial picture.

2.6 Study of Multi-Word Expressions

As has been mentioned before, one of the main aims of corpus linguistics is to look at recurrent patterns of language use. The frequent occurrence of lexical or grammatical patterns in a text collection is an evidence of what is typical in the language. For this reason, it is very useful to examine words' behavior and contexts, instead of only inspecting the frequencies of individual items. This can give us a good idea of the words' connotations and it enables us to see how they tend to co-occur with other items.

2.6.1 Collocation

Both corpus linguistics and literary stylistics are interested in the relationship between meaning and form (Mahlberg, 2007: 221). Mahlberg (2007) believes that a central descriptive category to characterize the association between meaning and form is the concept of collocation which describes the tendency of words to co-occur (ibid.: 222).

In this thesis, the term collocation is understood as “the relationship that a lexical item has with items that appear with greater than random probability in its textual context” (Hoey, 1991: 6). Lexical items which are involved in a collocation are always to some degree “mutually predictable” (Crystal, 2003: 162) and collocation occurs when one item 'calls up' another one in the mind of a native speaker: every mature native speaker will say *commit a murder*, but they cannot say something like *commit a task* (ibid.).

By a collocation we do not understand idiomatic expressions, because their structure allows little or no change and their meanings often cannot be predicted from the individual words. For these reasons, it may not be wise to analyze them as collocations but we should rather treat them as an individual group (ibid.). Similarly, we do not define as collocations chunks of texts like *it seems to me* or *lived happily ever after*, because they also behave differently and require a separate analytic approach (and will be discussed later on) (ibid.: 163).

Lipka (1992) also points out that a collocation is in many ways neutral. It is a combination of lexemes which is “independent of word class or syntactic structure” (Lipka, 1992: 166): if we take words like *open* and *window*, they will form a collocation, irrespective of whether *open* is a verb or an adjective (ibid.). Similarly, collocates do not have to be contiguous, as demonstrated by this example: '*They collect stamps'* and '*They collect many things, but chiefly stamps*' (ibid.). In fact, most software programs enable us to set the span in which we want to look for collocates of a word and we may therefore find collocates which are farther in the text from the given word.

When applying the study of collocations to a literary analysis, there are various aspects which we can focus on. We can, for example, pick a frequent or key word and then look at its typical context. More specifically, we could focus on a specific character by examining the collocates which surround the character's name, out of which adjectives and verbs tend to be the most telling ones. This can enable us to see how a character is presented in the text or what actions and features are associated with them (Adolphs, 2006: 67).

A similar approach was adopted by Kettemann (1995)[4], only instead of searching for collocates of characters' names, he decided to find out how personal pronouns *he* and *she* collocate, in

[4] Kettemann , B. (1995) ‘Concordancing in stylistics teaching’, in W. Grosser, J. Hogg and K. Hubmayer (eds) *Style: Literary and Non-Literary. Contemporary Trends in Cultural Stylistics*, New York:The Edwin Mellen Press, pp. 307–18.

order to highlight the differences in characterization of men and women in an early, emancipatory American short story. What he found out was that at the beginning of these stories, the gender depiction was quite stereotypical: the pronoun *she* collocated mostly with housekeeping verbs like *cooking* or *baking*, whereas *he* collocated with various lexical verbs, suggesting that men appeared to be more in control (Kettemann, 1995 cited in Adolphs, 2006: 68). However, this type of analysis tends to support only very basic interpretations and further analysis of the surrounding text is required.

2.6.2 Colligation, Semantic Preference, Semantic Prosody

Apart from collocation, there are other recurrent phrasal constructions which are recognized. Stubbs (2007) mentions Sinclair's model of extended lexical units, which apart from collocation include concepts of colligation, semantic preference and semantic prosody (Stubbs 2007: 178).

Colligation is the co-occurence of grammatical choices and is therefore closely connected to syntax (unlike collocation, which is the co-occurrence of individual word forms and thus concerns lexis). Semantic preference, also referred to as 'lexical field', is the relation between a word and lexical sets of semantically related word forms. Finally, semantic prosody is the discourse function of the unit: it describes the speaker's evaluative attitude or communicative purpose (ibid).

The study of semantic prosodies allows us to recognize 'shades' of modality. This recognition is typically not easily detected by intuition, rather it is supported by empirical corpus evidence (Adolphs, 2006: 71). The shading of a lexical item can be determined by finding its collocates: Adolphs (2006) looks, for example, at the word *happen* and by studying the concordance lines, she is able to see that the modality shading is negative, as the word mostly occurs in combinations like *the worst that can happen* (ibid: 72).

Semantic prosodies may be used in the analysis of point of view in fiction, as they contribute to the ways in which the characters' speech and thoughts are presented. As an example we could mention an analysis by Adolphs who focused at an extract from Woolf's *To the Lighthouse*. The analysis brought out the differences in the 'shading' of the text, which leaned towards uncertainty and negativity in the depiction of Mrs Ramsay and towards strength and certainty in the case of Mr Ramsay (ibid: 76).

2.6.3 Clusters

A cluster[5] can be defined as a "recurrent uninterrupted string of orthographic word forms" (Stubbs, 2007:166) or as "frequently occurring word sequences which follow each other more frequently than expected by chance" (Hyland, 2008: 5). In addition, they tend to vary across genres and disciplines and thus help to shape textual meanings (ibid). Clusters are especially relevant for corpus research, because they are identified purely on the basis of their frequency in the text and unlike idioms, they are semantically transparent (ibid: 6).

There are two basic ways how we can approach frequent clusters in electronic text analysis. On the one hand, we may select individual items which we are interested in (such as keywords) and study their typical co-text. On the other, we could also generate the most frequent clusters and then study their typical content. If we do it like that, then we can investigate if the high frequency of these words is conditioned by them forming recurrent phrasal constructions which have frequent and predictable functions in text (Stubbs, 2007: 166).

Apart from examining frequent clusters, as described above, we may also study key clusters. Key clusters are calculated similarly as keywords, only instead of comparing two lists of frequent words the software will compare the lists of frequent clusters. This means that key clusters are capable of revealing what is typical of the given text and are much less generally applicable. Similarly, key clusters are not likely to include frequent sequences of grammatical words like *if it were a* or *as if he had been* (Mahlberg, 2007b: 9). However, the main limitation is that most programmes with free access do not enable the calculation of key clusters. Therefore, the first approach (i.e. examining clusters of specific words) will be used in the thesis, yet the main focus should not fall on grammatical clusters, because they are likely to appear in number of texts as explained.

2.7 Linguistic Studies of *The Catcher in the Rye*

Since the time of its publication, J. D. Salinger's novel *The Catcher in the Rye* (1951) has been subject of both praise and criticism (Graham, 2007: xi). The criticism stemmed from the fact that the readers of that time, as well as reviewers, would find the language of the novel offensive or obscene and they would condemn the idea that the book's "hero" should be a teenager who

[5] Other common terms for such sequences are 'lexical bundles' (Biber, Leech), 'n-grams', 'chains' or 'chunks'

drinks, smokes, and engages with a prostitute (ibid). The first official complaint against the novel has been raised in 1955 and most of its controversy usually relates to the question if such a novel should be read and studied in literature classes (ibid: 17).

As was mentioned already, it is the language of the novel that the controversy is due to. One of the complaints says that the novel "takes the Lord's name in vain 295 times" and "uses blatant blasphemy 587 times" (Laser and Fruman, 1963: 127). The contents of the novel are recognized as unacceptable mostly by the parents of the students, perhaps afraid that their children would sympathize with the main protagonist and that they might justify or even copy his behavior, that is, drinking, smoking and failing classes. Shortly after the book was published, the situation was so tense that teachers would even lose their jobs as a penalty for assigning this text to their students (Graham, 2007: 18). All of this may be partly justified by the fact that the novel was published after the World War II, when America was a very conservative country.

What is probably more interesting however, is that the nature of the book remains problematic to this day, as it continues to be withdrawn from high school reading lists on the basis of its "sexual content and offensive language" (ibid.: xii). But in spite of this controversy, the book is nowadays seen as one of the most famous American novels of the twentieth century and it continues to attract generations of readers, which suggests that even after six decades, there are aspects which are still relevant for young readers.

Holden Caulfield, the narrator, is without doubt the most prominent character in the novel, whose authenticity rests greatly on his very distinctive voice. Moreover, this voice is usually seen as the main source of the novel's humor, which is the key aspect of its massive success (ibid.: 39). A number of reviewers praised Salinger's style of writing and how it conveys the comic element. Graham cites R.D.Charques who notes that the style is "a little showy" but "intelligent, humorous, acute and sympathetic" (ibid.). Other reviewers, like Harvey Breit, were afraid that this distinctive writing style is so dominant that it may distract the reader from important issues which the text raises (ibid.).

When focusing at Holden's narration, Graham stresses that Holden's language evokes intimacy and informality of speech rather than writing (Graham, 2007: 6). At the same time, the narrator tends to leave certain things unsaid and ambiguous, and the use of such technique gives his readers interpretative freedom, which is otherwise quite difficult to achieve in first-person narrative. This can be observed in the scene when Holden is unsure if his teacher molested him or not: “[Mr. Antolini] is sort of petting me or patting me on the goddam head” – Holden is not certain which word choice is appropriate, as “petting” has more sexual connotations, while the word “patting” is preferred in friendly situations (ibid.). Just like Holden, the reader is uncertain and therefore is forced to decide on his own as to what is in fact going on.

Another technique used quite commonly by Salinger goes against the use of ambiguity discussed above. On many occasions, Holden's speech is very repetitive and he tends to explain something which does not need explanation: “Lift up, willya? You’re on my towel,” Stradlater said. I was sitting on his stupid towel’. Apart from describing a situation in more detail than required, Holden also tends to repeat and stress the words which relate to his feelings. Graham, even though she focuses mostly at the plot, cannot miss that “words related to 'worry' appear six times in the first paragraph of Chapter 6” (Graham, 2007: 22). This, in fact, proves that in this case, the protagonist's speech is so distinct that some patterns may be recognized easily even by intuitive reading.

Studying the language and style of *The Catcher in the Rye* may be, however, justified not only on the basis of literary interest, but also on the basis of linguistic significance. In fact, Costello (1959) even suggests that the novel might be potentially, in coming decades, studied from a sociological point of view. He claims that the text may function as a typical representative of teenage vernacular of the 1950s and suggests that the text may be approached in a similar fashion as *The Adventures of Huckleberry Finn*, which is nowadays, among other things, seen as a valuable study of 1884 dialect (Costello, 1959: 172). In fact, already in 1958, Gwynn and Blotner (1958) claimed that “it is not inconceivable that some day Holden Caulfield may be as well known an American boy as Huck Finn” (Gwynn and Blotner, 1958: 29). This claim about *Catcher*'s significance may be justified also because most critics who reviewed the book in the time of its publication indeed considered the language to be authentic (Costello, 1959: 172). Costello, nevertheless, continues to stress that Salinger's task was an artistic one and that his goal was to create an individual character, rather than reproducing teenage speech in general. He achieved this task by giving Holden “typical teenage speech” which is “overlaid with strong personal idiosyncrasies” (ibid.: 173).

Costello focuses on these personal idiosyncrasies and observes that Holden tends to end utterances with phrases like “and all”, “or something” or “or anything” and goes on to show that they do not have a consistent linguistic function and that their use is often arbitrary. He also notes that the second most common idiosyncrasy is affirmation like “really” or phrases like “it really does” or “if you want to know the truth” (ibid.:173-5).

