Bakalářská práce

Chat language in World of Warcraft

Studijní program: B0114A300068 Anglický jazyk se zaměřením na vzdělávání

Studijní obory: Anglický jazyk se zaměřením na vzdělávání

Tělesná výchova se zaměřením na vzdělávání

Autor práce: **Jan Zástava**

Vedoucí práce: Mgr. Petra Peldová, Ph.D.
Katedra anglického jazyka

Liberec 2023

Zadání bakalářské práce

Chat language in World of Warcraft

Jméno a příjmení:	**Jan Zástava**
Osobní číslo:	P19000045
Studijní program:	B0114A300068 Anglický jazyk se zaměřením na vzdělávání
Specializace:	Anglický jazyk se zaměřením na vzdělávání Tělesná výchova se zaměřením na vzdělávání
Zadávající katedra:	Katedra anglického jazyka
Akademický rok:	2020/2021

Zásady pro vypracování:

Hlavním cílem této bakalářské práce je jazyková analýza internetového chatu ve hře World of Warcraft pomocí nástrojů a prvků korpusové lingvistiky. K analýze bude vytvořen korpus hráčského chatu dané hry, na kterém budou pomocí seznamu frekvenčních a klíčových slov, kolokací, a n-gramů identifikována specifika daného jazyka.

Bakalářská práce je rozdělena na dvě části: teoretickou a praktickou. Teoretická část se zaměřuje na definování terminologie v oblasti korpusové lingvistiky, čtenáře seznámí s tématikou použití zkratek v chatech a počítačových hrách. Dále bude ve zkrácené verzi nastíněna příběhová linie samotné hry. Praktická část popíše postupy a metodiku při samostatné analýze vytvořeného korpusu a zaměří se jazyková specifika ve vyjadřování hráčů ve hře World of Warcraft.

Rozsah grafických prací:

Rozsah pracovní zprávy:

Forma zpracování práce: tištěná/elektronická

Jazyk práce: angličtina

Seznam odborné literatury:

1. Čermák, František. 2017. *Korpus a korpusová lingvistika.* Praha: Karlova Univerzita.
2. McEnery, Tony and Andrew Hardie. 2012. *Corpus linguistics: method, theory and practice.* Cambridge: Cambridge University Press.
3. Baker, Paul. 2006. *Using Corpora in Discourse Analysis.* London / New York: Continuum.
4. Gablasova, Dana and Vaclav Brezina and Tony McEnery. 2017. "Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence." *Language Learning.* Volume 67, Issue 1, Pages 155-179.
5. Kameníková, Lucie. 2013. "Jazyk chatu (lexikální analýza)." Master's thesis. Palacký University Olomouc. Philosophical Faculty.
6. Nevařilová, Nela. 2008. "Zkratky v chatové komunikaci." Bachelor's thesis. Tomas Bata University. Faculty of Humanities.

Vedoucí práce: **Mgr. Petra Peldová, Ph.D.**
Katedra anglického jazyka

Datum zadání práce: **16. června 2021**
Předpokládaný termín odevzdání: **15. července 2022**

L.S.

prof. RNDr. Jan Picek, CSc.   **Mgr. Zénó Vernyik, Ph.D.**
děkan   **vedoucí katedry**

V Liberci dne 4. července 2021

Prohlášení

Prohlašuji, že svou bakalářskou práci jsem vypracoval samostatně jako původní dílo s použitím uvedené literatury a na základě konzultací s vedoucím mé bakalářské práce a konzultantem.

Jsem si vědom toho, že na mou bakalářskou práci se plně vztahuje zákon č. 121/2000 Sb., o právu autorském, zejména § 60 – školní dílo.

Beru na vědomí, že Technická univerzita v Liberci nezasahuje do mých autorských práv užitím mé bakalářské práce pro vnitřní potřebu Technické univerzity v Liberci.

Užiji-li bakalářskou práci nebo poskytnu-li licenci k jejímu využití, jsem si vědom povinnosti informovat o této skutečnosti Technickou univerzitu v Liberci; v tomto případě má Technická univerzita v Liberci právo ode mne požadovat úhradu nákladů, které vynaložila na vytvoření díla, až do jejich skutečné výše.

Současně čestně prohlašuji, že text elektronické podoby práce vložený do IS/STAG se shoduje s textem tištěné podoby práce.

Beru na vědomí, že má bakalářská práce bude zveřejněna Technickou univerzitou v Liberci v souladu s § 47b zákona č. 111/1998 Sb., o vysokých školách a o změně a doplnění dalších zákonů (zákon o vysokých školách), ve znění pozdějších předpisů.

Jsem si vědom následků, které podle zákona o vysokých školách mohou vyplývat z porušení tohoto prohlášení.

4. července 2023 **Jan Zástava**

Acknowledgements

First and foremost, I am deeply grateful to my supervisor, Mgr. Petra Peldová, PhD, for her invaluable guidance, constructive criticism, and valuable advice throughout the writing process of this bachelor's thesis. Her expertise and dedication have significantly shaped the outcome of this work. I would also like to extend my sincere thanks to my friends and my family for their unwavering support in all aspects of my life. Their encouragement and presence have been a constant source of strength and inspiration. I am profoundly appreciative of their love, understanding, and belief in my abilities.

Anotace:

Bakalářská práce se zaměřuje na provedení jazykové analýzy, konkrétně korpusové analýzy, s cílem prozkoumat jazyková specifika chatové komunikace ve hře World of Warcraft, především se zaměřením na zkrácená slova. Pro dosažení tohoto cíle byl vytvořen korpus WoWChatlog obsahující zprávy hráčů. Práce je rozdělena na teoretickou a praktickou část. Teoretická část definuje klíčové pojmy v oblasti korpusové lingvistiky, které jsou následně využity v praktické části. Dále se tato část zabývá počítačově zprostředkovanou komunikací a zkratkami, a také poskytuje stručný popis samotné hry. V praktické části jsou nejprve formulovány výzkumné otázky, které jsou následovány popisem tvorby korpusu WoWChatlog a metodologií využívanou při výzkumu. Dále je v této části prováděna samotná jazyková analýza a v poslední kapitole jsou prezentovány odpovědi na jednotlivé výzkumné otázky a vyvozeny závěry.

Klíčová slova:

Korpusová lingvistika, zkratky, World of Warcraft, chatová komunikace, videohra

Abstract:

The bachelor's thesis conducts a language analysis, specifically corpus analysis, in order to explore the language specifics of chat communication in the video game World of Warcraft, with a particular focus on abbreviations. To achieve this objective, the WoWChatlog corpus was created by collecting messages from players. The thesis is divided into theoretical and practical parts. The theoretical part defines key terms in the field of corpus linguistics, which are subsequently utilized in the practical part. Additionally, this section addresses computer-mediated communication and abbreviations and provides a brief description of the game itself. In the practical part, research questions are formulated; further, it describes the process of creation of the WoWChatlog corpus and the methodology employed in the study. Furthermore, the corpus analysis is conducted in this section, and the final chapter presents the findings addressing each research question and draws conclusions.

Keywords:

Contents

List of Figures

List of Tables

List of Graphs

List of abbreviations

@oms – atoms

2date – today

BG – battleground

BNC – British National Corpus

COCA – Corpus of Contemporary American English

CQS – computer query system

cuz – because

CZ – Czech Republic

D8 – date

DMG – damage

DPS – damage pes second

E-mail – electronic mail

exam – examination

FFA – free for all

flu – influenza

fone – phone

fridge – refrigerator

GBID – goldbid

GF – girlfriend

GW2 – Guild Wars 2

hc – heroic

IMCO – in my considered opinion

IMHO – in my humble opinion

IMNSHO – in my not so humble opinion

KWIC – keywords in context

LA – Lost Arc

lab – laboratory

LF – looking for

LFG – looking for group

LFM – looking for more, LF1M – looking for one more…

LoL – League of Legends

luv – love

LVL – level

maths – mathematics

MMORPG – massively multiplayer online role-playing game

MOBA – multiplayer online battle arena

net – internet, bot - robot

NP – no problem

NPC – non-playable character

OMG – Oh my God!

PoS – parts of speech

PVE – player versus environment

PVP – player versus player

SE – SketchEngine

SL – Shadow Labs

THKS, THX, TNX – thanks

TMRW – tomorrow

TUL – Technical University of Liberec

W – whisper

WoW – World of Warcraft

WT – want to

WTB – want to buy

WTS – want to sell

X – kiss

Y – yes

Introduction

Over the last 30 years, another viewpoint from which one can analyze language has evolved. This subdiscipline of applied linguistics is called Corpus linguistics. (Szudarski, 2018, 1). It deals with the construction and the analysis of corpora and offers a distinct perspective from which the researcher is able to analyze patterns of language use across various contexts of the language (ibid.). As McEnery and Hardie (2012, 228) state, "Corpus linguistics is ultimately about finding out about the nature and usage of the language." Corpus tends to be defined as a collection of texts, usually of large sizes, compiled together for the purposes of linguistics research with the aim of drawing conclusions about a particular language variety (Adolphs, 2006, 3). Utilizing a corpus analysis tool is essential due to the diverse and extensive nature of corpora (Szudarski, 2018, 19).

In a similar manner to corpus linguistics, but at a larger scale, the popularity of video games rose. According to a study executed in 2009 (Gentile, 2009, 596), almost 90% of American children between the age of 8 to 18 play video games at least occasionally. Since then, the popularity has only risen in the United States of America and Europe or Asia (Statista, 2023). Not only do kids shelter themselves in the entertainment of computer games, but plenty of adults seek some kind of escape from reality. In a single-player game, the individual is the only one with the ability to willingly respond to stimuli and with consciousness. Multi-player games, on the other hand, are crowded with players with the possibility of interacting with each other. As for some of these multi-player games, their player roster can reach up to millions. Interaction between players in these games can be essential for the complete experience of the gameplay. In some games, communication occurs in the form of an in-game chat, where players can write and see messages sent by other players (Suznjevic and Matijasevic, 2010, 1). The possibilities of a particular chat and its distinctive features vary depending on the game.

This work is divided into two parts: theoretical and practical part. In the theoretical part at the beginning, underlying principles and concepts relevant to corpus linguistics are discussed with the aim of introducing the topic to the reader. Furthermore, key analytical techniques and approaches to electronic text analysis are introduced. Thirdly, the reader is presented with a brief introduction to the so-called internet language or, in other words, computer-mediated communication and its different varieties. In addition, *Chapter 4* deals with the definition,

origins, and different classifications of abbreviations. As the title suggests, this work examines the language used in chat communication in a massively popular video game, World of Warcraft. Therefore, *Chapter 5* offers a terse game development history, basic gameplay principles, and an overview of chief objectives. The following and final part of the theoretical section seeks to introduce the in-game chat itself. The practical part of this work discusses the whole process of methodology and data collection, the course of the language analysis, and conclusions drawn from the corpus analysis.

For the purpose of this work, textual data from an in-game World of Warcraft chat consisting of messages sent by online players have been collected. Messages were taken from the chat by an in-game chat command "/chatlog", collecting all chat messages sent after the command. The collected text has been edited so that the unnecessary or undesired data, for instance, the name of the sender or the date and time the message was sent, were removed, and only the text messages would remain in the *WoWChatlog* corpus.

The chief purpose of this work is to present a theoretical background, perform a corpus analysis of WoWChatlog and potentially reveal its language peculiarities specific to the communication of players of World of Warcraft with particular emphasis on abbreviations, their use, and frequency. The analysis conducted in this research utilizes various features to examine the WoWChatlog. These features include wordlist analysis, which identifies the most commonly used words, keywords analysis, which generates and investigates a list of items based on their uniqueness to this corpus, and N-grams analysis enabling a more profound understanding of phrase usage involving 2-, 3-, and 4-word combinations. Additionally, the concordances feature is implemented to provide contextual examples of specific items. The working assumption of this thesis is that players tend to omit some parts of regular-speech language to enhance the visibility of the important parts of the message, and furthermore, they tend to excessively use abbreviations to compress the message with the intention of reducing the writing and reading time while still preserving meaning.