Costello also addresses the controversy which surrounds the language of the main character, which many would call offensive or obscene. It seems however, that Holden is actually quite careful about his language and “does not use vulgarity in a self-conscious way” (ibid.: 175). For example, the word 'fuck' is not even once used as a part of Holden's speech and it appears in the novel only when Holden disapprovingly discusses its wide appearance on the walls. However, Holden does use expressions like 'sonuvabitch' or 'bastard', typically when he refers to 'phonies', and his language truly may be seen as blasphemous, as his “favorite” words are 'goddam' and 'hell' (ibid.). Using slang is also typical of Holden's speech, and the meaning of slang expressions is usually not stable: to be “killed” by something can be both good ('That story just about killed me') and bad ('Then she turned her back on me again. It nearly killed me.') (ibid.: 177).

Another crucial feature of Holden's language is that it combines these colloquial expressions with advanced vocabulary, which often results in comic effect. This advanced vocabulary reflects the fact that he is well-read and educated, perhaps even “overtaught”, which results in the use of hyper-correct forms like “She'd give Allie or I a push.” (Costello, 1959: 180). Nevertheless, these structures are seen as another proof that Holden's speech is supposed to imitate a spoken language, rather than written. Costello claims: “I doubt if a student who is 'good in English' would ever create such a sentence in writing” and similarly, he finds it “impossible to imagine Holden taking pen in hand and actually writing 'Spencer'd' or 'I'd've'” (ibid.). All of this confirms that Holden's narration is supposed to be authentic artistic rendering of informal, colloquial, teenage American spoken language.

3 Material and Method

3.1 Material

This thesis is going to analyze frequent and key words of J. D. Salinger's novel *The Catcher in the Rye* and the text of the novel is therefore going to function as the target corpus. The reference corpus will allow us to generate the keywords of the target corpus. At the same time, the reference corpus ought to be larger than the target corpus (Adolphs, 2006: 44) and it is going to function as a language norm in the specific context (ibid, 66). For these reasons, I decided to compose the reference corpus out of five contemporary novels which share some fundamental similarities with Salinger's novel. The reference corpus consists of the following texts: *Rats Saw God* (1996) by Rom Thomas, *It's Kind of a Funny Story* (2006) by Ned Vizzini *Someday This Pain Will Be Useful to You* (2007) by Peter Cameron, *The Absolutely True Diary of a Part-Time Indian* (2007) by Sherman Alexie and *Life of a Loser Wanted* (2014) by Lou Zuhr.

Both the target and the reference corpora include novels which: are written with informal American English, have a strong presence of male, teenage narrator, are predominantly monological in their nature and the audience of these books consists mostly of teenage or young adult readers. However, the crucial difference between the two corpora is the fact that Salinger's novel was first published in 1951 and the target corpus could consequently be seen as a representation of the colloquial teenage speech of the 1950s. On the other hand, the reference corpus consists of books which were written relatively recently and as a result, they may illustrate how contemporary teenage American language looks like.

Corpus	**Types**	**Tokens**
Target corpus (*Catcher in the Rye*)	**3989**	**77574**
Reference corpus	**15071**	**272638**
- *Rats Saw God*	8583	61164
- *It's Kind of a Funny Story*	5977	88091
- *Someday This Pain Will Be Useful to You*	5645	60845
- *The Absolutely True Diary of a Part-Time Indian*	4335	49327
- *Life of a Loser Wanted*	2720	13211

Table 1: Corpora used in the thesis

3.2 Method

The research method used in this thesis is a keyword analysis. In order to generate keywords, i.e. words which are statistically relevant, a software needs to compare the expected frequencies of word lists in the target and the reference corpora (more discussion on the definition of keywords and keyword analysis in Chapter 2.5).

The software which will be used for this calculation in the thesis is AntConc 3.4.4, which was developed by Laurence Anthony and is available for free download. The main tools which will be used during the research is the word list tool, which displays the most frequent words, and the keyword list tool, which generates keywords. The method chosen for the keyword generation is log-likelihood calculation. Another measure of statistical significance which identifies keywords is a chi-square test, however, Adolphs (2006) mentions that this type of calculation "can produce distorted results if the expected frequencies of individual items are low" (Adolphs, 2006: 50), so log-likelihood analysis was preferred.

In the analysis, top hundred resultant keywords (sorted by the log-likelihood value) will be selected and sorted into three basic groups: proper names, grammatical keywords and lexical keywords. We are then going to study the behavior of the chosen words, focusing especially at their tendency to co-occur with different words or to form text-specific clusters. This can be done either by using the collocates tool, which enables us to set the span in which we want to search for a collocate of the given word, or by the clusters tool, where we can set the cluster size and chose if we want to search clusters on the right or left side from the given word.

The setting of the software was selected to facilitate the work with the resultant keywords. For our purposes, token definition includes only letters, but no punctuation. As a result, contracted forms like *I'd* will not appear on the word list; instead, they will be split into *I* and *d*, which can give us more accurate results. The software is also set to treat all data as lower case, as otherwise the data would be much harder to sort, as we would have different frequencies for the same word when spelled with an upper case letter, e.g. at the beginning of a sentence.

4 Analysis

4.1 Frequent words

This part of the thesis presents an analysis of the most frequent words of J. D. Salinger's novel. As was already suggested in the previous section (Chapter 2.7), the language used in *The Catcher in the Rye* has multiple functions. It is, on the one hand, intended to shape specific personal speech of the main protagonist and those personal idiosyncrasies will be apparent mostly from analyzing keywords. On the other hand, Salinger's writing style is supposed to imitate the spoken language of American teenagers in general and this tendency may be in fact apparent from studying the most frequent words.

rank	**freq.**	**word**
1	4219	i
2	2629	the
3	2082	and
4	1723	to
5	1714	a
6	1597	was
7	1578	it
8	1391	you
9	1333	t
10	1298	he
11	1034	in
12	1028	of
13	987	all
14	980	she
15	956	that
16	820	s
17	726	me
18	705	said
19	605	her
20	600	my

Table 2: List of 20 most frequent words in *Catcher*

From only briefly looking at the results, we can see that majority of these words are grammatical ones, which was quite predictable, as these words are very common in the English language. We can also see that most of these words could be associated more with the spoken language rather than with written: by far the most frequent word is *I*, but the first person singular is also included in objective *me* and possessive *my* which all appear on the list, confirming its mostly subjective nature. At the same time, personal pronouns show that the narrator addresses another character or the reader (*you*), while pronouns like *he* or *she* show that the narrator reports on other characters. This kind of interpersonal interaction is to be expected when dealing with spoken discourse[6]. The spontaneity of conversation taking place in real time also results in the use of reduction effort-saving devices, such as contractions: "reduced enclitic forms of the verb (e.g. *it's*, *we'll*) and of the negative particle (e.g. *isn't*, *can't*)" (Biber et al. 1999: 1048). The contracted forms (*'s*, *'t*) are also attested on the *Catcher* frequency list. Finally, the conjunction *and* is again tied with spoken language, as it suggests that there is a great deal of coordination rather than subordination. This may be seen as a manifestation of the 'add-on' strategy (Biber et al., 1999: 1068, 1078) due to limited planning in real conversation (ex. 1).

1. I read a lot of classical books, like The Return of the Native **and all**, **and** I like them, **and** I read a lot of war books and mysteries **and all**, but they don't knock me out too much.

The hypothesis that the language of the novel is similar to speaking can be further supported when looking at a different corpus, which consists of spoken language data. In order to make such a comparison possible, we generated a frequency list from the spoken American English corpus of Santa Barbara [7].

rank	**freq.**	**word**
1	9073	i
2	7928	the
3	7145	and
4	6475	you
5	6051	s
6	5762	it
7	5573	that
8	5032	a
9	4609	to
10	3383	t
11	3064	of
12	2946	he
13	2757	in
14	2722	they

[6] Biber et al. (1999: 1042) comment on the frequency and functions of personal pronouns in conversation: "The user of personal pronouns (by far the most common class of pronouns) normally assumes that we share knowledge of the intended reference of *you, she, it,* etc. This sharing of situational knowledge is most obvious in the case of first and second person pronouns (especially *I* and *you*) which, referring directly to participants in the conversation, are the most common in this variety."

[7] Available at http://www.linguistics.ucsb.edu/research/santa-barbara-corpus, June 29th 2017.

15	2689	was
16	2415	know
17	2337	yeah
18	2154	is
19	2140	like
20	2085	we

Table 3: List of 20 most frequent words of the Santa Barbara corpus of spoken English

When comparing Tables 2 and 3, it indeed becomes clear that they are very similar: in fact words *I*, *the* and *and* are among the top three word-forms on both of these lists and other words like *you*, *he* or *was* are represented in both tables as well. On the other hand, a closer look at the Santa Barbara corpus shows that *Catcher* does not display all the typical features of informal spoken language, e.g. repeats, false starts or filled pauses (cf. the use of the pronouns *I*, *me*, *my* in ex. 2 a. and b.).

> 2. a. If you really want to hear about it, the first thing you'll probably want to know is where **I** was born, and what **my** lousy childhood was like, and how **my** parents were occupied and all before they had **me** (*Catcher*)
>
> b. Roy: ... Yeah, **I don't know, I mean** -- **I**- **I** don't know if our drought here will ever break. **I** wonder if this is just isn't, (Santa Barbara, file SBC003)

What is perhaps quite interesting is that *you* is much more frequent in the Santa Barbara corpus which may be seen as an evidence of *Catcher* being more of monological nature, as the narrator tends to speak more about himself (hence all the first person pronouns) rather than frequently addressing others (hence *you* is not as frequent as one may expect it to be). The high frequency of *you* in the Santa Barbara corpus is also due to the addressee-oriented discourse marker *you know* (ex. 3). In the *Catcher* corpus, *you* is often used to address the reader (ex. 2 a.) or to refer to the general human agent (often preceded by *if* or *when*, ex. 4).

> 3. I don't know how to say it. But **you know**, they do it for a living. **you know**, ... most people that you would get to trim your horse do it .. all the time. (Santa Barbara, file SBC001)
>
> 4. I mean I could shoot the old bull to old Spencer and think about those ducks at the same time. It's funny. **You** don't have to think too hard when **you** talk to a teacher. (*Catcher*)

However, *Catcher* does have some features which are typical of almost all narratives: the high frequency of "said" suggests that there is a high number of reporting clauses. The most frequent immediate left collocate of *said* in the Catcher is the first person pronoun *I* (ex. 5), its frequency of 297 instances exceeding by far that of the pronouns *he* (125 instances) and *she* (120 instances).

5. “I know I did,” **I said**. **I said** it very fast because I wanted to stop him before he started reading that out loud. (*Catcher*)

4.2 Keyword Analysis

This section provides an analysis of keywords in J. D. Salinger’s novel. Keywords are generated by comparing word lists of *Catcher in the Rye* with the word list of the reference corpus which consists of five present-day novels written for teenage readership and with a dominant male teenage narrator.

The first hundred keywords will be then separated into three basic categories as classified by Scott and Tribble (2006) and described previously in Chapter 2.5: proper nouns, lexical words and grammatical words.