Theoretical part

1. Corpus linguistics

1.1.   Types of Corpora

It can prove to be a challenging task to name or define all types of corpora: As Schmitt (2010, 91) states: "It could be said that there are as many types of corpora as there are research topics in linguistics." However, some corpora types are worth noting, usually an opposing pair of types, for instance, Monolingual and Multilingual corpora, according to the number of languages included in a particular corpus (Wozniak, 2017, 4). There are corpora for both written and spoken speech. Another aspect, according to which corpora subdivide into general and specialized, is a topic. A Topic is another aspect of the corpora division. General corpora do not contain texts of one specific topic or origin but all kinds and varieties. Specialized corpora, on the other hand, contain one specific topic or the origin of the texts from one particular source (Čermák, 2017, 74).

General corpora contain a number of different texts, usually containing a large number of tokens. The size and variety of these texts are essential for the results of the analysis to be considered somewhat generalized (Bennet, 2010, 13). General corpora strive to provide users or readers with an optimal representation of a given language while recognizing the limitation that no single corpus can encompass the entirety of a language's possibilities. General corpora can be used as a representation of both written and spoken speech. For instance, *British National Corpus 2014*, a successor to the original *BNC*, represents a model example of both written and spoken language in Britain containing a lot of different texts ranging from newspapers, magazines, and novels to scholarly articles as well as all different kinds of conversations and other means of spoken speech (Brezina et al., 2021, 605).

General corpora are composed in a manner so they can be subdivided into sections, which can be further considered as specialized corpora. Specialized corpora aim to represent a particular type of language or language used by a particular group of people. For instance, the previously mentioned corpus *BNC 2014* is composed of different varieties of texts or groups of texts of the same kind, such as governmental documents, novels, and short stories or texts from popular magazines. Each one of those examples can be considered a specialized corpus and

represents a particular aspect of a language in which the researcher might be interested (Adolphs, 2006, 17).

1.2.   Corpus analysis tools

Acquiring or building a large corpus is a prerequisite of a thorough analysis in order for the results to be representative. Therefore, the researcher alone cannot carry out the analysis. To put that into perspective, the researcher is given 20 occurrences of a word in a 1000-word corpus. Under these circumstances, the researcher can analyze the result individually, whereas having to read and analyze 1000 or more results is unquestionably not an option. In this case, corpus analysis tools, often referred to as corpus query systems, come to the researcher's aid (Kilgarriff et al., 2004, 106). Nowadays, there is a variety of corpus analysis tools available on the internet to choose from. There are two main types of corpus query systems – computer-based tools and online tools. With the rising popularity of computers and later easily accessible faster internet connection, also the popularity of online systems has risen. Wordsmith tool is a computer-based tool in contrast to Sketch Engine as an example of an online corpus query system. Furthermore, they are categorized into simple/advanced tools by the features offered to the user. Corpus analysis tools offer plenty of functions and features, which range from simple functions, for instance, wordlist, keywords, collocations, and concordance, to more complicated or advanced, e.g., sketch-diffs, word sketches, or N-grams. Another division is corpus-related and corpus-independent tools. These vary in the ability to access only one specific corpus because the particular software was designed for the purposes of the analysis of that specific corpus (corpus-related) or the ability to access any corpus of choice depending on which the user chooses (corpus-independent) (Kilgarriff, Iztok, 2012, 5).

2. Corpus analysis

2.1.   Frequency data

In corpus linguistics, frequency data is fundamental information allowing the researcher to review the very basics of the corpus. There are two main ways to represent frequency data: raw frequency and relative frequency (McEnery and Wilson, 2001, 82). Raw frequency simply counts the number of times a word or phrase appears in a corpus, while relative frequency expresses this count as a proportion of the total number of words or tokens. Both measures have advantages and disadvantages, and the choice of which measure to use depends on the research question being addressed and the nature of the corpus under investigation (McEnery and Wilson, 2001, 83).

Raw frequency is often used to identify the most frequent words in a corpus and to compare the frequency of specific items across different subcorpora. Raw frequency can be useful in identifying the key features of a corpus, as well as for identifying differences between subcorpora (McEnery and Wilson, 2001, 83). Such an approach is implemented in this very research. For instance, the raw frequency could be used to identify the most frequent words in a political speech or a medical journal article, providing insight into the specific language used in those contexts.

On the other hand, relative frequency allows for comparisons of the relative importance of different items within a corpus, regardless of corpus size or composition. Relative frequency takes into account the overall size of the corpus and allows for a more accurate comparison of the frequency of items in different parts of the corpus (Baker, 2006, 68). This measure is particularly beneficial for identifying patterns of language use across different genres or registers, such as in the comparison of spoken and written language.

In conclusion, raw and relative frequencies are two important measures for analyzing frequency data in corpus linguistics. While raw frequency is useful for identifying the most frequent words and comparing frequency across subcorpora, relative frequency allows for comparisons of the relative importance of different items within a corpus and across different genres or registers. Researchers should carefully consider which measure to use depending on their research question and the nature of the corpus being investigated.

2.2.   Wordlist

The term wordlist refers to a list of lexical expressions along with their frequency information, according to which the most frequent ones can be determined and placed in the list in a descending or ascending order. One fundamental question arises concerning wordlist creation: "What should be counted as one item?" (Adolphs, 2006, 40). Commonly, there are five units of counting: Token, Type, Lemma, Flemma, and level 6-word family (Dang, 2019, 5). The term Token is referring to the number of all words in a corpus. Word type refers to every single unique use of a word. Lemma (produce) includes a stem (produc-) as well as the word's inflected forms (produce, producing, produced). Lemma does not include different parts of speech. Flemma, on the other hand, while being very similar to lemma, does not separate parts of speech from one another (production) (Dang, 2019, 6). For instance, in the sentence, *Children are playing videogames because it is fun even though it is unhealthy*; there are 13 separate words but only 11 unique words (types) in this sentence because the words *it* and *is* are used both twice. Only ten lemmas are present in this sentence because the word *be* is used in two forms – *is* and *are*. Obtaining frequency information in a target corpus is a fundamental part of corpus analysis and is a standard starting point for analysis. Sorting by frequency, the most frequently occurring word would be first, and the least would be the last or vice versa in a wordlist, along with the exact frequency of a particular item. There are different sorting criteria, for instance, an alphabetical order, a lemmatized format, or grammatical tags (Adolphs, 2006, 55). Since corpora are usually tremendously large collections of texts containing hundreds of thousands or even millions of words, corpus analysis tools are used to determine the most frequent items and to calculate their exact occurrence frequency (Čermák, 2017, 264).

Two opposing categories are contained in virtually all wordlists: lexical and grammatical words. The distinction between lexical and grammatical words is as follows. Grammatical words include articles, prepositions, conjunctions, and auxiliary verbs, which build a sentence structure. Lexical words, on the other hand, convey meaning and include nouns, verbs, adjectives, and adverbs (Corver and Riemsdijk, 2001, 1). Grammatical words, on the other hand, are usually short, with relatively little meaning, and often occur in sequences where their order and use are predictable, while lexical words are typically longer, with more specific meaning, and are more variable in terms of their usage and co-occurrence patterns (Krejtz et al., 2016, 223). This distinction is essential in corpus linguistics in order for researchers to identify patterns and frequencies of use of different types of words in a given language or text. By analyzing large amounts of text data, corpus linguists can explore the roles of lexical

and grammatical words in a language, including their distribution, collocation patterns, and semantic relationships. Another category, which is rare in both written or spoken corpora but appears in chat communication, is abbreviations, which will be discussed in *Chapter 7.3*.

2.3. Keywords

Keywords are items that occur considerably more or less in the target corpus in comparison with a larger reference corpus. Items that are more frequently used in the target corpus than in the reference corpus are called positive keywords. Conversely, the term negative keywords refers to items that occur with a lesser frequency in the target corpus than in the reference corpus. Comparison of all items in the wordlist of the target corpus with the corresponding identical pair in the reference corpus wordlist is necessary for establishing a keywords list (Adolphs, 2006, 44). For a valid list of keywords, it is essential to choose an appropriate reference corpus. Otherwise, the results may vary or might be invalid. For instance, comparing radio conversations of army officers with a corpus consisting exclusively of written texts, the different nature of written and spoken corpora might influence the validity of the output (ibid.).

2.4. Collocations

A thorough analysis of 2-word phrases in a corpus can be achieved through the examination of collocations. Collocation is a combination of words based on the (usual) co-occurrence of two words. When one particular word is used, it is statistically very likely that there will be one or more words that co-occur with this word more than other words (Bennet, 2010, 8). Two words collocate with each other if one meaningfully occurs with another more likely than with others (Čermák, 2017, 264). Collocations can be established through either sole consideration of frequency or, more commonly, through the utilization of a statistical metric known as an association measure. Association measures, also referred to as collocation measures, are statistical metrics that determine the degree of association between words by evaluating various aspects of their co-occurrence relationship (Gablasova et al., 2017, 159). For instance, the word *deal* as a noun form collocates with adjectives such as *great*, *big,* and *good*. These three adjectives co-occur with the word *deal* statistically more than other adjectives (perfect, ideal, beautiful…). Various collocations tend to have different associations with the word. An action or an event of great significance might refer to the phrase *a big deal*. On the other hand, the phrase *a good deal* often means a very profitable trade or a bargain, and last but not least, the phrase *a great deal* generally refers to a plurality or an amount. (Bennet, 2010, 9).

Collocations can be established using either frequency or a statistical measure known as an association measure, which is more commonly employed. Association measures, also referred to as collocation measures, are statistical calculations that determine the strength of association between words by analyzing various aspects of their co-occurrence relationship. It is worth noting that there are numerous association measures available, and each may generate a slightly different list of collocates (Brezina, 2018, 67).

2.5. N-grams

N-grams are a paramount tool in corpus linguistics for analyzing the structure of language. An n-gram is a contiguous sequence of n items, which are typically words but can also be letters or other linguistic units. By analyzing the frequency of occurrence of n-grams in a corpus, linguists can gain insights into the patterns of language use, such as the distribution of syntactic structures and the collocation of words. The use of n-grams in corpus linguistics has been further popularized by the availability of large digital corpora, which can contain millions or even billions of words. With the aid of computer software, researchers can efficiently generate n-gram frequency lists and perform statistical analysis on the data. N-grams can be generated for any given length of n, with the most common being bigrams (n=2) and trigrams (n=3). However, also longer n-grams can be analyzed, depending on the research question at hand. As computing power has increased, it has become possible to use larger N-grams, which can capture more complex patterns in language use. N-grams are useful for a wide range of linguistic applications, including language modeling, text classification, and many others. (Jurafsky and Martin 2000, 194). However, there are also limitations to the use of N-grams, such as the issue of data sparsity, where certain combinations of words may not occur frequently enough in the corpus to be reliably predicted (Manning and Schütze 1999, 201). Despite limitations, N-grams remain an essential tool in computational linguistics and continue to be used in a variety of applications.

3. Computer-Mediated Communication

To a certain degree, Computer-Mediated Communication (CMC) or Internet-based language shares some similarities with real-life verbal communication (Kerr and Hiltz, 2013, 3). Chat dialogue represents neither exclusively written nor spoken form of dialogue but stands somewhere in between and further also shares the receptive elements of reading (Hadziahmetovic et al., 2016, 1). The advantage of chat communication is that chat participants have the opportunity to interact with individuals from all over the world without being constrained by limitations of time and space (Kameníková, 2013, 11). In contrast with the spoken language, online chatting has its limitations. The absence of suprasegmental and non-verbal aspects of spoken communication appears to be the chief limitation. The possibility to use emoticons, also referred to as emoji or smileys, is a typical feature of an online chat. The lack of non-verbal aspects in the chat interaction can be partially restored with the option to use emoticons, with which the sender can express, for instance, their facial expression or a desired emotion (Derks and Bos, 2007, 843). Emoticons are further mentioned in *Chapter 4*. Real-time online chat has a very simple and straightforward concept. Two or more discussion participants join the conversation and then communicate between themselves in real-time in the form of short written messages. In order for the online chat to work, all the participants must be online. The written message is then sent to all other participants immediately (Hård af Segerstad, 2003, 125).