Rank	Keyness	Word
1	753.793	all
2	738.580	goddam
3	554.545	old
4	501.027	hell
5	393.976	d
6	356.517	he
7	346.680	phoebe
8	304.309	damn
9	289.403	stradlater
10	249.545	was
11	245.567	very
12	243.206	anyway
13	226.096	ackley
14	225.596	sort
15	204.960	though
16	188.036	boy
17	180.877	pencey
18	178.365	t
19	162.789	sally
20	153.745	antolini
21	150.731	jane
22	145.345	anything
23	144.771	didn
24	144.504	went
25	141.821	she
26	141.388	lousy
27	139.682	guy
28	139.452	while
29	139.005	sudden
30	136.676	kept
31	135.176	around
32	131.294	started

Rank	Keyness	Word
51	96.003	stuff
52	93.453	sore
53	92.802	him
54	90.704	kidding
55	88.917	hat
56	86.443	hardly
57	82.063	if
58	81.407	nice
59	80.544	b
60	79.135	finally
61	78.380	caulfield
62	76.002	pretty
63	75.415	practically
64	75.401	when
65	70.715	sake
66	67.010	funny
67	65.982	really
68	63.307	crumby
69	63.307	luce
70	61.642	over
71	61.287	something
72	61.059	guys
73	60.883	said
74	60.292	hunting
75	60.292	maurice
76	60.150	i
77	59.734	near
78	59.169	always
79	58.595	corny
80	58.464	listen
81	57.060	gave
82	54.633	certainly

33	127.953	somebody	83	54.263	sonuvabitch
34	126.614	holden	84	53.231	they
35	126.523	terrific	85	53.098	go
36	125.976	dough	86	52.926	crazy
37	125.097	even	87	52.722	whole
38	120.437	wouldn	88	52.665	told
39	119.080	it	89	51.533	or
40	117.191	till	90	51.515	about
41	116.902	bastard	91	51.248	helluva
42	116.160	mean	92	49.838	ernie
43	114.555	allie	93	49.616	kid
44	111.736	quite	94	49.242	coat
45	108.575	ya	95	48.234	madman
46	105.511	spencer	96	48.234	suitcases
47	103.773	got	97	48.234	whooton
48	103.643	too	98	46.929	gloves
49	96.873	phony	99	46.725	nobody
50	96.468	chrissake	100	45.219	horsing

Table 4: Top 100 Keyword of *Catcher* ranked by keyness (log-likelihood)

4.2.1 Analysis of Proper Nouns

As was explained earlier in Chapter 2.5.1, proper nouns are very likely to appear on a keyword list, as it is quite predictable that names of places and characters would be repeated to a great extent in the target corpus. At the same time, it is improbable that the same proper nouns would appear on the reference corpus as well and such occurrence would be purely coincidental. Generally, the main function of proper nouns is that they shape the fictional universe, as they introduce the main characters and places where the action takes place.

keyness	**freq.**	**word**
346,680	115	phoebe
289,403	96	stradlater
226,096	75	ackley
180,877	60	pencey
162,789	54	sally
153,745	51	antolini
150,731	50	jane
126,614	42	holden
118,870	38	d.b.[8]

[8] The acronym D.B. did not appear on the keyword list, as the token definition does not include punctuation. This name was only discovered when looking at the contraction *'d*. The search for the term D.B. showed that there are 38 occurrences of the name. The value of log-likelihood was then calculated at http://ucrel.lancs.ac.uk/llwizard.html, accessed July 17th 2017, by entering the following values: the frequency of the term is 38 in the target corpus, zero in the reference corpus. The size of the target corpus is 3989 tokens, the size of the reference corpus is 15071 tokens.

114,555	38	allie
105,511	35	spencer
78,380	26	caulfield
63,307	21	luce
60,292	20	maurice
49,838	19	ernie
48,234	16	whooton

Table 5: Proper nouns[9] within the top 100 keywords ranked by keyness (log-likelihood)

In the *Catcher* corpus, the proper nouns refer mostly to character names. The most frequent one is *Phoebe*, the narrator's sister. Her name typically occurs with the adjective *old*, in fact, out of the total 115 occurrences of the name *Phoebe*, it is modified by this adjective in 69 cases. When referring to Phoebe, *old* always marks the name as a term of endearment; the narrator uses it lovingly and kindly, as can be seen in ex. 6.

> 6. You'd like her. I mean if you tell **old Phoebe** something, she knows exactly what the hell you're talking about.

Another phrase expressing affection which is used repeatedly by Holden is *my kid sister Phoebe*, which is again informal[10] and therefore characteristic of the narrator's speech.

> 7. While I was changing my shirt, I damn near gave **my kid sister Phoebe** a buzz, though. I certainly felt like talking to her on the phone.

Finally, it should be noted that most of the time Holden ruminates on and reminisces about other people and as a result, he seldom uses vocatives and if he does, the utterance tends to be emotionally charged. For example, he addresses Phoebe directly only twice and in both cases it is as a part of a desperate exclamation (ex. 8a, b). This lack of vocatives is, however, mainly due to the monological nature of the novel: in most cases, the narrator retells his story and transcribes only those conversations which hold larger relevance to him.

> 8. a. **God, Phoebe**! I can't explain.
> b. **Oh, God, Phoebe**, don't ask me. I'm sick of everybody asking me that," I said.

When looking at the character names which appeared on the list, an interesting tendency can be observed. The narrator seems to prefer to call female characters by their first name, as is the case of *Phoebe*, *Sally* and *Jane*. Male characters are, on the other hand, often referred to by their

[9] In the table, proper nouns are written with lower case initial letters due to AntConc being set not to distinguish between upper and lower case letters. Therefore, the results will not take into consideration the upper case letters at the beginnings of sentences, making the results more precise. In addition, proper names are easily distinguised even if written with lower case letters.

[10] Kid - adjective, kid sister/brother (informal), a person's younger sister/brother: <http://www.oxfordlearnersdictionaries.com/definition/american_english/kid_3>, July 17th 2017.

last name. This would be predictable in the cases when he refers to adults (*Antolini*, *Spencer*), however, he also refers to his college and childhood friends by their last names: *Stradlater*, *Ackley* or *Luce*. In fact, the only male characters on the list which are described by their first names are *Holden* (typically in the direct speech of other characters) and his deceased brother *Allie*. In addition, there appears the name *Ernie*, however, after studying the concordance lines, it becomes clear that it mostly refers to the nightclub called *Ernie's*.

Apart from character names, there are some proper nouns which refer to places. In the table, there are two names of schools which the narrator attended: *Pencey* and *Whooton*. The most frequent prepositions which co-occur with *Pencey* are *at*, *to* and *out*, but they do not necessarily always function as space relators, as *Pencey* represents the institution rather than the actual school building. The differences in the usage of prepositions can be seen in the following examples.

> 9. a. (…), then took the bus back **to Pencey**.
> b. "Oh, do you go **to Pencey**?" she said.

The preposition *out* is used almost exclusively in relation to the narrator's termination of studies. When looking at the cluster *out of Pencey*, it is typically preceded by words which are informal (10a) and/or emotionally charged (10b)

> 10. a I said I'd **flunked out of Pencey**, though.
> b. All of a sudden, I decided what I'd really do, I'd **get the hell out of Pencey**—right that same night and all.

Also, it can be observed that the word forms of the verb 'leave' – in particular *left* and *leaving* - collocate with *Pencey*. This suggests that this event had a huge impact on the narrator, as he keeps returning to it throughout the novel. The frequent use of *since* or *when* (*I left Pencey*) implies that leaving the school is a fixed point in time after which everything changed and it is, in fact, the event which the narrator decided to start his narrative with:

> 11. Where I want to start telling is **the day I left Pencey Prep**.

Whooton is the name of school which Holden attended before Pencey. However, its usage is almost surprisingly monotonous: out of the 16 occurrences in total, a half form a phrase *when* sb *(I/we/you) was (were) at Whooton*. Other instances, also, carry a degree of nostalgia, as the name *Whooton* co-occurs with expressions of time like *once* or *used to*, usually followed by colloquial *this* with 'false definite function' (Dušková et al., 4.4).

> 12. **Once, at the Whooton School, this** other boy, Raymond Goldfarb, and I bought a pint of Scotch and drank it in the chapel one Saturday night, where nobody'd see us.

4.2.2 Analysis of Grammatical Words

Grammatical, or function words, have little meaning on their own and their main function is to show grammatical relationships in and between sentences (Scott and Tribble, 2006: 96). They consist of closed word classes such as prepositions, determiners, conjunctions and pronouns (ibid: 23). Grammatical keywords should be studied carefully, since they reveal stylistic features of the text but are, at the same time, especially easy to overlook in intuitive reading. This is due to their high frequency in the English language in general (further discussion on the topic in 2.4 and 2.5.3).

keyness	**freq.**	**word**
753,793	987	all
377,260	466	d[11]
356,517	1298	he
249,545	1597	was
243,206	149	anyway
204,960	191	though
178,365	1333	t
145,345	205	anything
144,771	400	didn
141,821	980	she
139,452	152	while
135,176	238	around
127,953	101	somebody
120,437	139	wouldn
119,080	1578	it
117,191	44	till
108,575	65	ya
103,773	257	got
92,802	360	him
82,063	438	if
75,401	395	when
61,642	224	over
61,287	220	something
60,150	4219	i
53,231	498	they
51,533	362	or
51,515	438	about
46,725	51	nobody

Table 6: Grammatical words within the top 100 keywords ranked by keyness (log-likelihood)

[11] The total number of occurrences for *d* is 506, however, it sometimes occurs as a part of a name (*J. D. Salinger*, *D.B.*) or as a school grade. The data in this table (that is keyness 377,26 and frequency 466) are only for 'd as a verb contraction (search term +*'d). The value of log-likelihood was again calculated at <http://ucrel.lancs.ac.uk/llwizard.html> by comparing the frequencies of 'd used as a verb contraction in the target corpus (466) with its frequency in the reference corpus (474), as well as comparing the sizes of both corpora (3989 tokens target corpus, 15071 reference corpus).

4.2.2.1 General Extenders

By far the most significant keyword in the entire corpus is the indefinite pronoun *all*, which occurs 987 times. Exploration of the concordance lines shows that the reason for its high frequency is Holden's tendency to end his utterances with the words *and all* (392 hits).

Expressions like *and all*, *or something* or *or anything,*[12] which all appear repeatedly at the end of utterances in the corpus, are very typical of spoken language and moreover, they are especially common with adolescent speakers (Stenström et al, 2002: 88). The high presence of general extenders in *Catcher* may be explained also by the informal character of the novel: Stenström stresses that vague expressions are closely connected with the (in)formality of the situation; the less formal the situation, the higher the degree of vagueness (ibid.: 86). Adolphs (2006) adds that "vague language is a particular feature of unplanned discourse" (Adolphs, 2006: 107). Salinger was probably aware of all of this and as he intended to make the speech of his protagonist authentic, it would be only reasonable to imitate the features of spoken informal language.

Nevertheless, it should be noted that not all occurrences of *and all* appear sentence finally as general extenders, as there are some instances when *and* functions as a coordinator connecting two sentences and it is followed by *all* only accidentally (ex. 13), though these instances are rather rare.

> 13. He put down his razor, **and all of a sudden** jerked his arms up and sort of broke my hold on him.