There are many internet chat varieties, each with different interaction possibilities and limitations. Crystal (2006, 10-17) and Lindh (2009, 5-6) have both listed situations where internet communication can be conducted in with thorough descriptions of different forms of language used.

- o *Electronic Mail (E-mail)* refers to electronic messages with the utilization of computer systems. Nowadays, emails consist mainly of messages between private mailboxes. E-mails are diverse in both language style and length; therefore, there is no general concept of describing them. The structure of e-mails, on the other hand, can be depicted into three parts, namely header, body of the message and greetings and farewells (Crystal, 2006, 10-11).
- o *Chatgroups* are virtual chat rooms on the Internet in which participants communicate with each other. These "rooms" are designated for a discussion about a certain topic, and whoever is keen on this topic can participate.

Chatrooms divide into two opposing groups, synchronous - messages can be sent and read in real-time, and asynchronous – sent messages are stored and can be viewed by users on demand (Crystal, 2006, 11).

o *"Virtual Worlds* are imaginary environments which people can enter to engage in text-based fantasy social interaction"(Crystal, 2006, 12). Unlike World of Warcraft, although similar, these virtual environments are strictly text-only based. Origins of virtual worlds date back to the 1970s role-playing game "Dungeons and Dragons."

o *World Wide Web (WWW)* refers to all public pages and websites accessible on every device connected to the internet. Commonly abbreviated to WEB or W3. Web involves billions of internet pages containing a variety of computerized data ranging from a tremendous amount of different texts and multimedia files to interactive software. (Crystal, 2006, 13-17)

o *Instant messaging* is a conversation, unlike chatgroups, usually between two known-to-each-other participants. Instant messaging, as the name implies, is synchronous and is commonly conducted with the help of software designed for it (Lindh, 2009, 5).

o *Blogging* is mostly a personal diary published in sections in intervals, similar to adding a description to a personal diary day by day. In the case of blogs or its non-abbreviated name, weblogs, the intervals between the addition of a new part of the blog is solely up to the publisher (Lindh, 2009, 5).

4. Abbreviations

Abbreviations can be defined as shortenings of words in different manners. Abbreviations are formed through various shortening methods, including initialisms, omissions, clippings, nonstandard spelling, and genuine novelties (Crystal, 2008, 37-53).

Abbreviations are not a modern phenomenon. The presence of abbreviations in Latin proves that humankind learned that their usage expedites the writing process and significantly saves space on the writing medium as early as before the 10th century (Zerkina et al., 2015, 140). The purpose for abbreviating words in these times was that the medium on which the text was recorded was either quantitatively limited or writing on it was challenging. Further abbreviations were used to speed up the writing process for the purposes of recording the spoken speech (Hrbáček, 1979, 19). Nowadays, abbreviations are used in any type of communication. Abbreviated words are surprisingly ubiquitous as abbreviations can be encountered in not only "living languages" but also in "dead languages." For instance, the previously mentioned language Latin includes abbreviations broadly used in present medicine, such as *ac* – ante cibum, meaning *before meal*, or *os* – oculus sinister, meaning *left eye* (Zerkina et al., 2015, 140).

Now the formation styles of abbreviations are going to be discussed, along with corresponding examples and their non-abbreviated forms. Firstly, **initialisms or acronyms** are words or combinations of words abbreviated to initial letters. There are two groups of initialisms distinguished in pronunciation. The first group includes abbreviations such as CIA or BBC pronounced by spelling letters used in the abbreviation. This group is called *alphabetisms*. However, there are some initialisms called *acronyms*, including, for instance, *NATO* or *UNESCO*, which are pronounced as a regular word (Plag, 2003, 162). Common jargon is nowadays influenced by internet language to such an extent that initialisms are sometimes used in spoken dialogue, especially regarding teenagers, who spend arguably the most time chatting (YouGov, 2017). This category of abbreviations can be divided into types:

- o Single words: Y (yes), W (whisper)
- o Compound elements: GF (girlfriend), BG (battleground)
- o Words in phrases: NP (no problem), OMG (Oh my God), WTS (want to sell)

(Crystal, 2008, 42)

Omissions are abbreviated words shortened by omitting one or more letters from various parts of the word. The vast majority of words can be shortened by omission, sometimes in various fashions meaning there can be different omissions of the same word since there are no rules for omitting. Omitted letters are mostly vowels for a number of reasons. One of which is the fact that the alphabet existed without vowels in the past, and only the consonants conveyed the meaning. However, there are a number of abbreviations with consonants also being omitted. Examples:

- o THKS, THX, TNX (thanks)
- o TMRW (tomorrow)
- o LVL (level)

(Crystal, 2008, 46)

Similar to omissions, **clippings** are a category of abbreviations where words are shortened by omitting a substantial part of the word. This category can be divided into four clipping types according to which part of the word was clipped. The first type is *Initial clipping,* also referred to as *foreclipping*, where the abbreviation is formed by omitting the introductory (initial) part of the word. Opposing the foreclipping, *Final clipping* forms an abbreviation by omitting the final part of the original word. In the case of *Medial clipping*, similar to Final clipping, the final part of the word is clipped with the exception of the final *–s* of the word, which is preserved. Lastly, *Ambiclipping* combines the attributes of both Initial and Final clipping and omits the beginning and the end of the original word. Examples:

- o Initial clipping: net (internet), bot (robot)
- o Final clipping: exam (examination), lab (laboratory)
- o Medial clipping: hc (heroic), mobs (mobile objects)
- o Ambiclipping: flu (influenza), fridge (refrigerator)

(Crystal, 2008, 51)

Nonstandard spelling refers to a technique of texting which consists of a conscious manipulation of spelling. This technique is the chief irritant to "haters" of internet texting; they argue that it is merely an error of spelling. However, the irregular spelling that can be observed in text messages generally gives the expression of conscious changes in spelling. Examples:

- o cuz (because)
- o fone (phone)

o luv (love)

(Crystal, 2008, 48)

Logograms and pictograms are a noticeable feature of Internet chat. Logograms represent the use of signs and symbols that represent whole words or concepts. In this case, meaning is conveyed by single letters, typographic symbols, numerals or combinations of numbers with the parts of the words (Crystal, 2008, 37). Logograms may stand alone in a text, or they may be used in combination with other letters, numerals, or symbols. Pictograms, on the other hand, use a visual shape to represent an object or concept, for instance, smileys or emoticons. Examples:

o X (kiss)
o D8 (date)
o 2date (today)
o @oms (atoms)

(Crystal, 2008, 48)

Genuine novelties can also be described as a form of a language play, where the user takes an already known abbreviation and modifies it further in his own manner. Therefore, the meaning of some novelties may be challenging to resolve, even for an experienced "texter." As an example, a known and simple abbreviation is taken – IMO (in my opinion). Possible examples of user modification:

o IMHO (in my humble opinion)
o IMNSHO (in my not-so-humble opinion)
o IMCO (in my considered opinion)

(Crystal, 2008, 48)

5. World of Warcraft

Before the capability to comprehend the players' interaction and communication that takes place in this video game, one must first be introduced to the basics of the game. Origins date back to 1994 when *Blizzard Entertainment* released a new real-time strategy game called Warcraft, in which a player completes specific objectives, for instance, constructing items, obtaining materials, or building an army to gain more power and later on to defeat the opponent. The game amazed people all around the world. This encouraged the developers to present two sequels (Ray, 2022).

In 2004, Blizzard came up with World of Warcraft, which is a massively multiplayer online role-playing game or a more commonly used term: *MMORPG*. The game is set in the Warcraft fantasy universe, primarily on the planet of Azeroth. However, new worlds such as Outland or Draenor were added to the game with new expansions. At the beginning of the game, the player first chooses to play either the Alliance or the Horde. Each of these two factions consists of several races. After reaching the maximum level, players do not have the opportunity to level up their characters anymore, but they try to acquire better equipment through PvP or PvE content. This phase of the game is called "endgame." PvP content consists of several areas where groups of the same faction battle against players of the opposing faction in order to gain better equipment. PvE content consists of a wide variety of 5-player dungeons and raids, in which players cooperate in groups consisting of up to 40 players to slay particularly powerful foes, ultimately to obtain specific equipment for more advanced and more difficult PvE content (Wowhead, 2021).

Since the release, Blizzard Entertainment has released nine major expansions: The Burning Crusade, Wrath of the Lich King, Cataclysm, Mists of Pandaria, Warlords of Draenor, Legion, Battle for Azeroth, Shadowlands, and the most recent expansion Dragonflight. With each expansion, more content was added to the game, such as new characters, zones, classes, professions, end-game content, or new storylines to explore (Blizzard, 2022).

5.1. In-game WoW chat

In video games similar to World of Warcraft, such as Guild Wars 2 or Blizzard's other *MMORPG,* Diablo III, interacting with other players is essential for the full experience of the entire content of the game. There is a 255-character limit for every message to prevent the lack of clarity in the chat since the chat window is in restricted size. Firstly, players can use chat to

interact with each other solely for their own entertainment. The second option, however, is much more common. In this case, players communicate with one simple purpose – to obtain better equipment or team up with other players (Rusaw, 2011, 77). Whether it is to find players to form a group or a raid to defeat mighty foes and complete dungeons or to find a worthy and reasonably cheap artisan, who could craft the equipment they need, both options include real-time interaction between players. World of Warcraft in-game chat is a very specific means of communication. In contrast with other instant messaging systems, such as *Whatsapp*, Blizzard as a developer of this game, does not offer a spectrum of cartoon smileys or the possibility to send pictures while interacting with other players, possibly to maintain the clarity and simplicity of the chat. However, there are some cartoon objects that can be used in WoW chat, which are displayed in *Figure 3,* for instance, square, circle, diamond, or triangle emoticons. Besides illustrated shape emoticons, players are able to use punctuation emoji (pictograms), as mentioned in *Chapter 3* and *Chapter 4.*

In-game communication is directed into channels. Some channels are created by the developers, and some are created by players. Every player can create his own channel with the chat command "/join [channel name]" (excluding the square brackets). Other players can then join that particular channel using the same command. (Nardi et al., 2007, 1) *World* is the chief player-created and -moderated channel with almost no topic restrictions. A player can get banned from using the world chat channel only when a player tagged as a moderator bans him. There are four public Blizzard-created channels *General*, *Trade*, *LocalDefense,* and *LFG* (Nardi et al., 2007, 2).

Despite the WoW chat communication being mostly synchronous, there is a possibility to send an in-game e-mail, which is akin to e-mails on the internet. A player can send a mail to any other character from the same faction as the sender. Similar to internet emails, the player is able to send an attachment with the message in the form of in-game currency or items. (Lindh, 2009, 5).

Practical Part

6. Data & Methodology

In the practical part of this work, the process of the *WoWChatlog* corpus is described, and a corpus analysis of the corpus is conducted with a particular focus on abbreviations within the chat messages with the intention of investigating specific aspects of World of Warcraft chat jargon. The following chapters provide a thorough and comprehensive description of the analysis procedure along with drawn conclusions at the very end.

The **research questions** for this thesis are as follows:

1. In which aspects does the chat communication in World of Warcraft differ from the everyday language represented by The British National Corpus?
2. To what extent do players implement abbreviations in their chat messages?
3. How effectively can abbreviations reduce the length of the original message?
4. Can World of Warcraft players communicate solely through abbreviations?

6.1. Creation of WoWChatlog corpus

In order to perform a corpus analysis, as the name suggests, a corpus needs to be obtained or built. This section of the bachelor's thesis deals with the creation of a text corpus for the purposes of the thesis, which consists of in-game chat messages sent by players of the *MMORPG* videogame World of Warcraft. Baker (2006, 43) states that a prerequisite of language corpora creation is to obtain a text or a collection of texts which contains a sufficient amount of words. The building and processing of a corpus consist of a sequence of various steps, such as text collection itself or sorting the texts into different documents and editing, which is a necessary step in order to form a corpus representative for the desired language area of communication that the corpus should contain (Čermák, 2017, 38). The text collection is the initial part of this process. In the case of *WoWChatlog*, an in-game chat command "/chatlog" was utilized, which initiates the collection of every message in the chat, along with other data that might be useful for whoever requires the chat to be saved. *Figure 1* presents the authentic appearance of a World of Warcraft chat directly screenshotted from the game. For better visualization, a chat log transcript of this chat with the exact same messages is displayed in *Figure 2.* This very figure also shows highlighted chat elements in various colors. For instance,

yellow is the date and the exact dispatch time of the particular message; pink indicates the channel in which the message is sent, the nickname of the sender is highlighted in green; the name of the server is gray, and, last but not least, the message itself is highlighted in blue. Despite the examples in this work being anonymous, in this case, real character names are used so that it is easier for a reader to visually compare the chat and the transcript. Individual character names provided in the transcript were taken with the owner's consent and permission. For the purposes of this paper´s research, some elements of the text were erased. Firstly, the channel highlighted in purple and the server name highlighted in gray hold no value for this analysis; thus, these pieces of information are omitted in the final corpus. Furthermore, the date and the exact dispatch time of the message highlighted in yellow are irrelevant for the purposes of the language analysis. Finally, yet importantly, the name/nickname of the sender is again not important and frankly unethical to store and display. Therefore, examples used in this work, as mentioned above, are presented anonymously ultimately to protect the sender.
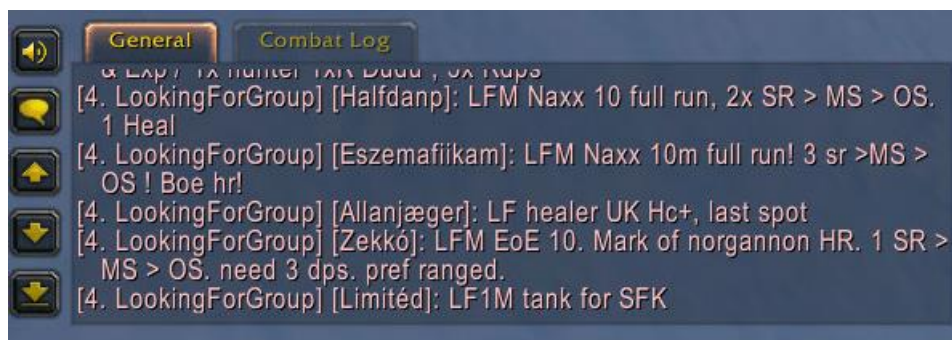
Figure 1: In-game WoW chat window

| Date and time | [Channel name] | Character-Server: | Chat message |

| Date and time | [Channel name] | Character-Server | Chat message |

3/4 12:51:04.101 [4. LookingForGroup] Halfdanp-Mograine: LFM Naxx 10 full run, 2x SR > MS > OS. 1 Heal

3/4 12:51:09.446 [4. LookingForGroup] Eszemafiikam-Mograine: LFM Naxx 10m full run! 3 sr >MS > OS ! Boe hr!

3/4 12:51:14.706 [4. LookingForGroup] Allanjæger-Mograine: LF healer UK Hc+, last spot

3/4 12:51:31.284 [4. LookingForGroup] Zekkó-Mograine: LFM EoE 10. Mark of norgannon HR. 1 SR > MS > OS. need 3 dps. pref ranged.

3/4 12:51:35.937 [4. LookingForGroup] Limitéd-Mograine: LF1M tank for SFK

Figure 2: Chat to chatlog transcription

Visual chat elements, such as cartoon objects, were also excluded from the corpus. These emoji transcribe from chat to chatlog document in a word form rather than their visual appearance in the in-game chat. They do not convey any meaning; players use them solely for improved visibility of their message comparable function to a real-life highlighter, and thus transcribed cartoon emoji such as star, square, moon, etc., are unwelcome in the final corpus. Examples of this chat-chatlog text transcription are shown in the following messages sent by my character; see upcoming *Figure 3*, which demonstrates the visual in-game representation and transcription of emoji to chatlog.

In-game chat:

[Shootashellz] says:

Chatlog emoji transcription (excluding date and time):

Shootashellz says: {circle}{triangle }{square}{moon}{star}{skull}{diamond}{cross}

Figure 3: In-game chat cartoon objects and chatlog transcription

After collecting a sufficient amount of text and filtering undesirable data in the word processing program Microsoft Word, a document containing purely messages from World of Warcraft players was created. The next step was to convert the messages into a simple text

document in which every new line represents a single message. A short excerpt of a final *WoWChatlog* corpus is presented below in *Figure 4.*

```
LFM AC N. Healer + DPS!!
LF tank ramp
LFM AC N. Healer + DPS!!
LF1M SL N - Healer
LFM tank Botanica Arca Mecha
LF2M SHH HC Tank+Mage Blaster Group
WTB Copious amounts of mage water
WTS Shadow Labs  XP/Rep Boost lvl 65-70| (70 mobs Rep/run|50g/run Summon available)  |LF heal and 1 dps for mechanar normal
WTS Shadow Lab boost 65 lvl+/50k exp(700+ rep) ne set 5 runs (got summon)
LF tank and DPS for SV Norm
```

Figure 4: WoWChatlog text sample

As a result of collecting and editing textual data from the World of Warcraft chat, a corpus *WoWChatlog* was established and is composed of 138 059 tokens and 99841 individual words, as can be seen in *Table 1*. The document alone contains 17 451 messages and 2867 unique word types used by various players. Out of all 2867 types, over a third (1037) appear in the corpus merely once. It suggests that the content of various messages in WoW chat seems to be repetitive and similar, even from different senders.

Table 1: Word counts of WoWChatlog

Tokens	138 059
Words	99 841
Types	2867
Chat Messages	17 451

With the corpus formed, the next phase prior to the analysis itself was choosing an appropriate and suitable corpus analysis tool. The corpus analysis tool of choice, *SketchEngine,* is software accessible via an internet browser. SketchEngine belongs to a category of paid corpus analysis tools. However, the Technical University of Liberec allows all students to benefit from this software using an institutional login via TUL email. From the author's standpoint, SE is the most user-friendly corpus analysis software available and accessible online. It offers a variety of features and functions paramount for the course of the entire

language analysis of a corpus, for instance, Wordlist, Keywords, Word Sketches providing collocations and a feature of comparing collocations of two words, N-grams, Concordance, etc.

6.2.　　Analysis Methodology

As mentioned in the chapter *Introduction*, the chief aim of this work is to analyze the language used in the World of Warcraft chat, focusing specifically on abbreviations. The thesis' working assumption is that some sentence elements essential in everyday speech tend to be omitted in World of Warcraft chat to enhance the clarity and visibility of the crucial parts of the message. Moreover, an additional assumption is that WoW players tend to include an abundance of abbreviations in their chat messages in order for the message to be as short as possible while still maintaining the meaning and thus reducing the writing and reading time of the message. The methodology implemented for the corpus analysis of *WoWChatlog* in this bachelor's thesis involves several distinct procedures. Firstly, the analysis commences by examining the items present in the wordlist. The general preview of the 100 most frequent words of *WoWChatlog* is discussed in the initial chapter of the analysis, followed by a brief comparison with the corresponding list of the 100 most frequent words within the *British National Corpus* (*BNC*). Subsequently, these 100 words from *WoWChatlog* and *BNC* are classified into three categories: lexical words, grammatical words, and abbreviations. Each category is individually inspected, investigated, and discussed in dedicated chapters. Moreover, the results of the wordlist analysis of *WoWChatlog* are continuously compared to the corresponding word category within the *BNC*'s 100 most frequent words. Additionally, the most frequent abbreviations utilized in *WoWChatlog* are explored, and these abbreviations are further categorized based on their formation style. Moving on to the subsequent chapter, the analysis delves into the investigation of the 50 strongest keywords present in *WoWChatlog*. To establish the keywords list, the *BNC*, as a representative sample of the English language, is chosen as a reference corpus. Thirdly, the analysis proceeds to examine the presence of N-grams within the *WoWChatlog*. Specifically, the focus is placed on identifying and examining phrases consisting of 2, 3, or 4 words, each analyzed in its own separate subchapter. Furthermore, the N-grams are categorized based on the number of abbreviated words in the phrase. Finally, all the gathered insights and findings obtained during the research process are evaluated, and each research question is addressed, leading to the formulation of conclusions in the last chapter of this bachelor's thesis.

7. Wordlist analysis

Commonly, the initial stage of a corpus analysis is an establishment of a Wordlist. The analysis of a wordlist is an important tool in corpus linguistics, as it allows the researcher to investigate the most frequently used words within the study corpus, as mentioned in *Chapter 2.2*. In this part of the bachelor's thesis, a wordlist analysis of *WoWChatlog* corpus is presented and discussed. The wordlist analysis aims to investigate the different categories of words within the corpus, such as lexical words, grammatical words, and abbreviations, and compare them with the wordlist of the *British National Corpus*. The comparison with the *BNC* will enable the identification of any significant differences or similarities between the language usage in the *WoWChatlog* corpus and that of a larger and more diverse corpus. Furthermore, the analysis will delve deeper into the formation of abbreviations according to the research discussed in *Chapter 4*. By conducting the analysis of abbreviations, a more comprehensive understanding of the language used in digital communication within the World of Warcraft, context can be achieved. This paragraph is followed by *Table 2: WOWTOP100* containing a list of the 100 most frequent words of *WoWChatlog*,

Table 2: WOWTOP100

| #	Item	#	Item	#	Item	#	Item	#	Item
1	lfm	21	sp	41	lvl	61	heroic	81	sethekk
2	tank	22	sl	42	sr	62	per	82	a
3	dps	23	the	43	mechanar	63	res	83	shattered
4	lf	24	sv	44	ub	64	on	84	at
5	hc	25	sm	45	summ	65	h	85	lab
6	heal	26	wtb	46	halls	66	slave	86	only
7	for	27	shh	47	cath	67	pens	87	lfg
8	need	28	bm	48	botanica	68	lf3m	88	reset
9	n	29	to	49	shadow	69	sum	89	rest
10	boost	30	lf2m	50	mt	70	exp	90	arca
11	and	31	bf	51	loot	71	is	91	deadmines
12	normal	32	summon	52	ac	72	ramps	92	arcatraz
13	wts	33	rep	53	ramparts	73	we	93	hunter
14	last	34	of	54	or	74	stockade	94	have
15	run	35	mage	55	zg	75	mobs	95	min
16	w	36	runs	56	got	76	blood	96	no
17	healer	37	daily	57	full	77	stratholme	97	more
18	spot	38	all	58	ready	78	armory	98	nether
19	lf1m	39	norm	59	can	79	farm	99	in
20	me	40	ohf	60	hr	80	guild	100	going

Using the corpus analysis tool, the Wordlist of *WoWChatlog* corpus was obtained. The observation of the wordlist revealed, as we can see in *Table 2,* that the first item on the list and

the most frequent word is, in fact, an abbreviation *lfm*. It suggests that users want to expedite the process of writing and reading the message. The fact that the most frequent item in the corpus is an abbreviation is a remarkable revelation, as it is highly uncommon for a corpus to have an abbreviation as the most frequently used word (Baker, 2006, 53). The above-mentioned *BNC* contains mostly grammatical words in the first 20 words on their wordlist, as can be seen in *Table 3,* which shows the 100 most frequent words of *BNC*. It can be assumed that also other general corpora share the same attribute for the 20 most frequent words on their wordlist. Baker (2006, 53) states that almost all forms of language have a high proportion of grammatical words. For instance, the written part of *BNC*, see *Table 3*, has, along with other grammatical words, personal pronouns or definite (1st) and indefinite (5th) articles among the first twenty places in the rankings of most frequent words, see *Table 3: BNCTOP100*. Furthermore, the most frequent items of *BNC* have been searched in the wordlist of *WoWChatlog*: The definite article is the 23rd and the indefinite article 82nd in the *WoWChatlog* frequency list. In addition, the personal pronoun "I", which is the 12th item in the frequency list of *BNC written*, is, in fact, on more than 150th place in *WoWChatlog*, which suggests that these particular sentence elements are not as frequently used, and crucial as in regular everyday speech.