In most cases, *and all* functions as informal general extender and there are some passages when its repeated usage is especially noticeable, appearing almost in every other sentence (illustrated in ex. 14). Most of the time, the expression is used almost arbitrarily at the end of clauses and if deleted, the coherence would be preserved. In fact, it seems that this general extender is intentionally making the utterance vaguer and as a result, the speaker's attitude comes out as careless and lazy. In addition, there is a visible tendency for general extenders to co-occur with other vague expressions like *stuff* (109 hits) or *sort of* (179 hits).

> 14. I can't always pray when I feel like it. In the first place, I'm **sort of** an atheist. I like Jesus **and all**, but I don't care too much for most of the other **stuff** in the Bible. Take

[12] There is no generally accepted term for these expressions. They have been referred to as: "set marking tags (Dines 1980), vague category identifiers (Channell 1994), approximators (Erman 2001), general extenders (Overstreet 1999), discourse extenders (Norrby and Winter 2002), extension particles (Dubois 1992) and more" (Cheshire, 2007: 156). In the thesis, the expressions will be called general extenders, a term used, amongst others, by Overstreet 1999 and Cheshire 2007.

> the Disciples, for instance. They annoy the hell out of me, if you want to know the truth. They were all right after Jesus was dead **and all**, but while He was alive, they were about as much use to Him as a hole in the head. All they did was keep letting Him down. I like almost anybody in the Bible better than the Disciples. If you want to know the truth, the guy I like best in the Bible, next to Jesus, was that lunatic **and all**, that lived in the tombs and kept cutting himself with stones. I like him ten times as much as the Disciples, that poor bastard. I used to get in quite a few arguments about it, when I was at Whooton School, with this boy that lived down the corridor, Arthur Childs. Old Childs was a Quaker **and all**, and he read the Bible all the time. He was a very nice kid, and I liked him, but I could never see eye to eye with him on a lot of **stuff** in the Bible, especially the Disciples. He kept telling me if I didn't like the Disciples, then I didn't like Jesus **and all**.

Sometimes when using *and all*, the narrator hints that he deliberately leaves some information unsaid: there is the implication that there is more to say, but that the narrator dismisses it as unimportant and hence not worth mentioning. This attitude is especially visible in the opening sentence of the novel (ex. 15a). However, there is often no such implication and *and all* usually appears without any apparent function (15b) (this tendency was also observed by Costello (1959), as discussed in 2.7).

> 15. a. If you really want to hear about it, the first thing you'll probably want to know is where I was born, and what my lousy childhood was like, and how my parents were occupied **and all** before they had me, **and all that David Copperfield kind of crap**, but I don't feel like going into it, if you want to know the truth.
>
> b. It was Monday **and all**, and pretty near Christmas.

Other instances of *and all* could be seen as having evaluative function and in this case, it usually collocates with *all* used as an adverb which precedes the general extender; this repeated usage of the word *all* again explains its high occurrence in the corpus:

> 16. a. Anyway, the corridor was **all linoleum and all**.
>
> b. She was worried that it might make her legs lousy—**all thick and all**.

However, *and all* is certainly not the only general extender which the narrator uses. First of all, there is a number of expressions which are extensions of *and all*: in particular *and all that crap* (7 hits) or *and all that stuff* (4 hits). Secondly, there are other general extenders to be found, such as *or anything* (102 hits), *or something* (100 hits) or *and everything* (16 hits[13]).

[13] The word *everything* does not appear amongst the top 100 keywords (unlike *all*, *anything* and *something*) and therefore the phrase *and everything* is mentioned only for illustration, since it serves the same communicative purpose as other, more frequent general extenders and hence it contributes to the overall style of the book.

Or anything occurs typically in negative sentences, either expressing emotional reaction to the preceding utterance, mostly a surprise or irritation (17a), but more commonly it functions as a means of clarification: the narrator uses it in order to avoid any misunderstandings (17b).

> 17. a. He always looked all right, Stradlater, but for instance, you should've seen the razor he shaved himself with. It was always rusty as hell and full of lather and hairs and crap. He never cleaned it **or anything**.
>
> b. But he wasn't a bastard **or anything**. He was a very nice guy.

By ending clauses with *or something,* the narrator implies that he is not absolutely sure about the situation which he describes; by adding *or something* he makes the utterance less explicit as the conjunction *or* directly offers an alternative. This attitude can be recognized in example 18a: the narrator describes what he thinks his schoolmate Ackley does on Saturday nights, although it is more than likely that he has no idea how he spends his evenings and made this assumption only because Ackley does not go out often and has acne. Similarly, in 18b the narrator comments on the sign on the wall being inscribed with a sharp object, probably a knife, though he admits he is not sure.

> 18. a. The reason I asked was because Ackley never did anything on Saturday night, except stay in his room and squeeze his pimples **or something**.
>
> b. I went down by a different staircase, and I saw another "Fuck you" on the wall. I tried to rub it off with my hand again, but this one was scratched on, with a knife **or something**. It wouldn't come off.

Another function of *or something* appears in invitations: this general extender indicates, again, that there is an alternative option. In this case however, the speaker does not use it to expresses uncertainty, but he does it in order to make the proposal sound more casual. In this way, if the addressee rejects, the speaker would be more likely to avoid embarrassment. This casualness, which is however rather forced and functions as a defense mechanism, is especially clear in ex. 19a, as the invitation is preceded by *if she'd care to have (a hot chocolate)*. On the other hand, the fear of rejection is visible in 19b in *I said to her finally*, which suggests that the speaker first had to find the courage to invite the girl for a drink.

> 19. a. I asked her if she'd care to have a hot chocolate **or something** with me, but she said no, thank you.
>
> b. "Do you want to get a table inside and have a drink **or something**?" I said to her finally.

Or something also tends to appear as a part of a simile. Holden's comparisons are typically exaggerated, quite peculiar and certainly amusing, but the use of *or something* even deepens the comical effect as it indicates a degree of emotional distance and his lack of interest, as can be seen in ex. 20.

> 20. a. He started handling my exam paper **like it was a turd or something**.
>
> b. He put my goddam paper down then and **looked at me like he'd just beaten hell out of me in ping-pong or something**.

Finally, and this is perhaps most surprising, *or something* is, in majority of cases, employed in contexts where no other alternative is implied, as the content of those sentences is particularly specific. In these cases the general extender is used more or less arbitrarily and its main function is to shape the narrator's speech as intentionally vague and a bit careless (as is also the case of *and all*).

> 21. a. I'd have the damn gloves right in my hand and all, but I'd feel I ought to sock the guy in the jaw **or something**—break his goddam jaw.
>
> b. I call people a "prince" quite often when I'm horsing around. It keeps me from getting bored **or something**.

4.2.2.2 Generic Language

We can observe that the narrator tends to make generic statements a lot. This tendency can be discovered just by looking at some of the pronouns in the table like *somebody* and *nobody*: when studying the concordance lines in which they appear, it becomes clear that their high frequency in the corpus is partly due to their frequent usage in generic sentences.

If these pronouns (*somebody*, *nobody*) are employed in a sentence with generic meaning, they typically collocate with generic *they, he* and *you*. Some statements which the narrator makes could be truly seen as universal ones (ex. 22).

> 22. **You** can't teach **somebody** how to really dance.

However, the vast majority of generic sentences which are to be found in the *Catcher* corpus behave a little differently. The narrator, very frequently, makes a specific and personal statement and then he proceeds to turn it into a general one. First, this strategy forces the reader to relate more with the narrator, as he is led to consider the statement as if it were a universal truth. At the same time, by saying the same thing twice (first specifically, then more generally in ex. 23a, and vice versa in 23b), the narrator stresses his point.

> 23. a. Naturally, I never told him I thought he was a terrific whistler. I mean **you** don't just go up to **somebody** and say, "You're a terrific whistler."

> b. What I think is, **you**’re supposed to leave **somebody** alone if **he**’s at least being interesting and **he**’s getting all excited about **something**. I like it when **somebody** gets excited about **something**.

Then there are some instances which do not communicate the narrator’s life experience, but rather his opinions and ideas. His discussion on *Romeo and Juliet* is especially repetitive, also mixing the specific (names of the characters, their qualities) with the general. This example is particularly interesting: the statement *it drives me crazy if somebody gets killed (...) and it’s somebody else’s fault* is rather general and vague, but at the same time, we know very specifically that he’s talking about Mercutio, since the *somebody* who is killed is *smart and entertaining*.

> 24. “All those Montagues and Capulets, they’re all right — especially Juliet — but Mercutio, he was — it’s hard to explain. He was very **smart and entertaining** and all. The thing is, it drives me crazy if **somebody** gets killed — especially **somebody very smart and entertaining** and all — and it’s **somebody else’s** fault. Romeo and Juliet, at least it was their own fault.”

Also, it should be noted that sometimes the narrator chooses *he* as a reference to *somebody* (ex. 23b), but more frequently he employs *they*, a variant which does not indicate gender (ex. 25).

> 25. I don’t know if you’ve ever done it, but it’s sort of hard to sit around waiting for **somebody** to say **something** when **they**’re thinking and all.

Finally, it is this generalization and repetition, combined with casual language, which makes all the utterances especially amusing. The comical effect lies in that the narrator likes to make universally true statements based on his own personal experiences (26a). Perhaps the second reason why these utterances are so humorous is that the use of generic sentences is completely redundant in most cases, since the narrator describes events which are extremely specific (26b).

> 26. a. “You chose to write about them for the optional essay question. Would you care to hear what you had to say?”
>
> “No, sir, not very much,” I said.
>
> He read it anyway, though. **You** can’t stop a teacher when **they** want to do **something**. **They** just do it.
>
> b. What he did was, Richard Kinsella, he’d start telling **you** all about that stuff — then all of a sudden he’d start telling **you** about this letter his mother got from his uncle, and how his uncle got polio and all when he was forty-two years old, and how he wouldn’t let anybody come to see him in the hospital because he didn’t want anybody to see him with a brace on. It didn’t have much to do with the farm — I admit it — but it was nice. It’s nice when **somebody** tells **you** about **their** uncle. Especially when **they** start out telling you about their father’s farm and then all of a sudden get more interested in **their** uncle.

Negative sentences with generic meanings are less common, but they do appear. These sentences behave similarly to positive ones: we can expect generic *they* or *you* and the only difference is that the negative sentences comprise negative pronouns such as *nobody* or *anybody*:

27. a. I'd have this rule that **nobody** could do **anything** phony when **they** visited me. If **anybody** tried to do **anything** phony, **they** couldn't stay.

b. **You** always got these very lumpy mashed potatoes on steak night, and for dessert **you** got Brown Betty, which **nobody** ate, except maybe the little kids in the lower school that didn't know any better—and guys like Ackley that ate everything.

Finally, it should be stressed that definitely not every instance of these pronouns means that we are dealing with a generic reference. The indefinite pronouns quite commonly have non-generic reference, cf. ex. 28.

> 28. I'd only read about three pages, though, when I heard **somebody** coming through the shower curtains.

4.2.2.3 Contracted forms of verbs

General extenders, as well as the repetitiveness which is often present in generic statements, are both typical of spoken language. However, we should also look at other grammatical words whose usage would support the claim that Holden's speech is supposed to imitate spoken informal language.

When looking at the table of grammatical keywords, it is rather easy to notice that reduced forms, in particular *'d* (which stands for 'would' and 'had') and *'t* ('not'), display high frequencies of occurrence in the corpus. This does not seem very surprising as these forms are to be expected in colloquial language, however, they sometimes behave quite unexpectedly.