Table 3: BNCTOP100

#	Item	#	Item	#	Item	#	Item	#	Item
1	the	21	are	41	do	61	its	81	new
2	of	22	have	42	been	62	could	82	these
3	and	23	this	43	has	63	into	83	also
4	to	24	not	44	their	64	then	84	any
5	a	25	but	45	if	65	time	85	first
6	in	26	from	46	will	66	other	86	very
7	that	27	had	47	would	67	him	87	know
8	it	28	they	48	what	68	two	88	people
9	is	29	his	49	so	69	only	89	see
10	was	30	or	50	can	70	like	90	after
11	for	31	which	51	no	71	my	91	should
12	i	32	she	52	when	72	than	92	such
13	on	33	we	53	more	73	well	93	where
14	you	34	an	54	up	74	now	94	because
15	with	35	there	55	who	75	did	95	how
16	as	36	n't	56	out	76	your	96	most
17	be	37	were	57	said	77	me	97	back
18	he	38	her	58	about	78	over	98	get
19	at	39	one	59	them	79	may	99	way
20	by	40	all	60	some	80	just	100	our

Studying the wordlist of the *WoWChatlog* using SketchEngine, it was further discovered that the corpus consists of 2868 unique words (types), which expand to slightly over 100 thousand individual words. It has been observed that the word with the tenth highest frequency out of a total of 2868 words within the corpus occurs with a frequency of less than one-third that of the most frequent word. If a linear pattern of declining frequency were to exist here, it would result in items listed at the 15th most frequent word or lower having zero occurrences, which is logically impossible. It is noteworthy that this phenomenon is not unique to the *WoWChatlog* corpus, but it is a typical attribute of corpora of all kinds and sizes (Brezina, 2018, 44). The principle is described as Zipf's law, which is also known as the principle of least effort or the law of diminishing returns. It states that in any corpus of natural language, the frequency of a word is inversely proportional to its rank in the frequency table. In other words, the most common word in a corpus will occur about twice as often as the second most common word, three times as often as the third most common word, and so forth (Brezina, 2018, 44). This law has been observed in various languages. Despite being a statistical regularity, Zipf's law is still not fully understood, and its underlying mechanisms remain an active topic of research. This phenomenon is displayed in *Graph 1: Item frequency decline* showing the decrease of the frequency of the most frequent items occurring in *WoWChatlog* as well as a linear decrease of trend frequency.
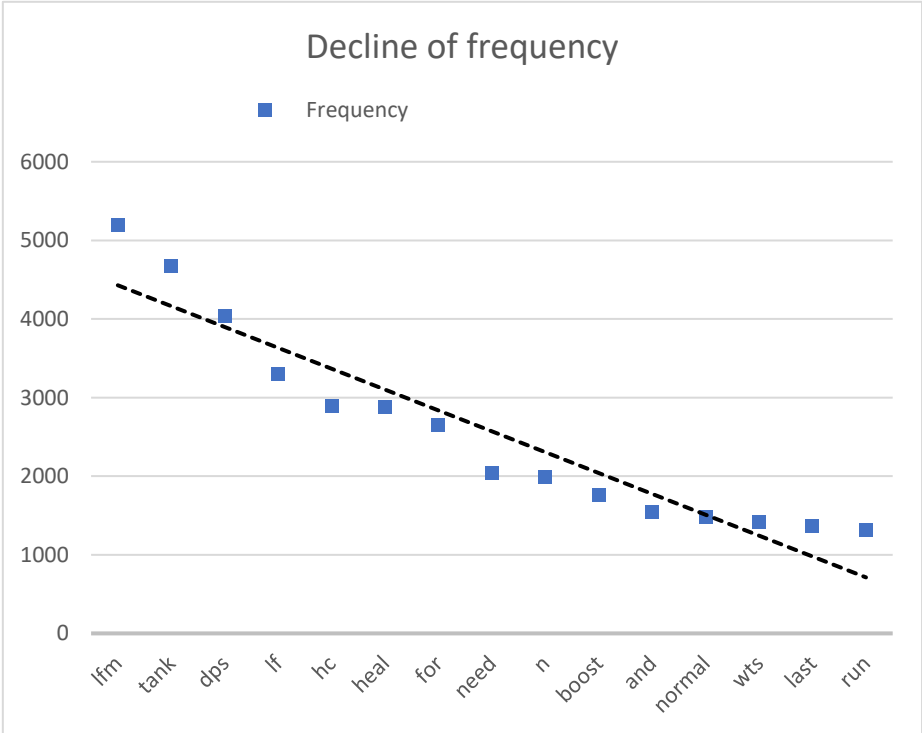
Graph 1: Item frequency decline

Division of the items into lexical and grammatical words in the wordlist that remain after the separation of abbreviations brought yet another intriguing discovery. Out of the 100 most frequent words of *WoWChatlog*, for the purposes of this work, also called WOWTOP100, there are 43 lexical words and only 16 grammatical words; the remainder are, of course, 41 abbreviations, which add up to 100 words. This division can be seen in *Graph 2.*
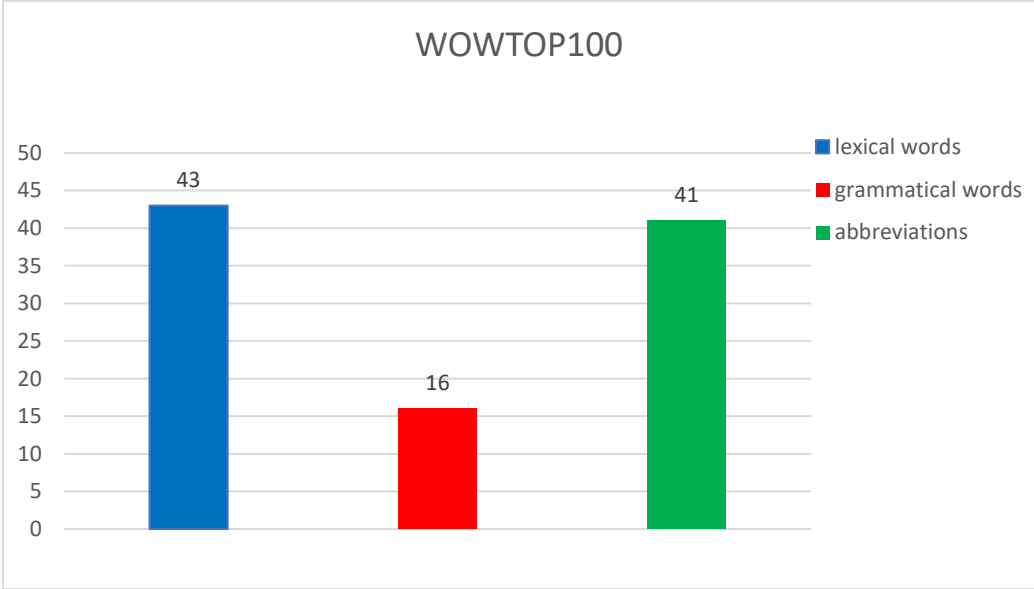
Graph 2: Word category distribution - WOWTOP100

For comparison, *British National Corpus* mentioned above contains a comparable quantity of lexical words (42) and a significantly higher amount of grammatical words (58) in *BNC*TOP100, approximately 2:3 in favor of grammatical words in contrast to WOWTOP100 where the ratio (lexical vs. grammatical words ratio) is highly in favor of lexical words, see *Graph 3.* Assigning a precise quantity of grammatical and lexical words and categorizing them by their parts of speech proved to be a difficult task considering that a single word in the English language can correspond to multiple categories of parts of speech, and it is virtually impossible to designate the PoS of every occurrence in the corpus. Therefore, the amounts are roughly estimated, and the exact numbers may vary slightly, and the deviation is displayed as error bars in *Graph 3*. Nevertheless, the comparison holds equal value, and itself alone suggests that in WoW, players tend to omit some grammatically significant elements of a sentence and replace them with lexical words or abbreviations in order to convey the fundamental meaning while preserving the message compressed.
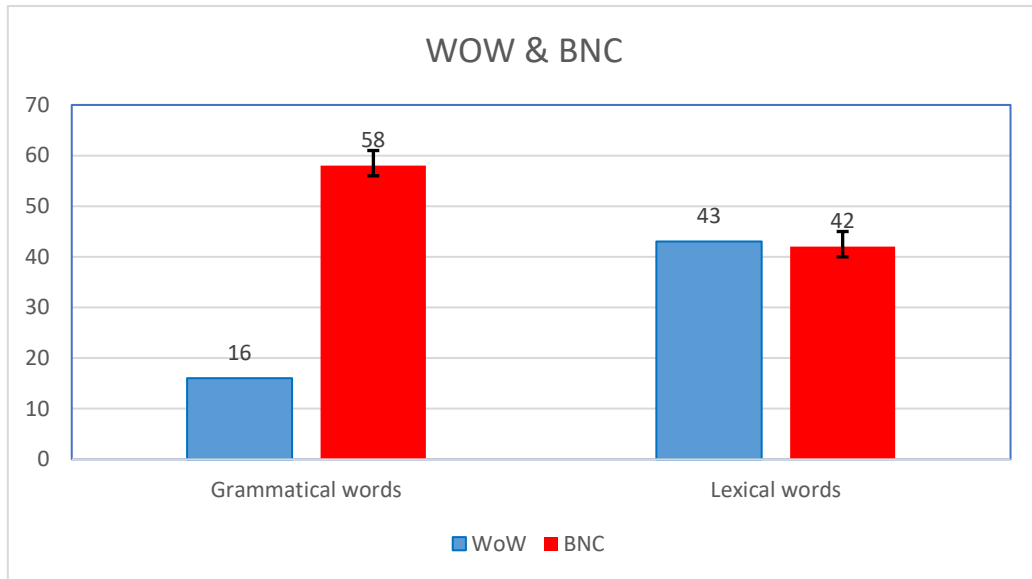
Graph 3: Word category distribution - comparison

7.1. Lexical words

Categorizing items from the wordlist by different parts of speech is equally problematic in the case of *WoWChatlog* as it was with the *BNC*. However, as a result of being strongly familiar with the items on the wordlist, it became possible to differentiate their manner of utilization and thereby classify them by their corresponding parts of speech category, ultimately facilitating the division. The predominant category is lexical words, with the occurrence of 43 items in the list of *WoWChatlog*'s 100 most frequent words, see *Table 4*.

Table 4: Lexical words of WOWTOP100

#	Item	PoS	#	Item	PoS	#	Item	PoS
1	tank	noun	16	botanica	proper noun	31	armory	noun
2	heal	verb	17	shadow	noun	32	farm	noun
3	need	verb	18	loot	verb	33	guild	noun
4	boost	verb	19	ramparts	noun	34	sethekk	proper noun
5	normal	adjective	20	got	verb	35	shattered	verb
6	last	adverb	21	full	adverb	36	only	adverb
7	run	verb	22	ready	adjective	37	reset	verb
8	healer	noun	23	can	verb	38	deadmines	proper noun
9	spot	noun	24	heroic	adjective	39	arcatraz	proper noun
10	summon	verb	25	slave	noun	40	hunter	noun
11	mage	noun	26	pens	noun	41	have	verb
12	runs	verb	27	is	verb	42	nether	adjective
13	daily	adverb	28	stockade	noun	43	going	verb
14	mechanar	proper noun	29	blood	noun			
15	halls	noun	30	stratholme	proper noun			

Lexical words from both wordlists were divided into groups by their parts of speech and then compared. Lexical words of *BNCTOP100* are shown in *Table 5*. Looking at the individual groups, significant deviations begin to emerge. Firstly, a high occurrence of verbs in both *WOW-* and *BNCTOP100* lists was observed, namely 14 verbs in the former and 21 verbs in the latter. The number of items in the category of adverbs on each of these lists indicates a significant disparity. In the latter, 21 verbs, together with 17 adverbs, represent the vast majority of items. However, the amount (4) of instances of adverbs in this WOWTOP100 presents a considerable decrease from the number of the instances in the *BNC*TOP100, presumably because adverbs are not crucial for the message to convey meaning and WoW players tend to be cautious about the unnecessary content of their messages.

Table 5: Lexical words of BNCTOP100

#	Item	PoS	#	Item	PoS	#	Item	PoS
1	is	Verb	15	would	Verb	29	new	Adjective
2	was	Verb	16	can	Verb	30	also	Adverb
3	be	Verb	17	up	Adverb	31	very	adverb
4	are	Verb	18	out	Adverb	32	know	Verb
5	have	Verb	19	said	Verb	33	people	Noun
6	had	Verb	20	could	Verb	34	see	Verb
7	not	Adverb	21	then	Adverb	35	should	Verb
8	there	Adverb	22	time	Noun	36	most	Adverb
9	n't	Adverb	23	only	Adverb	37	back	Adverb
10	were	Verb	24	well	Adverb	38	get	Verb
11	do	Verb	25	now	Adverb	39	way	Noun
12	been	Verb	26	did	Verb	40	no	Adverb
13	has	Verb	27	may	Verb	41	where	Adverb
14	will	Verb	28	just	Adverb	42	how	Adverb

Another comparable category of lexical words is nouns. Within the 43 lexical words in the WOWTOP100, nouns are represented by roughly half of all the items. In the BNCTOP100, however, nouns occur merely three times. Categorizing nouns into distinct noun groups, such as common and proper nouns, can prove to be a challenging task in this analysis. Several nouns included in this list are classified as common nouns, such as *halls*, *stockade,* or *blood*. However, players do not utilize them in their original sense, which can be found in the dictionary. Instead, these words are used as abbreviated designations of multi-word names of dungeons or in-game locations (*Sethekk Halls*, *Stormwind Stockade,* or *Blood Furnace* dungeons). Nevertheless, my vast experience of the game and the in-game chat made the division of nouns into common and proper nouns possible. There are six instances of proper nouns and 15 items that can be classified as common nouns in the WOWTOP100; however, eight of these 15 common noun instances refer to an in-game location, as mentioned above. With this new perspective, the list would exhibit a 14-item domination of proper nouns and a minority of common nouns with only seven occurrences. On the other hand, in the *BNC*TOP100, there are no occurrences of proper nouns, and thus the noun group is composed of exclusively common nouns. Such deviation from general English suggests that WoW communication tends to contain not only common but also proper nouns on a significantly more frequent basis than everyday spoken language. This occurrence can be attributed to the greater capacity of nouns to effectively convey meaning. A category of lexical words that has not yet been discussed is the group of adjectives that were also among the minorities, as the WOWTOP100 contains only four

adjectives, and the *BNC*TOP100 contains only a single occurrence, see *Table 4* and *Table 5*. However, there is one item in particular of all the lexical words in the WOWTOP100 that should be discussed further. This item is the adjective "only," and it is the single instance of an item that is included in both lists of most frequent lexical words. This word is commonly used to denote amount, quantity, or value[1]. This observation highlights the unique lexicon utilized in the World of Warcraft community and provides insight into the specific linguistic adaptations that have emerged within this specialized context.

Finally, looking at some lexical words individually, verbs in particular, an extraordinary peculiarity has been noticed. Several lexical items categorized as verbs in the wordlist exhibit a low frequency of use in their original parts of speech and a high frequency in other non-official parts of speech, namely nouns. For instance, the verbs *to heal*, *to summon,* and *to reset* are primarily employed as nouns in the corpus, as indicated by the examples A, B, and C provided below. One potential reason for this phenomenon is that in order for a sentence containing a verb to be comprehensible, it should also contain a subject and an object. By utilizing these verbs as nouns, even though they are not conventionally listed as such in dictionaries, players can avoid the obligation of incorporating additional words that may lengthen the message, ultimately making it more challenging for other players to read the message. In example A, the lexical item "heal" is employed as the object of the verb "need." In example B, the word summon is used as a noun and is accompanied by the abbreviated adjective "ready." Lastly, in example C, the prepositional phrase "per reset" complements the phrase "15g", which denotes the amount of in-game currency the player charges for the service he/she offers.

> Example A: LFM ZF need <u>heal</u> + tank
>
> Example B: LF healer to mech normal last spot <u>summon</u> rdy
>
> Example C: WTS boost SM cath and armory 15g per <u>reset</u>

The lexical words found in *WoWChatlog* and *BNC* wordlists display a complete deviation from a standard English written language, and it can be attributed to the specific nature of the *WoWChatlog* corpus as compared to the general nature of the *BNC*. The former is a specialized corpus, exclusively consisting of communication between players in a particular game, while the latter is a general corpus. This difference in context can lead to a discrepancy

---

[1] https://dictionary.cambridge.org/dictionary/english/only

in the frequency and usage of certain lexical items, as well as the prevalence of jargon and technical terminology specific to the game.

7.2. Grammatical words

In addition to the lexical words discussed in *section 7.1* above, there is another category of words worth investigating known as grammatical words. Grammatical words are essential elements of language and serve as the structural framework for constructing coherent and meaningful sentences. These words are often referred to as function words, as they indicate the grammatical relationship between different parts of a sentence. Examples of grammatical words include conjunctions, prepositions, articles, and pronouns, among others (Cruz-Pavía et al., 2021, 802). In the *WoWChatlog* corpus, there are 16 grammatical words, as can be seen in *Table 6*, which can be divided into six different parts of speech categories.

Table 6: Grammatical words of WOWTOP100

#	Item	PoS	#	Item	PoS	#	Item	PoS
1	for	preposition	7	all	pronoun	13	at	preposition
2	and	conjunction	8	or	conjunction	14	no	quantifier
3	me	pronoun	9	per	preposition	15	more	determiner
4	the	article	10	on	preposition	16	in	preposition
5	to	preposition	11	we	pronoun			
6	of	preposition	12	a	article			

These include seven prepositions, two conjunctions, both definite and indefinite articles, three pronouns, and, lastly, both determiners and quantifiers, each with only one instance. The BNCTOP100, on the other hand, predominantly contains grammatical words with the amount 58 items which can be divided into a variety of nine parts of speech. Although the frequency of grammatical words in WoWTOP100 is virtually a fourth of that in BNC, contrary to the lexical words, all of the grammatical words in the former list were also present in the latter with only one exception (per), see *Table 7*.

Table 7: Grammatical words of BNCTOP100

#	Item	PoS	#	Item	PoS	#	Item	PoS
1	the	Article	21	they	Pronoun	41	its	Pronoun
2	of	Preposition	22	his	Pronoun	42	into	Preposition
3	and	Conjunction	23	or	Conjunction	43	other	Adjective
4	to	Preposition	24	which	Pronoun	44	him	Pronoun
5	a	Article	25	she	Pronoun	45	two	Numeral
6	in	Preposition	26	we	Pronoun	46	like	Preposition
7	that	Conjunction	27	an	Article	47	my	Pronoun
8	it	Pronoun	28	her	Pronoun	48	than	Preposition
9	for	Preposition	29	one	Numeral	49	your	Pronoun
10	i	Pronoun	30	all	Pronoun	50	me	Pronoun
11	on	Preposition	31	their	Pronoun	51	over	Preposition
12	you	Pronoun	32	if	Conjunction	52	these	Pronoun
13	with	Preposition	33	what	Conjunction	53	any	Determiner
14	as	Preposition	34	so	Conjunction	54	first	Numeral
15	he	Pronoun	35	when	Conjunction	55	after	Preposition
16	at	Preposition	36	more	Determiner	56	such	determiner
17	by	Preposition	37	who	Pronoun	57	because	Conjunction
18	this	Pronoun	38	about	Preposition	58	our	Pronoun
19	but	Conjunction	39	them	Pronoun			
20	from	Preposition	40	some	Pronoun			

This suggests that while the choice of lexical words may vary significantly depending on the context and style of communication, the selection of grammatical words might remain relatively consistent even across different genres and registers of English, which is again proved by Baker (2006, 53). These words are necessary for sentence formation and serve as the glue that holds a sentence together. Without grammatical words, a sentence may still convey some meaning thanks to lexical words, but it would lack coherence and clarity. In the author's experience, this is particularly important in online gaming, where players often rely on clear and concise messages to coordinate actions and make strategic decisions. In conclusion, while the occurrence of grammatical words in the World of Warcraft chatlog tends to be significantly lower than in everyday English and their inclusion remains essential for effective chat communication.
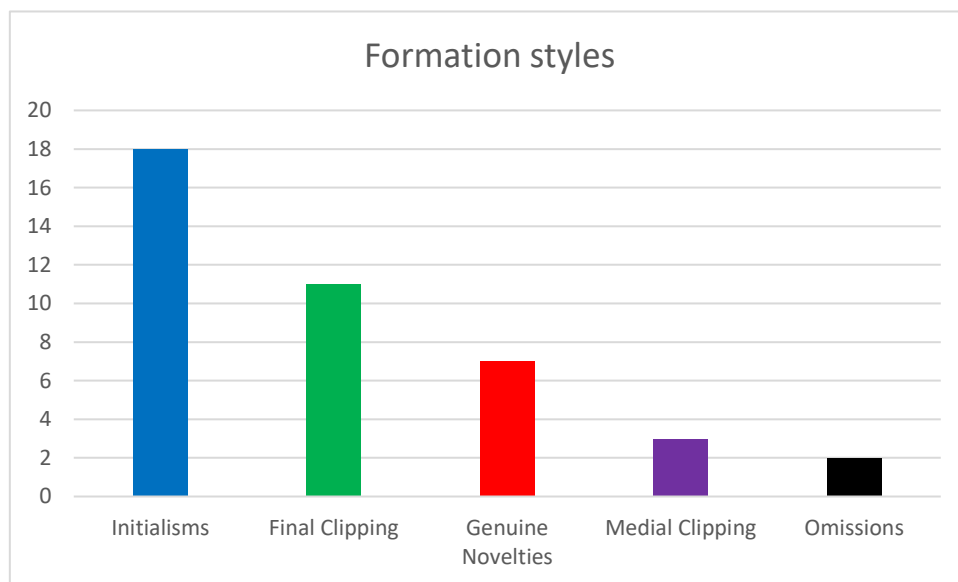
7.3. Abbreviations

In this paragraph, an analysis of the most frequent abbreviations used in this context and their formation style are to be revealed and discussed. The corpus analysis revealed that out of the top 100 most frequent words used in *WoWChatlog*, 41 were abbreviations. For comparison, in the written part of *BNC*, only five abbreviations, such as *p* (page), *mr* (mister), *mrs* (missus), *st* (saint), and *dr* (doctor), can be found in the list of the 1000 most frequent words. In the same scenario, *BNC 2014* spoken contains only four abbreviated words, namely *dunno* (abbreviated informal spoken form of "do not know"), *tv* (television), *innit* (abbreviated informal spoken form of "is not it"), and *u* (you). All 41 abbreviations present in the WOWTOP100 list are displayed in *Table 8,* along with their formation style category.