But first, let us look at the usage which we understand as unmarked. In example 29, we can see that the contracted form *'d* follows a personal pronoun and *'t* a negated modal verb, and they function as means of language economy.

> 29. Anyway, I couldn**'t** get that off my mind, so finally what I figured I**'d** do, I figured I**'d** better sneak home and see her, in case I died and all. I had my door key with me and all, and I figured what I**'d** do, I**'d** sneak in the apartment (…).

However, the reduced forms are particularly frequent: it seems that every time the narrator is allowed to make a contraction, he does so. In example 29, the only *had* which is not contracted is a lexical, one which does not allow reduction. This hypothesis is confirmed by the frequency data below.

I decided to study more deeply the contraction *'d,* and since the first person singular is by far the most common pronoun in the novel, I chose to run the search for the term *I'd*. The form *I'd*, meaning 'I would' or 'I had' appears 212 times. However, the full form *I would* is used only 8 times and in most of these cases, contraction would be impossible: in example 30, the verb *would* functions as a proform and it consequently cannot be reduced:

> 30. Then all of a sudden, out of a clear blue sky, old Sally said, "Look. I have to know. Are you or aren't you coming over to help me trim the tree Christmas Eve? I have to know." She was still being snotty on account of her ankles when she was skating.
>
> "I wrote you **I would**. You've asked me that about twenty times. Sure, I am."

Sometimes, the narrator uses the full verb form in order to avoid excessive reductions. In two cases *I would've* (ex. 31.a) is preferred to *I'd've*, even though the nonstandard form *I'd've* can be found four times elsewhere in the corpus (ex. 31.b).

> 31. a. I told him how **I would've** done exactly the same thing if I'd been in his place
>
> b. He had hold of my wrists, too, so I couldn't take another sock at him. **I'd've** killed him.

The frequency of *I had* in the corpus is 89. The lower ratio of reduction is due to the fact that the verb 'have' has more functions than 'would'. *Had* is used as a lexical verb (ex. 32.a), which cannot be contracted, in almost 60 clauses. In 24 instances the verb 'have' operates as a modal verb expressing obligation (ex. 32.b). Reduction of *had* is similarly unlikely when it has a causative function and when it acts as a proform (ex. 32.c). There is only one example of *had* being used as an auxiliary and hence potentially reducible (ex. 32d). All of this proves that if a verb can be reduced, the narrator is likely to contract the verb form.

> 32. a. All **I had** was three singles and five quarters and a nickel left — boy, I spent a fortune since I left Pencey.
>
> b. **I had to** go to the hospital and all after I hurt my hand.
>
> c. I didn't put my hands on her shoulders again or anything because if **I had** she really would've beat it on me.
>
> d. Not that I'd have done much about it even if **I had** known.

The frequent use of *I'd* also uncovers the speaker's marked tendency to use past perfect. Up to this point, the analysis showed results which indicated that Holden's speech is informal, a bit lazy and full of colloquial expressions. The perfect aspect, on the other hand, is not particularly common in spoken American English. The analysis of the concordance lines shows that past perfect is often employed in sentences in which the past simple would be sufficient; the use of past perfect is therefore redundant and the sentences display a degree of hypercorrection (ex.

33.a, cf. also 34.a-d). At the same time, the co-occurrence of informal colloquial language with the perfect aspect is also responsible for the humor in the novel (ex. 33.b). Generally, the combination of the formal and informal layers of language greatly contributes to the narrator's style, and it will be further discussed in Chapter 4.2.3, which deals with lexical words.

> 33. a. I was afraid some teacher would catch me rubbing it off and would think **I'd written** it.
>
> b. I was sorry as hell **I'd kidded** her.

There is one more thing worth commenting when discussing contractions. The contractions do not necessarily have to be preceded by a personal pronoun. The narrator often places the reduced verbs after compounds (ex. 34a), proper nouns (34b), common nouns (34c), adverbs (34d) and after another contraction, which results in such forms as *I'd've* and *I wouldn't've*.

> 34. a. **Somebody'd written** "Fuck you" on the wall. (…) I kept wanting to kill **whoever'd written** it.
>
> b. The only trouble was, the cold made my nose hurt, and right under my upper lip, where old **Stradlater'd** laid one on me. **He'd smacked my lip** right on my teeth, and it was pretty sore.
>
> c. After I got all packed, I sort of counted my dough. I don't remember exactly how much I had, but I was pretty loaded. My **grandmother'd just sent me a wad** about a week before.
>
> d. I mean I started thinking that even if he was a flit he **certainly'd** been very nice to me. I thought how he hadn't minded it when I'd called him up so late, and how he'd told me to come right over if I felt like it.

Finally, the contraction *'d* can occasionally stand for 'did', which also aims to imitate the spoken discourse. Nevertheless, this type of contraction appears only in direct speech (ex. 35).

> 35. a. "Leave it alone. **Why'd** he push you down the stairs?"
>
> b. "**What'd** she say?"

4.2.2.4 Non-standard spelling

Non-standard spelling is another strategy which is used to bring the narrator's and other characters' speech closer to informal spoken English. The pronoun 'you' is frequently spelt as *ya* and this kind of spelling is present only in direct speech. The non-standard spellings are not used to characterize a particular speaker; they occur both in the direct speech of the narrator and in the speech of other characters.

In close proximity of *ya* we can find other non-standard spellings of 'you' which imitate the pronunciation: *where'dja* for 'where did you' and *didja* for 'did you' (ex. 36).

> 36. a. "**Where'dja** get that hat?" Stradlater said. He meant my hunting hat. He'd never seen it before.
>
> I was out of breath anyway, so I quit horsing around. I took off my hat and looked at it for about the ninetieth time. "I got it in New York this morning. For a buck. **Ya** like it?"
>
> Stradlater nodded. "Sharp," he said. He was only flattering me, though, because right away he said, "Listen. Are **ya** gonna write that composition for me? I have to know."
>
> b. **Didja** have your lunch? **Ya** had your lunch yet?" I asked her.

In situations which are more emotionally tense, spellings reflecting pronunciation play a much larger role. In example 37, the first speaker, Maurice, is angry with Holden and he is not careful with pronunciation (*I tole ya*). Under normal circumstances, we can imagine that Holden would speak similarly carelessly (as shown in ex. 36), but in this particular situation, Holden starts speaking much more carefully (as reflected in standard spelling and lack of contracted forms) in order to create a distance between him and Maurice.

> 37. "It's ten bucks, chief. **I tole ya that**. Ten bucks for a throw, fifteen bucks till noon. **I tole ya that.**"
>
> "**You did not tell me that**. **You** said five bucks a throw. **You** said fifteen bucks till noon, all right, but I distinctly heard **you** — "

In other cases, Holden's use of informal language is highlighted by the non-standard spelling. In ex. 38, the spelling (*trimma goddarn tree for ya*) together with repetitiveness contributes to the impression of drunken speech.

> 38. "Yeah. **Listen**. **Listen, hey**. I'll come over Christmas Eve. **Okay? Trimma goddarn tree for ya**. **Okay? Okay, hey, Sally?"**
>
> "Yes. You're drunk. Go to bed now. Where are you? Who's with you?"
>
> "Sally? I'll come over and **trimma tree for ya**, **okay? Okay, hey?"**
>
> "Yes. Go to bed now. Where are you? Who's with you?"
>
> "Nobody. Me, myself and I." Boy was I drunk! I was even still holding onto my guts. "**They got me**. Rocky's mob **got me**. **You know that? Sally, you know that?**"
>
> "I can't hear you. Go to bed now. I have to go. Call me tomorrow."
>
> "**Hey, Sally! You want me trimma tree for ya? Ya want me to? Huh?**"

The non-standard *ya* does not always appear as a single word, but rather as a part of complex expressions, such as *willya* for 'will you' (12 hits). Different nonstandard spellings often occur in close proximity. In example 39. a., *willya* is preceded by *letcha up* ('let you up'). In fact, the search for '*tcha' returns 12 results, including *don'tcha* - ex. 39b. (5 hits), *can'tcha* (3 hits), *letcha* (2 hits), *ain'tcha* (1 hit) and *wutchamacallit* ('what do you call it'/'what is it called", 1 hit).

The second most frequent compound with *ya* is *wuddaya* ('what do you') with 11 occurrences, 3 of which is the composite *wuddayacallit* ('what do you call it'). As was mentioned, these words are likely to attract other non-standard spellings and this can be seen in example 39c.: *wuddaya* is followed by *tryna* ('trying to').

> 39. a. He said it over again. "Holden. If I **letcha up**, **willya** keep your mouth shut?"
>
> b. "Why the hell **don'tcha** shut up when I **tellya** to?"
>
> c. "**Wuddaya** mean what the hell am I doing? I was **tryna** sleep before you guys started making all that noise. What the hell was the fight about, anyhow?"

4.2.3 Analysis of Lexical Words

By lexical words (more detailed definition in 2.5.2.), we generally understand open word classes. This analysis works with the division of lexical/grammatical words as proposed by Scott and Tribble (2006), suggesting that lexical words consist of nouns, lexical verbs, adjectives and adverbs, while grammatical words consist of prepositions, determiners, conjunctions, auxiliaries and pronouns (ibid: 23). The keywords in the analysis have been divided into tables in accordance with this model only with one slight change: the table of lexical keywords also includes interjections, which are not mentioned by Scott and Tribble (2006) in the discussion. However, as interjections do not fulfil any grammatical function, they are understood as lexical words, even though their lexical meaning is questionable.

keyness	freq.	word
738,580	245	goddam
554,545	397	old
501,027	234	hell
304,309	126	damn
245,567	298	very
225,596	179	sort
188,036	146	boy
144,504	169	went
141,388	50	lousy
139,682	177	guy
139,005	71	sudden
136,676	120	kept
131,294	151	started
126,523	45	terrific
125,976	47	dough
125,097	238	even
116,902	49	bastard
116,160	183	mean
111,736	89	quite

keyness	freq.	word
79,135	62	finally
76,002	119	pretty
75,415	42	practically
70,715	28	sake
67,010	73	funny
65,982	228	really
63,307	21	crumby
61,059	81	guys
60,883	705	said
60,292	20	hunting
59,734	52	near
59,169	147	always
58,595	22	corny
58,464	50	listen
57,060	78	gave
54,633	37	certainly
54,263	18	sonuvabitch
53,098	245	go
52,926	77	crazy

103,643	244	too
96,873	35	phony
96,468	32	chrissake
96,003	107	stuff
93,453	31	sore
90,704	41	kidding
88,917	39	hat
86,443	54	hardly
81,407	82	nice

52,722	80	whole
52,665	134	told
51,248	17	helluva
49,616	82	kid
49,242	33	coat
48,234	16	madman
48,234	16	suitcases
46,929	18	gloves
45,219	15	horsing

Table 7: Lexical words within the top 100 keywords ranked by keyness (log-likelihood)

4.2.3.1 Controversial vocabulary

This part of the analysis is going to identify the main lexical elements which may explain what it was precisely that caused the controversy which surrounded the novel.

Firstly, we are going to focus on imprecations which appear relatively high on the keyword list. There is a degree of variation, as can be seen in *goddam* and *damn*, *hell* and *helluva* and *Chrissake* and *God's sake*.