Table 8: Abbreviations of WOWTOP100

#	Item	Formation Style	#	Item	Formation Style
1	lfm	Genuine Novelty	22	sr	Initialism
2	dps	Initialism	23	ub	Initialism
3	lf	Initialism	24	summ	Final Clipping
4	hc	Medial Clipping	25	cath	Final Clipping
5	n	Initialism	26	mt	Initialism
6	wts	Genuine Novelty	27	ac	Initialism
7	w	Initialism	28	zg	Initialism
8	lf1m	Genuine Novelty	29	hr	Initialism
9	sp	Initialism	30	res	Final Clipping
10	sl	Initialism	31	h	Initialism
11	sv	Initialism	32	lf3m	Initialism
12	sm	Initialisms	33	sum	Final Clipping
13	wtb	Genuine Novelty	34	exp	Final Clipping
14	shh	Omission	35	ramps	Medial Clipping
15	bm	Initialism	36	mobs	Medial Clipping
16	lf2m	Initialism	37	lab	Final Clipping
17	bf	Initialism	38	lfg	Initialism
18	rep	Final Clipping	39	rest	Final Clipping
19	norm	Final Clipping	40	arca	Final Clipping
20	ohf	Initialism	41	min	Final Clipping
21	lvl	Omission			

These abbreviations were manually sorted into five different categories according to the formation style research included in *Chapter 4* of the theoretical part: initialisms, medial clippings, final clippings, omissions, and genuine novelties. A substantial amount of these abbreviations (43%) were initialisms, consisting of two or more letters, and players read and

pronounce them as individual letters. A total of 34% of abbreviations on this list were formed by clipping, namely three abbreviations, 7% out of all 41 abbreviations, through medial clipping, which involves omitting the middle portion of a word, and 11 abbreviations (27%) through final clipping, which involves removing the final part of a word. There were only two instances (5%) of formation by omission, which involves leaving out one or more letters from a word not exclusively next to each other. Finally, 17% of abbreviations were classified as genuine novelties, as they were created by the player taking an already known abbreviation (LF – Looking For; WT – Want To) and modifying it in his/her own manner (lfm, lf1m, lf2m, lf3m, lfg, wts, wtb). *Graph 4: Formation style of abbreviations in WOWTOP100* exhibits the distribution of formation styles categories of abbreviations.[2]

Graph 4: Formation style of abbreviations in WOWTOP100

The findings indicate that initialisms tend to be the most commonly implemented formation style of abbreviations used in chat messages from players of World of Warcraft. These initialisms are derived from non-abbreviated forms that are virtually all two or three words long, condensed into an equal number of letters. The frequent use of initialisms is likely due to the fact that other formation styles, such as medial or final clipping, are not capable of omitting this many letters, compressing the text to such a degree while still preserving the meaning of the message. Initialisms, unlike other formation styles of abbreviations, preserve only a single letter from the original non-abbreviated words included in the particular abbreviation and thus are the most suitable for World of Warcraft players. By using initialisms,

---

[2] Percentage values in this paragraph are rounded to whole numbers.

World of Warcraft players are able to save both writing and reading time, as well as make room in the message for other important information. This suggests that players of World of Warcraft prioritize efficiency and brevity in their chat messages, as the use of initialisms allows for a more concise and streamlined form of communication. It is imperative to note that the use of abbreviations in chat messages is not limited to World of Warcraft, and this study's findings may have implications for other online gaming communities or digital communication contexts where time and message length are of significance.

As *Figure 1* shows, the in-game chat window is rather small and is capable of displaying only a few messages at a time. Players of World of Warcraft are, therefore, very cautious about the content of their messages. Protracted and essential information is usually shortened into abbreviations in order to hasten the reading since the message might be displayed in the game chat only for a very short time, no more than a couple of seconds during peak playtime of the day. Therefore, it is essential for the players to conduct their communication in messages that are as brief and concise as possible, usually omitting grammatical words leaving predominantly lexical items which convey the meaning.
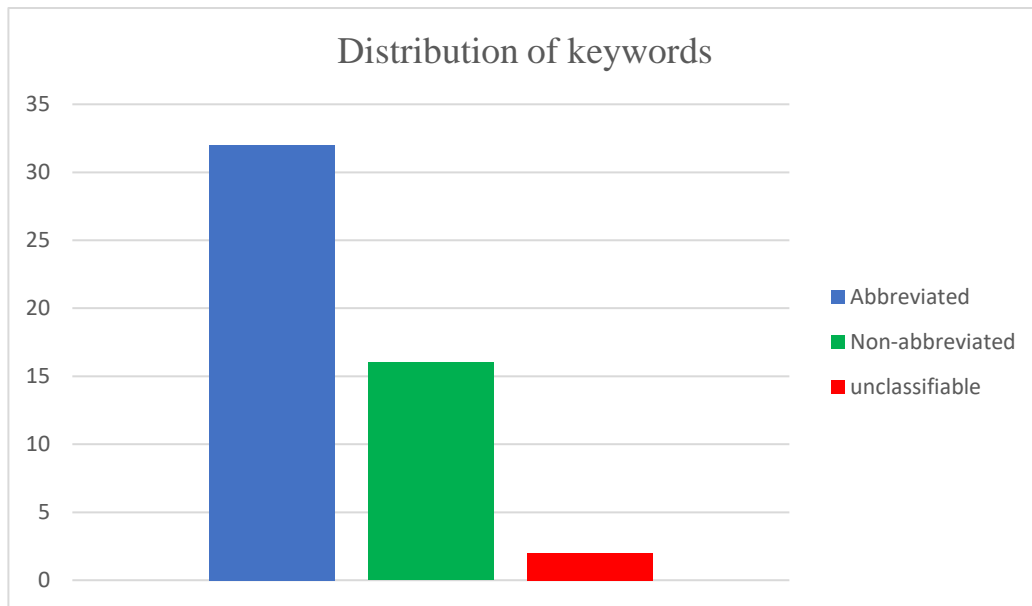
8. Keywords analysis

This section of the Bachelor's thesis focuses on the analysis of keywords in a *WoWChatlog* corpus created for the purpose of this research. The aim of this analysis is to explore the language patterns and identify frequently used words and abbreviations in the context of World of Warcraft chat, which are less common in general English. When creating a keyword list, it is essential to designate a proper, larger reference corpus (Sardinha, 2000, 7). For the purposes of the keywords analysis, once again, the *British National Corpus* was selected, this time to be the reference corpus. In *Table 9* below, the first 50 items on the keywords list of *WoWChatlog* are displayed.

Table 9: 50 strongest keywords of WoWChatlog

#	Item	#	Item	#	Item	#	Item	#	Item
1	lfm	11	*dp*	21	zg	31	arcatraz	41	*ffa*
2	dps	12	sl	22	mage	32	arca	42	cath
3	lf	13	ohf	23	heal	33	armory	43	*karazhan*
4	hc	14	lvl	24	sp	34	exp	44	*uldaman*
5	wts	15	sm	25	ub	35	stockade	45	*dmg*
6	lf1m	16	mechanar	26	lf3m	36	~~discordgg~~	46	*cz*
7	sv	17	summ	27	stratholme	37	deadmines	47	mt
8	wtb	18	botanica	28	sethekk	38	auchenai	48	sr
9	healer	19	bf	29	lfg	39	~~http~~	49	*gbid*
10	lf2m	20	bm	30	*gruul*	40	shh	50	*underbog*

Graph 5, shows the number of words in each of the categories, such as abbreviated, non-abbreviated, and a specific category called unclassifiable, which will be discussed later on in this chapter.

Graph 5: Distribution of the 50 strongest keywords

The analysis revealed that overall the keywords list is to a substantial degree similar to the previously analyzed wordlist. Despite the fact that this section does not deal with the comparison of wordlist and keywords, this information provides a context for the analysis. Similarly to the wordlist, the first 50 places on the keywords list exhibit a high frequency of occurrence of abbreviations. Moreover, there are only 16 non-abbreviated words, which is less than half of the abbreviations in this section of the list. Additionally, the first eight places on the list are occupied by abbreviations. This alone again suggests that the use of abbreviations is prevalent in the *WoWChatlog* corpus, possibly due to the highly specific and unorthodox nature of the context. It is worth noting that more or less the same words on the keywords list of *WoWChatlog* are also present on the wordlist. These words were discussed in *Chapter 7*, and therefore it is unnecessary to review them again.

However, some words were not present in the "WOWTOP100" list and are highlighted in orange in *Table 9*. These words are presumably not frequent enough to be in the WOWTOP100; on the other hand, they might be unique to this corpus to such a degree that their keyness is significantly stronger than other words in contrast with general English in *BNC* (reference corpus). There are four non-abbreviated words that have not been discussed earlier, all four of which denote a particular location, in this case, dungeon; therefore, they can be categorized as proper nouns and beyond the assumption that players tend to excessively use proper names in chat communication of World of Warcraft, they hold no substantial value for further analysis. Furthermore, there are four words not discussed in *Chapter 7*, namely the abbreviations *ffa*, *cz*, *dmg*, *gbid*. These abbreviations are highlighted in orange in *Table 9* and

were formed by initialism, final clipping, and the last two by omission, listed respectively. Among the 50 most prominent keywords, one intriguing item standing on rank 46 is a state code associated with the Czech Republic. It is worth noting that the presence of the Czech Republic's state code, among other strong keywords, raises curiosity. This may be attributed to a significant density of Czech native or Czech-speaking players or indicating a plausible connection between players from the Czech Republic in particular and their inclination to publicly exhibit or present their nationality. Lastly, merely two words on this list were assigned to the category of unclassifiable words. These words, namely *discordgg* and *http,* can be found crossed out in *Graph 5*. Discordgg and http (abbreviation for "Hypertext Transfer Protocol") are two parts of an internet link (http://discord.gg), which redirects the user to a web page or a program called Discord.gg, which is a speech-based communication software, in WoW, commonly utilized by the most "tryhard" players. Therefore, both HTTP and Discord.gg are ineligible for classification and thus have been excluded from the analysis.

9. N-grams analysis

This section of the bachelor's thesis delves into the n-gram analysis. The n-gram analysis helps researchers to understand the frequent sequences of words in the text, thereby identifying patterns that might not be immediately apparent (Andresen and Zinsmeister, 2017, 4). The analysis focuses on 2-, 3- and 4-grams, which will be examined individually. This approach aims to identify unusual peculiarities and, hopefully, patterns of World of Warcraft chat communication. Moreover, by identifying patterns in the text, inferences about the social dynamics of the World of Warcraft community may be conceivably drawn. For example, by examining frequent 2-grams, we may identify common and the most frequent two-word phrases used by players of WoW. Similarly, examining frequent 3-grams may provide insight into the structure and use of three-word phrases or abbreviations and their tendency to appear with various combinations of whole or shortened words in the chat. Further examination of 4-grams may enable the identification of more complex phrases in the social interactions of players, and thus it should provide insights into the communication dynamics of this particular game. By analyzing the n-grams, it will be possible to gain a deeper understanding of the way players communicate in-game and the manner they use language to achieve their goals. Ultimately, the insights gained from this analysis are expected to bring more clarity to the unique features of game-based communication and could be used to inform the development of more effective communication strategies in online gaming environments.

9.1.   2-grams

The first subchapter of n-grams analysis focuses exclusively on 2-grams, also known as bigrams. By examining pairs of consecutive words, common language patterns, dependencies between adjacent words, or, in this case, tendencies of abbreviations to co-occur together with other abbreviations or with non-abbreviated words might be uncovered. *Table 10* shows the 50 most frequent bigrams within the *WoWChatlog* corpus.