Goddam is a word with the second highest value of keyness in the corpus. It is typically employed as an adjective: a closer look at the corpus revealed that it is typically followed by a noun: 213 times out of 245, in other cases there are two consecutive adjectives. *Goddam* has predominantly negative connotations and sometimes is used to condemn the referent of the noun which it modifies (*goddam fool*) and even more frequently it is employed to indicate the frustration of the situation overall (*goddam hand*), though often there is a combination of both approaches (40a). However, the word is very commonly used as an intensifier and its sole purpose is emphasis rather than condemnation or criticism (40b).

> 40. a. I was getting excited **as hell**, the more I thought of it, and I sort of reached over and took old Sally's **goddam hand**. What a **goddam fool** I was.
>
> b. I was the **goddam manager** of the fencing team.

Damn often behaves similarly as *goddam*, as it also tends to be employed as an intensifier (e.g. *damn good*, *happy*, *nervous*, *mad*, *tired*) and its connotations can be both positive and negative. *Damn* may appear as an adjective as well, however, these cases are relatively rare when compared to *goddam*. In this respect, *damn* is much more versatile.

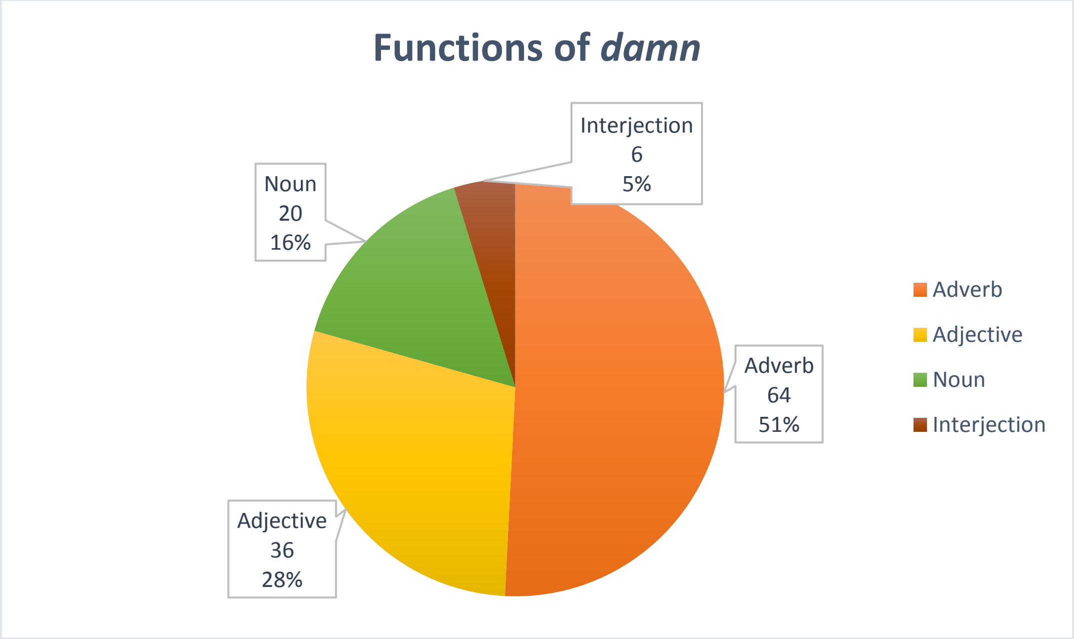

Figure 1: Functions of *damn*

The least frequent use of *damn* is an interjection and it typically expresses anger and irritation. However, *damn* never functions as an exclamation on its own, but only as a part of the phrase *god damn it* (ex. 41), though *damn* and *damn it* may be seen as equivalents.[14]

41. "Now, shut up, Holden, **God damn it** — I'm warning ya," he said (…).

Damn is also employed as a noun in the informal expression 'not give a damn' meaning 'not care about something'. The noun *damn* can be preceded by a degree modifier in this construction, e.g. *the type that doesn't give much of a damn if they lose their gloves*. The only other construction the noun *damn* occurs in in the corpus (2 hits) is 'worth a damn', e.g. *I couldn't pray worth a damn*.

Most importantly, *damn* operates as an adverb, and it usually modifies another one, as is the case of the frequent *damn near* (30 hits) or *damn well* (3 hits). *Damn near* is typically used in situations when something unwelcomed, for instance physical harm, almost happened but was avoided. Some of the collocations of *damn near* are listed in ex. 42, which also illustrates the fact that the narrator likes to make rather hyperbolical statements:

42. I **damn near**: fell down/over/off; broke my knee, broke my crazy neck, dropped dead, got killed,...

[14] < http://www.oxfordlearnersdictionaries.com/definition/american_english/damn_1>, July 26th 2017.

The word *hell* (234 hits) appears even more visibly as a part of specific language patterns, most notably as a comparison (with the function of an intensifier). Comparisons *as hell* (82 hits) and *like hell* (14 hits) together account for 41% of occurrences of *hell* in the corpus. Unlike *damn* and *goddam*, which were discussed before, *hell* does not seem to have a preference for positive nor negative connotations (ex. 43a and 43b, respectively). It should also be noted that although these comparisons mostly aim at people and their feelings or characteristics, they may occasionally describe inanimate objects and concepts, for instance *composition* which is *descriptive as hell*.

43. **a. positive adj. +** ***as hell***: cute/beautiful/charming/friendly/funny/pretty/excited/ suave/seductive/hot/kindhearted **as hell**

b. **negative adj. +** ***as hell***: sore/bored/sorry/drunk/nervous/sad/anxious/depressed/ embarrassed/lonesome/mad/scared/stupid **as hell**

Some word combinations are especially provocative: sometimes Holden combines words from religious contexts with the comparison *as hell*, which can make us realize that the language was perhaps criticized for being 'blasphemous' not only due to the large frequency of those words, but also because of how they were used. Nevertheless, the narrator does not seem to employ these words to provoke his readers, but it is rather a part of his idiolect: he uses it purely as an intensifier which is otherwise empty of meaning, for instance *modest/innocent* or even *religious as hell*. The point that *hell* as used in the novel is empty of meaning can be further proven by the fact that sometimes the narrator emphasizes two antonymous words by *as hell* (ex. 44).

44. cold/icy **as hell** X hot **as hell**; old **as hell** X young **as hell**

The high frequency of *hell* in the corpus can be further explained by its presence in formulaic expressions which typically consist of an interrogative pronoun followed by 'the hell' and are used for emphasis: *what the hell* (38 hits), *where the hell* (9), *why the hell* (8), *how the hell* (5), *who the hell* (5).

Another common cluster comprising of *hell* is *hell out* (*of*) (36 hits). This cluster usually appears as a part of the phrase 'get the hell out of somewhere' (9 hits), although other words apart from 'get' can be employed as well, such as 'bang', 'clear' or 'flunk' (ex. 45a). *Hell out of* also collocates with verbs expressing feelings (typically negative ones) and in this context, *hell out of* could be replaced by 'very much' (45b).

45. a. He **banged the hell out of** the room.

b. It/somebody **annoyed/fascinated/insulted/bothered/depressed/scared (the) hell out of** somebody.

The last cluster worth mentioning is *(just) for the hell of it* (9 hits). This phrase has usually rather positive connotations and it suggests that somebody does something only for enjoyment. Indeed, we can see that sometimes the activity described is innocent and often playful (46a), although the narrator also employs it in contexts where it is unexpected (cf. 46b)

> 46. a. I got bored sitting on that washbowl after a while, so I backed up a few feet and **started doing this tap dance, just for the hell of it**. I was just amusing myself. I can't really tap-dance or anything (…).
>
> b. I slept in the garage the night he died, and I **broke all the goddam windows with my fist, just for the hell of it**. I even tried to break all the windows on the station wagon we had that summer (…).My hand still hurts me once in a while when it rains and all, and I can't make a real fist any more.

The expression *helluva* ('hell of a') also has both positive and negative collocates (47a, b.). In addition, if we look at the expression *helluva time*, it is always context-dependent, as its meaning changes with the situation (47c, d).

> 47. a. **helluva** good sense of humor/kind face/pretty girl/humble guy
>
> b. **helluva** lot of trouble/long time/headache
>
> c. We had a **helluva time**. I think it was in Bloomingdale's. We went in the shoe department and we pretended she — old Phoebe — wanted to get a pair of those very high storm shoes, the kind that have about a million holes to lace up. We had the poor salesman guy going crazy.
>
> d. She was having a **helluva time** tightening her skate. She didn't have any gloves on or anything and her hands were all red and cold. I gave her a hand with it.

Finally, the controversy surrounding the language of the book also includes the criticism of vulgar expressions. In reality though, they are not particularly common in the text. In order to express anger or irritation, the narrator typically uses interjections *for Chrissake* (32 hits) and for *God's sake* (28 hits) and the already mentioned *goddam*. The only two swearwords found on the keyword list are *bastard* (49 hits) and *sonuvabitch* (18 hits) and only the second one functions as such all the time. *Bastard*, on the other hand, can either be employed as a swearword, usually surrounded by adjectives denoting a negative quality (ex. 48a), sometimes it is used more neutrally as a synonym for 'person' (ex. 48b), or it can be employed as an intensifier (ex. 48c).

> 48. a. phoniest/phony/nosy/crooked/stupid/showoff/lazy/rude **bastard**
>
> b. sexy/friendly/the only normal **bastard**
>
> c. somebody is drunk/getting drunk/shivering/sweating/limping **like/as a bastard**.

4.2.3.2 Formal and informal language

This part aims to uncover and describe probably the least apparent tendency in the text, which is the combination of (hyper)formal and informal layers language. This combination was already illustrated in the previous chapter (4.2.2) when looking at the high occurrence of the past perfect around colloquialisms, nevertheless, lexical elements may provide further examples to support this claim.

In the text, there are two more degree adverbs which function as intensifiers: *quite* (89 hits) and *pretty* (119 hits[15]). While the adverb *quite* may be associated both with formal discourse (typically academic prose) and with conversations (Biber et al. 1999, 545), *pretty* is an informal intensifier which is most frequent in spoken language (Leech and Svartvik, 2002, 217). The interaction between the two adverbs is illustrated in ex. 49. At the same time, it can be observed that *pretty*, although informal, may co-occur with Latinate adjectives like *sophisticated* or *intelligent*.

49. I was **pretty sadistic** with him **quite often**.

On the other hand, *certainly*, another frequently used adverb, displays a high degree of formality. There are few instances when its use is fully justified, as in ex. 50a, since Holden speaks to his teacher. Other times though, *certainly* seems to be a little redundant, either because it is surrounded by colloquial expressions (ex. 50b) or simply because the reader is used to the informal language and the sudden formality seems out of place (ex. 50c).

50. a. Do you blame me for flunking you, boy?" he said.

"No, **sir**! I **certainly** don't," I said.

b. "What the hell was the fight about, anyhow?" Ackley said, for about the fiftieth time. He **certainly** was a **bore** about that.

c. Then she stood up and pulled her dress over her head. **I certainly felt peculiar** when she did that.

This blending of different types of vocabulary in terms of formality can also be discovered by looking at the co-text of the most prominent colloquialisms, for example at the cluster *hell out of* which occurs in the sentence *She was ostracizing the hell out of me*.

Nevertheless, it should be stressed that even though there are sentences where it is the formal element which is intrusive, most frequently the language is markedly informal in rather formal contexts. It is exactly in those situations that the comical effect is most transparent. This

[15] some of the occurrences include *pretty* in its basic meaning, i.e. describing the external appearance of sb.

approach can be illustrated by the words *guy* (177 hits) and *guys* (81 hits) (ex. 51a). Example 51b illustrates the mixing of formal and informal vocabulary and vague language.

> 51. a. the navy/the elevator/the psychoanalyst/the salesman/spooky/touchy//very distinguished-looking **guy**.
>
> b. All these angels start coming out of the boxes **and everywhere**, **guys carrying crucifixes and stuff all over the place**, and the **whole bunch** of them — thousands of them — singing "Come All Ye Faithful!" **like mad**. **Big deal**. It's supposed to be **religious as hell**, I know, and very **pretty and all**, but I can't see anything religious or pretty, **for God's sake**, about a **bunch of actors** carrying crucifixes all over the stage.

4.2.3.3 Adjectives and Adverbs

The informality of the language is best apparent from looking at colloquial words which are typical of spoken American English: *lousy* (50 hits), *dough* (=money) (47 hits),*phony* (35 hits), *sore* (=angry) (31 hits), *kid*(*ding*) (41 hits),*practically* (42 hits), *corny* (22 hits), *crumby* (variant spelling of 'crummy') (21 hits) and *horse/ing around* (=behave in a silly way) (18 hits). Some of these words are common in colloquial speech even nowadays, e.g. *kidding* or *practically*, though most of them now sound rather outdated. This is especially the case of evaluative adjectives (*lousy*, *phony*, *corny*, *crumby*), as their popularity seems to be bound with the specific generation which uses them. This hypothesis can be partly confirmed if we try to search for these adjectives in the reference corpus: they return either 0 (*crumby*) or 1 hit (*lousy*, *phony*, *corny*). On the other hand, the reference corpus contains different evaluative adjectives, which are likely to sound more naturally to a present-day reader (e. g. *weird* in "*you are one weird dude"*).

It should be noted that the evaluative adjectives listed above are used by the narrator in order to malign other characters or mark the situation described as unfavorable (*phony advice*/*bastard*/*girls*/*guys*/*party*). Adjectives which are not particularly characteristic of informal spoken discourse include words like *nice* and*funny*. These expressions are used mostly to describe positive qualities of other characters.

4.2.3.4 Other personal idiosyncrasies

In this final part, we are going to uncover few more tendencies which are characteristic of Holden's idiolect. Firstly, there is his tendency to use the word *old* as a term of endearment much more frequently than to actually refer to someone's age (ex. 52a). Example 52b illustrates also that *old* tends to be used very informally, similarly to *guy(s)* and it is, again, very likely to collocate with other informal expressions.

52. a. **old** Phoebe/Sally/Spencer/Stradlater/Jane/Luce/Maurice/Ackley/Ernie/Thurmer

b. They were always showing Columbus discovering America, **having one helluva time** getting **old Ferdinand and Isabella** to lend him the **dough** to buy ships with, and then the sailors mutinying on him **and all**. Nobody gave **too much of a damn** about **old Columbus**.

The high frequency of *very* (298 hits) in the corpus uncovers a tendency which was also hinted at before, and that is that Holden's vocabulary is particularly repetitive. *Very* intensifies frequent, recurrent adjectives (ex. 53), and is often (14 hits) reduplicated to increase the degree (e.g. *he was very very tired or very very bored*). This gives the impression of a rather limited vocabulary.

53. **very** good/big/nice/funny/cold/depressed/hard/nervous/stupid/tiny/important/smart

Another kind of repetitiveness can be found in affirmations, which are realized by the word *really* (228 hits). *Really* typically (101 times) groups with auxiliaries and modals which function as proforms, out of which the most frequent ones are *really did* (23 hits) and *really was* (19 hits). These affirmations are mostly used after the narrator talks about something surprising which could be hard to believe, or simply for emphasis. Similar strategy is employed in clarifications, which are usually realized by *mean*, only in these cases there cannot be any proform, so either synonyms (54a) or repetitions (54b) have to be employed instead.

54. a. "Oh, well it's a **long story**, sir. **I mean** it's pretty **complicated**."

b. **She's very affectionate**. **I mean she's quite affectionate**, for a child. Sometimes **she's** even **too affectionate**.

5 Conclusions

The main purpose of this research was to identify and describe the main linguistic tendencies which shape the language of J. D. Salinger's novel *The Catcher in the Rye*. Perhaps the most apparent feature of the text is the imitation of spoken language, which was achieved mostly through the use of phonological devices (reductions and non-standard spellings imitating real-life speech). The immediacy of the spoken discourse was then approximated by repetitions and clarifications, and the lack of planning was reflected also by the frequent use of coordination and vague expressions. At the same time, the extracted keywords enabled us to discover more linguistic features which are not only characteristic of a spoken discourse, but which are also typical of Holden's speech as an individual. The most distinct elements which characterize the narrator's style include the mixture of higher and lower layers of vocabulary, hypercorrect use of the past perfect and, on the other hand, rather simplified vocabulary.

It also becomes clear that the narrator seems to prefer grammatical words for communicating the informality of his language over the lexical ones (general extenders and other vague expressions, excessive contractions of modals etc.). In addition, a lot of elements, which we would normally recognize as lexical ones (e.g. *hell*, *damn*), are empty of meaning and their only function is intensification. The lexical meaning can be similarly questioned in vague expressions like *sort* (*of*) and in other fillers, such as *practically*. As a result, a great number of key words do not carry full lexical information.

Grammatical elements and lexical words emptied of meaning are quite unlikely to disappear from language use and that is perhaps one of the main reasons why the text still appears to be accessible to present day generation. This is also confirmed by the fact that these words do occur in the reference corpus as well, though with much smaller frequencies. For instance, general extenders were found in the reference corpus but their occurrence was markedly lower when compared to the *Catcher* corpus.

This analysis also showed that the grammatical words and lexical words emptied of meaning behave in specific language and these patterns are also not very inclined to change (e.g. clusters around *hell* are the same in the reference corpus). Generally speaking, the words which have little or no lexical meanings are present on the keyword list to mark the speech of the protagonist as distinctly repetitive, but they cannot give us any concrete information on language change over time.

On the other hand, there is a number of lexical expressions which are typical only of Salinger's text. These are most importantly slang expressions (*horse around*) and evaluative adjectives (*lousy*), which seem to change in popularity quite quickly, as they are not to be found in the reference corpus. However, a synchronic study could show with certainty if these expressions are truly typical of the given time period or if they are popular only in Salinger's text. Translation study of the Czech translation could be useful as well, as it could serve as a good comparison of teen-language development in the two languages.

6 References and Sources

6.1 References

- Adolphs, S. (2006) *Introducing Electronic Text Analysis. A Practical Guide far Language and Literary Studies,* London: Routledge.
- Biber, D., Conrad, S. and Reppen R. (1998) *Corpus Linguistics: Investigating Language Structure and Use*, Cambridge: Cambridge University Press.
- Biber, D. et al (1999) *Longman Grammar of Spoken and Written English*, London: Longman.
- 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, 29–59.
- Costello, D. P. (1959) “The Language of 'The Catcher in the Rye'” in *American Speech*, Vol. 34, No. 3, pp. 172-181.
- Cheshire, J. (2007) “Discourse variation, grammaticalisation and stuff like that,” *Journal of Sociolinguistics* 11/2, 155-193. Oxford: Blackwell Publishing Ltd.
- Crystal, D. (2003) *The Cambridge Encyclopedia of the English Language*, Cambridge: Cambridge UP.
- Dušková, L. et al. *Elektronická mluvnice současné Angličtiny*. Available at http://emsa.ff.cuni.cz/. Accessed on July 12th 2017.
- Fischer-Starcke, B. (2009) “Keywords and frequent phrases of Jane Austen’s Pride and Prejudice. A corpus-stylistic analysis”, *International Journal of Corpus Linguistics* 14:4, 492–523.
- Fischer-Starcke, B. (2010) *Corpus Linguistics in Literary Analysis*, London: Continuum.
- Graham, S. (2007) J.D. Salinger's The Catcher in the Rye. London: Routledge.
- Gwynn, F. L. and Blotner, J. L. *The Fiction of J. D. Salinger*, Pittsburgh, University of Pittsburgh Press.
- Hoey, M. (1991) *Patterns of Lexis in Text*. Oxford: Oxford University Press.
- Hoey, M., Mahlbergm M., Stubbs, M., Teubert, W. (2007) *Text, Discourse and Corpora*, London: Continuum.
- Holmes, D. I. (1998) “The Evolution of Stylometry in Humanities scholarship”, in *Literary and Linguistic Computing* 13:3, pp. 111-17.
- Hyland, K. (2008) “As can be seen: Lexical bundles and disciplinary variation”, *English for Specific Purposes* 27(1), pp. 4-21.

- Laser, M. and Fruman, N. (1963) 'Not Suitable for Temple City' in *Studies in J. D. Salinger: Reviews, Essays, and Critiques of The Catcher in the Rye and Other Fiction*, New York: Odyssey Press.
- Leech, G. and Short, M. (1981) *Style in Fiction. A Linguistic Introduction to English Fictional Prose*, Harlow: Pearson Education.
- Leech, G. and Svartvik, J. (2002) *A Communicative Grammar of English*, London: Routledge.
- Lipka, L. (1992) *An Outline of English Lexicology : Lexical Structure, Word Semantics and Word-Formation*, Tübingen : Niemeyer.
- Mahlberg, M. (2007b) "Clusters, Key Clusters and Local Textual Functions in Dickens" , *Corpora*, 2 (1). pp. 1-31.
- Scott, M. and Tribble, C. (2006) *Textual Patterns. Key Words and Corpus Analysis in Language Education*. Amsterdam: John Benjamins.
- Scott, M.R. (1998) *WordSmith Tools Help Manual*. Version 3.0. Oxford: Oxford UP. http://www.lexically.net/wordsmith/version3/manual.pdf <Apr 28 2017>
- Stenström A., et al (2002) *Trends in Teenage Talk: Corpus Compilation, Analysis, and Findings*, Amsterdam: Benjamins.
- Stubbs, M. (2005) "Conrad in the computer: examples of quantitative stylistics methods", *Language and Literature*, 14, 1: 5-24.

6.2 Sources and Tools

- AntConc, version 3.4.4 by Laurence Anthony, available online at <http://www.laurenceanthony.net/software.html>, July 26th 2017.
- Alexie, S. (2014). *The Absolutely True Diary of a Part-time Indian*. New York: Little, Brown and Company.
- Cameron, P. (2007). *Someday This Pain Will Be Useful to You*. New York: Farrar, Straus and Giroux.
- Salinger, J. D. (2001). *The Catcher in the Rye*. Boston: Little, Brown.
- Thomas, R. (2013). *Rats saw God*. New York: Simon & Schuster BFYR.
- Vizzini, N. (2007). *It's Kind of a Funny Story*. New York: Miramax Books.
- Zuhr, L. (2014). *Life of a Loser - Wanted*. Createspace Independent Publishing Platform.
- Corpus of spoken American English of Santa Barbara available at <http://www.linguistics.ucsb.edu/research/santa-barbara-corpus>, July 26th 2017.

7 Resumé

Bakalářská práce podává korpusově-stylistickou analýzu románu J. D. Salingera *The Catcher in the Rye* (*Kdo chytá v žitě*), který poprvé vyšel roku 1951. Hlavním cílem práce bylo odhalit specifika daného textu, popsat styl vypravěče a identifikovat konkrétní rysy neformální mluvené americké angličtiny. Při práci s korpusem byla využita metoda keywords (klíčová slova). Klíčová slova byla identifikována na pozadí většího, referenčního korpusu, který se skládal z knih publikovaných mezi lety 1996 a 2014, které jsou psané podobným stylem jako román Salingerův. Práce tedy mimo jiné naznačuje i to, jak se za posledních šedesát let proměnil monolog amerického teenagera v populární literatuře.

První část práce podává teoretický úvod a definuje hlavní pojmy, které budou aplikovány v části praktické. Tato kapitola popisuje hlavní rysy korpusové stylistiky. Stylistika jako taková se zabývá především tím, jak jazyk studované knihy ovlivňuje její celkové vyznění, a tak spojuje studium literatury a jazyka. Stylistika často nachází využití i při studiu poezie či při identifikaci autorství literárních děl. Korpusová stylistika je chápána jako spojení korpusové lingvistiky a jejích metod s literární stylistikou. Tento přístup má řadu výhod. Elektronická analýza textu je objektivní a postup může být replikován jinými lingvisty, kteří tak mohou dojít ke stejným výsledkům nebo výsledky analýzy korigovat. Software, který se při takových analýzách používá, navíc zvládne přesně vygenerovat nejen frekvence slov, ale i jejich kolokáty, a tak odhalit jazykové vzorce, které by pouhé intuitivní čtení nemuselo zaznamenat. Manipulace s daty je navíc snadná a rychlá, což je důležité zejména v případech, kdy pracujeme s větším objemem dat. V neposlední řadě může elektronická analýza literárního textu nabídnout novou perspektivu pohledu na zkoumaný text, případně poskytnout přesné empirické důkazy pro již existující interpretace.

Korpusová stylistika na druhou stranu může být kritizována pro přílišnou selektivnost, vzhledem k tomu, že celý analytický proces je značně ovlivněn tím, co daný lingvista zkoumá. Toto riziko ale může být minimalizováno, pokud badatel bude pracovat se seznamy frekvenčních a klíčových slov, jelikož tato data jsou do značné míry objektivní. Nicméně konečné závěry studie jsou vždy výsledkem lingvisty, který své výsledky interpretuje a není proto vyloučené, že by jiný badatel mohl data interpretovat jinak. Dalším problémem může být to, že je tato metoda výzkumu relativně nová, a proto konkordanční programy často nemají všechny funkce, které by lingvista mohl chtít využít. Práce s daty navíc vždy vyžaduje manuální zkoumání výsledků, např. v případech, kdy se v textu vyskytují homografy.

Bakalářská práce dále zmiňuje výběr studií, které byly v tomto oboru doposud vytvořeny. Stubbs (2005) analyzuje témata v novele Josepha Conrada *Srdce temnoty* a Fischer-Starckeová (2009) se zaměřuje na klíčová slova v románu Jane Austenové *Pýcha a předsudek*. Culpeper (2009) využil metodu klíčových slov na zkoumání jazyka jednotlivých postav v *Romeovi a Julii*. Poslední zmíněnou studií je práce Mahlbergové (2007), která zkoumá klíčové ‚clustry' v Dickensových románech.
Práce také uvádí pojmy, se kterými budeme pracovat při samotném výzkumu. Klíčová slova jsou v této práci chápána jako slova, která jsou pro text „statisticky relevantní" (Culpeper, 2009: 30). Tato slova jsou extrahována na základě porovnání frekvenčních seznamů v cílovém a referenčním korpusu a ta slova, která budou vykazovat výrazně vyšší výskyt v jednom korpusu v porovnání s druhým, jsou slova klíčová. Klíčová slova jsou dále rozdělena do tří kategorií: slova gramatická, lexikální a vlastní jména. Gramatickými slovy rozumíme předložky, spojky, zájmena a pomocná slovesa a jejich zkoumání typicky odhaluje stylistické vlastnosti textu. Lexikální slova jsou naopak schopná identifikovat témata a obsah daného textu a skládají se ze substantiv, adjektiv, adverbií a lexikálních sloves. V neposlední řadě se práce zaměřuje na víceslovné výrazy, kde nejvýznamnější je vymezení rozdílu mezi kolokací a ‚clustrem'. Kolokací se chápe vztah, kdy se slova vzájemně přitahují a mají tendenci se spolu vyskytovat. Nicméně kolokací se nerozumí idiomatické výrazy, ani ‚clustry', které tvoří fixní slet několika za sebou jdoucích slov.

Teoretická část je zakončena debatou o dosavadních lingvistických studiích románu *Kdo chytá v žitě*. Grahamová (2007) i Costello (1959) pozorují, že jazyk hlavního hrdiny je často nejasný, což je důsledek toho, že řadu věcí čtenáři vůbec nesdělí. Na druhou stranu oba zaznamenali tendenci, která je zcela protichůdná, a sice že se vypravěč nápadně často opakuje v situacích, kde to vůbec není nezbytné. Zároveň oba naznačili, že by se román za několik desítek let mohl studovat podobně jako román Twaina *Dobrodružství Huckleberryho Finna*, tj. jakožto doklad o podobě mluveného jazyka určité věkové skupiny v oné konkrétní době.

Metodologická část uvádí texty, se kterými budeme při analýze pracovat. Cílový korpus tvoří pouze Salingerův román a referenční korpus se skládá z pěti knih vydaných mezi lety 1996 a 2014, které jsou psané podobným stylem, jako román Salingerův, a poslouží tedy jako norma, oproti které budeme Salingerův román zkoumat. Pro extrakci klíčových slov bude využit volně dostupný software AntConc vyvinutý Laurencem Anthonym. Při analýze budeme pracovat se sto klíčovými slovy seřazenými podle hodnoty keyness, která je udána statistickou kalkulací log-likelihood. Vygenerovaných sto slov následně rozdělíme na slova gramatická, lexikální a vlastní jména a zaměříme se na jejich kolokace a na to, jaké tvoří ‚clustry'.

Praktická část je uvedena seznamem frekventovaných slov. Vzhledem k tomu, že román je koncipován jakožto monolog hlavního hrdiny, dalo se předpokládat, že seznam frekventovaných slov bude reflektovat mluvený jazyk. Pro srovnání byl použit volně dostupný korpus mluvené angličtiny ze Santa Barbary, který skutečně tuto podobnost dobře ilustruje, a seznamy jsou si velmi podobné. Pro mluvený jazyk jsou typická především osobní zájmena a koordinační spojka *and*, která ukazuje na převahu souřadného souvětí. Salingerův román ale zároveň zřetelně disponuje stupněm organizovanosti a v textu chybí řada konverzačních prvků, jako například řečové neplynulosti (opakované začátky promluvy, repetice, opravy atd.) a výplňková slova.

Analýza vlastních jmen byla užitečná zejména pro utvoření představy o fiktivním světě, ve kterém se děj románu odehrává. Na seznamu figurují především jména hlavních postav. Názvy míst se vyskytují o poznání méně a obvykle označují školy, které vypravěč navštěvoval. Tato jména se typicky pojí s dalšími lexikálními prvky, které vypovídají především o obsahu děje.

Analýza gramatických klíčových slov odhalila řadu prvků, které jsou charakteristické jak pro mluvený jazyk, tak i pro osobitý styl vypravěče. Pravděpodobně nejnápadnějším rysem románu je vágní jazyk, který je realizován především vágními dovětky typu *and all*, které způsobují to, že výpovědi vypravěče působí značně nedbale. Tyto dovětky nemají jednotnou funkci a velmi často se v textu objevují arbitrárně. Dovětky s alternativní spojkou *or*, jako *or anything* a *or something* často naznačují možnost volby a jejich užívání je v důsledku o něco užší než transparentní *and all*.

Analýza také odhalila sklon vypravěče k užívání vět s generickým významem (typicky za užití zájmen *somebody*, *something* atd.). Tato tendence může být částečně vysvětlena tak, že vypravěč chce být čtenářem pochopen, a proto převádí své vlastní názory a zkušenosti na univerzální fakta. Tyto věty jsou ale inherentně velice obecné a jejich vyznění je značně nekonkrétní, a proto je možné naznačit souvislost těchto generických vět s vágním jazykem.

Mluvený jazyk je výrazně napodobován v přímé řeči postav, což je zřejmé z nespisovného psaní slov, např. *ya* (you), *willya* (will you), *can'tcha* (can't you) atd. Dalším indikátorem neformálnosti jazyka jsou stažené slovesné formy, které jsou v textu velmi frekventované, a to i na nestandardních místech, např. po substantivu či adverbiu. Redukce slovesných forem je navíc značně systematická a zkoumání konkordančních řádků dokázalo, že téměř každé sloveso umožňující redukci bude skutečně redukováno. Další zkoumání redukovaných forem také ukázalo, že jejich vysoká frekventovanost je způsobena přítomností předminulého času, který je v textu užíván nad míru a v situacích, kdy je minulý čas prostý zcela vyhovující.

Analýza lexikálních slov přinesla obdobně zajímavé výsledky. Od lexikálních slov se očekávalo, že rozpoznají hlavní témata a motivy zkoumaného textu, nicméně řada slov, jež řadíme k lexikálním, nemají téměř žádný lexikální význam. Toto je zřejmé zejména podíváme-li se na výrazy typu *hell* nebo *damn*: tyto výrazy fungují jako intenzifikátory, které kromě této intenzifikační funkce nenesou žádný jiný význam. Obě tato slova navíc tvoří typicky fixní ‚clustry' a obecně se chovají spíše jako slova gramatická. Podobná sémantická prázdnost je zřetelná i u slov, u kterých je toto chování méně typické, jako např. *bastard*, které funguje jako urážka, ale i jako synonymum pro neutrální výraz ‚člověk'. Podobná míra sémantické prázdnosti je viditelná i z vágního *sort* (*of*), nebo z výrazu *practically*, které funguje spíše jako výplň. Zbytek lexikálních slov pak tvoří převážně výrazy, které sice nejsou lexikálně prázdné, ale které ani neposkytují moc informací o obsahu textu, ale spíše charakterizují vypravěče. Jedná se typicky o hodnotící adjektiva, kterými vypravěč popisuje sebe a okolí, či adverbia, které popisují jeho pocity. Klíčová lexikální slova také také obsahují výrazy, které dále charakterizují idiolekt vypravěče, např. *old*, které funguje jako atribut vlastního jména.

Výsledky zde zmíněné popsaly jazyk vypravěče románu J. D. Salingera a zároveň identifikovaly rysy mluveného neformálního jazyka. Bylo překvapivé, že zkoumaný text nevykazoval výrazné znaky stárnutí, ačkoliv je pravděpodobné, že je to důsledek častého používání gramatických a lexikálně prázdných slov, která podléhají změnám méně často, než slova plnovýznamová. Tato domněnka může být částečně potvrzena, podíváme-li se na hodnotící adjektiva, které se v referenčním korpusu vyskytují minimálně. Pro další výzkum by mohla být přínosná synchronní studie, jež by mohla podat přesnější zprávu o tom, jaké prvky jsou skutečně dobové, a jaké jsou pouze typické pro Salingerova vypravěče. Analýza českého překladu by mohla být obdobně zajímavá, jelikož by mohla popsat vývoj jazyka teenagerů v obou jazycích.