Table 10: 2G-TOP50

#	N-gram	#	N-gram	#	N-gram
1	last spot	18	lf1m tank	35	stockade boost
2	lf tank	19	lf dps	36	lf1m dps
3	w me	20	slave pens	37	need healer
4	need tank	21	tank heal	38	wts stratholme
5	lfm tank	22	heal and	39	healer for
6	dps for	23	per run	40	wts shadow
7	tank and	24	need heal	41	and heal
8	sm boost	25	need all	42	wtb sm
9	lf heal	26	tank last	43	boost cath
10	tank for	27	sp hc	44	hc lf
11	and dps	28	lf healer	45	lfm sl
12	heal for	29	stratholme boost	46	got summon
13	wts sm	30	shadow lab	47	heal dps
14	hc lfm	31	need dps	48	dps sv
15	me lfm	32	lfm heal	49	hc tank
16	dps lfm	33	normal lfm	50	healer and
17	sl hc	34	spot lfm		

Orange – Phrases containing only non-abbreviated words

Green – Phrases containing only abbreviations

The list of 50 most frequent 2-grams was named *2G-TOP50* for the purposes of this analysis. In the following subchapters of N-grams analysis (3-grams analysis and 4-grams analysis), there are two more tables containing lists named in the same manner (*Table 11* and *Table 12*). The words highlighted in orange are 2-grams consisting of solely non-abbreviated words. There are 19 instances of bigrams, including only non-abbreviated words in *2G-TOP50*. This indicates that pairs of co-occurring non-abbreviated words are considerably common in *WoWChatlog* regardless of the constant use of abbreviations. However, despite consisting of pairs of non-abbreviated words, understanding these phrases can prove to be a challenging task for a person not familiar with the context of World of Warcraft and its specific terminology. The difficulty lies in the number of words whose meaning in the game chat varies from the one

outside the game, or the words do not have a real-life interpretation. These include, for instance, (20) *slave pens*, (26) *tank last*, (29) *stratholme boost*, which are shown with context in examples D, E, and F extracted from the concordance feature of SketchEngine.

Example D: LFM Slave Pens - need 1 healer and 2 dps Be geared!

Example E: LFM Blackrock Depths 52-60lvl, need tank LAST SPOT

Example F: WTS Stratholme boost Run 10 min 45-60 lvl

In example D, the phrase *slave pens* is merely a name of an in-game dungeon, presumably connected with slavery and penitentiary content-wise. This phrase can be reasonably comprehensible, although one might not recognize it stands for a location in-game within the context of the message. Example E shows a phrase *tank last,* which includes two words relatively common in everyday speech. The word *last* functions in the same manner as in contemporary discourse. However, the word *tank* does not carry the same meaning. According to the Cambridge Dictionary[3] can be a noun representing a gas or liquid container or a military vehicle. *Tank* can also be a verb signifying a number of actions, also depicted on the Cambridge Dictionary website. In World of Warcraft chat, this word cannot be identified as any of the previously mentioned definitions. Instead, the word *tank* is used as a designation for durable classes, as described in *Chapter 5* of the theoretical part. Example F demonstrates the use of the phrase *stratholme boost*. The word stratholme does not carry any meaning or connotation in contemporary English. Thus, it can be deduced that this word was invented during the course of producing this game. These three examples highlight the importance of being familiar with the game terminology to a significant extent in order to comprehend short phrases, even without the use of abbreviations.

Further, the *2G-TOP50* list contains 11 phrases or 2-grams consisting of two abbreviations which are highlighted in green in *Table 10*. Understanding these bigrams without further exposure to the communication and basics of WoW is virtually impossible. On the other hand, while being the most challenging to comprehend, it has the potential to compress the message the most and still convey the original meaning. Examples G, H, and I serve as illustrations demonstrating the compression and length differences between non-abbreviated bigrams and their shortened counterparts. These examples allow the reader to observe how the use of abbreviations can effectively save time and condense the message. The letter counts for

---

[3] https://dictionary.cambridge.org/dictionary/english/tank

each example, displayed below the corresponding bigrams, serve to highlight the noticeable differences in length between the bigrams with abbreviations and their protracted counterparts.

Example G: wts sm – want to sell scarlet monastery

         5   x   30

Example H: lfm sl – looking for more shadow labyrinth

         5   x   33

Example I: dps sv – damage per second Steamvaults

         5   x   29

Examples G, H and I, all identified as *Initialisms,* display relatively similar compression ratios of approximately 1:6 letters. This suggests that short combinations of 2 *Initialisms* in World of Warcraft are capable of compressing the original meaning of the message in up to six times shorter message and thus making the writing and reading time six times faster, which is an important aspect for players of WoW. This phenomenon might be one of the motives of players to implement abbreviations in their messages.

Lastly, the most common group of 2-grams in the *2G-TOP50* are 20 phrases that utilize a combination of an abbreviated word and one non-abbreviated word. The position of abbreviated and non-abbreviated words within the 2-gram phrases is relatively evenly distributed, with a slight predominance of bigrams initiating with the abbreviation. Specifically, among the examined instances, there were 12 occurrences where the abbreviation is positioned at the beginning of the 2-gram, whereas eight instances where phrases are concluded with the abbreviation.

9.2.   3-grams

The second subchapter of the N-gram analysis aims to investigate the most frequent 3-grams, also called trigrams, of *WoWChatlog* which are displayed in *Table 11* below. The analysis of 3-grams offers valuable insights into the structure and relationships within three-word phrases for this thesis, encompassing the interplay between constituent words, the presence of abbreviations, and patterns observed in their usage.

Table 11: 3G-TOP50

#	N-gram	#	N-gram	#	N-gram
1	wts sm boost	18	lfm auchenai crypts	35	sethekk halls hc
2	w me lfm	19	wts stockade boost	36	lf heal for
3	tank last spot	20	last spot lf	37	dps last spot
4	wts stratholme boost	21	cipher of damnation	38	the cipher of
5	last spot lfm	22	auchenai crypts hc	39	lfm sethekk halls
6	sm boost cath	23	lfm zg full	40	wtb stockade boost
7	shadow lab boost	24	need tank last	41	zg full run
8	lf tank for	25	and heal for	42	for the shattered
9	lf tank and	26	lf1m tank for	43	lfm tank and
10	wts shadow lab	27	mobs sp boost	44	last spot for
11	wtb sm boost	28	sp boost optimal	45	for more info
12	heal and dps	29	boost optimal for	46	lf1m tank sl
13	tank and heal	30	tank and healer	47	me for more
14	tank sl hc	31	lf tank heal	48	need tank heal
15	tank and dps	32	healer and dps	49	and dps lfm
16	need tank and	33	sl hc lfm	50	ring of blood
17	w me lf	34	tank heal dps		

Orange – Phrases containing only non-abbreviated words

Green – Phrases containing only abbreviations

Listed trigrams have been assigned to their corresponding category according to the structure of the phrase. Phrases are highlighted in the same manner containing only non-abbreviated words are again highlighted in orange, and phrases with only abbreviations are green. 3-grams containing both abbreviated and non–abbreviated words are plain, without any highlighting. Within the 3G-TOP50, consisting of the 50 most frequently occurring 3-grams in the *WoWChatlog* corpus, a total of 15 trigrams were identified without any abbreviations. This subset of trigrams exhibits a slightly lower frequency compared to the corresponding subset in the 2G-TOP50, although the difference is not statistically significant. Notably, of these 15 trigrams, 12 incorporate at least one grammatical word. This finding is intriguing, considering the discussion on the substantially reduced usage of grammatical words in *WoWChatlog*

compared to everyday speech in *Chapter 7.2*. Surprisingly, 80% of the 50 most frequent trigrams containing only non-abbreviated words encompass at least one grammatical word. This observation implies that a certain level of the grammatical structure of *WoWChatlog* messages may persist, potentially contributing to enhanced comprehensibility even after significant compression of various sentence elements to facilitate brevity in messaging.

In 3G-TOP50, trigrams composed of solely abbreviations are represented only by a single occurrence, namely *sl hc lfm*. This trigram of seven letters means: "shadow labyrinth heroic looking for more." The non-abbreviated version of this trigram contains 40 letters, again proving the compression capacity of abbreviations of up to six times the original size. However, the occurrence of only one instance of trigrams solely composed of abbreviations within 3GTOP50 indicates that despite the prevalent use of abbreviations in *WoWChatlog*, there is a tendency to incorporate grammatical or lexical words alongside abbreviations in messages comprising three or more words. This pattern likely arises from the need to maintain comprehensibility, as employing solely abbreviations in such messages would likely result in incomprehensible communication.

Conversely, a much more common group of trigrams is one composed of a combination of abbreviated and non-abbreviated words. Namely, 68% of trigrams in 3G-TOP50 comprise at least one abbreviation and one non-abbreviated word. The cause of this phenomenon can be attributed to the potential of this type of trigram to achieve significant message compression while maintaining a satisfactory level of comprehensibility due to drawing benefits from both abbreviations and non-abbreviated words. The word structure of the three-word phrases in 3G-TOP50 varies to such an extent that the identification of their usage pattern is not plausible. 14 of these 34 trigrams contain an abbreviation as the initial word of the phrase. Additionally, eight occurrences of abbreviations in the concluding position of trigrams are observed, and 12 trigrams contain two abbreviations and only one non-abbreviated word. Despite the inability of the analysis to identify any distinguishable patterns of abbreviation use within 3-grams in *WoWChatlog*, the examination of word structure in 3-word phrases within World of Warcraft chat or any game chat presents an intriguing avenue for future research. This line of inquiry has the potential to offer valuable insights and a more profound understanding of the intricate landscape of gaming abbreviations.

9.3.   4-grams

Lastly, in the N-grams analysis chapter, the most frequent 4-word phrases are examined to reveal patterns or possibly rules for their formation. Furthermore, the most frequent 4-grams are assigned to a specific group according to the structure of the phrase from the perspective of the presence or absence of an abbreviation. These groups can be found highlighted in various colors in *Table 12* below.

Table 12: 4G-TOP50

#	N-gram	#	N-gram	#	N-gram
1	wts sm boost cath	18	tank for bm norm	35	lf tank for bm
2	wts shadow lab boost	19	hc tank last spot	36	the shattered n summ
3	lfm auchenai crypts hc	20	lf1m ohf normal heal	37	lf heal and dps
4	need tank last spot	21	lf ssh rep run	38	hc shattered halls need
5	sp boost optimal for	22	ssh rep run need	39	lf tank sp hc
6	mobs sp boost optimal	23	deathblow to the legion	40	lfm tank and heal
7	the cipher of damnation	24	rep run need tank	41	reset free for all
8	tank and heal for	25	lfm daily hc shattered	42	lfm zg full clear
9	lf1m tank sl hc	26	daily hc shattered halls	43	need healer and dps
10	wts stratholme boost run	27	sethekk halls hc tank	44	dps for the shattered
11	lfm sethekk halls hc	28	tank last spot lfm	45	halls hc need tank
12	lfm tank sl hc	29	lf tank and dps	46	sethekk halls hc need
13	lfm zg full run	30	zg full run need	47	~~sign up at https~~
14	lf heal ub hero	31	me for more info	48	need tank and healer
15	w me for more	32	run need tank heal	49	tank for arca hc
16	lf tank and heal	33	dps sv n farm	50	lfm shh normal need
17	tank sl hc lfm	34	for the shattered n		

Orange – Phrases containing only non-abbreviated words

Green – Phrases containing only one abbreviation

The predominant category of 4-grams found in this list comprises phrases that incorporate two or more abbreviations. Remarkably, these particular 4-grams constitute 50% of the entire 4G-TOP50 list. Further analysis reveals that nearly three-quarters (18) of these phrases are composed of two abbreviations along with two non-abbreviated words, while the remaining seven instances contain three abbreviations and a single non-abbreviated word. In the previous *Chapter 9.2*, it was discussed that phrases consisting solely of abbreviations are relatively infrequent in *WoWChatlog*. However, when an additional non-abbreviated word is introduced into the phrase, its occurrence becomes substantially more frequent in *WoWChatlog* messages. This finding again suggests that the inclusion of non-abbreviated words is crucial for

enhancing the comprehensibility of longer phrases within the chat discourse of World of Warcraft.

The group of 4-grams that follows as the second most frequent category includes phrases containing only a single abbreviation. Within the 4G-TOP50, a total of 16 such phrases are present. Among these, ten phrases initiate with the abbreviation, while three phrases conclude with the abbreviation. Additionally, three phrases incorporate an abbreviation within the middle portion of the 4-gram. This observation presumably suggests that abbreviations are more likely to appear at the beginning rather than at the end of multi-word phrases.

Lastly, in *Table 12,* there are eight 4-grams, some of which rank significantly high on the list, despite their relatively smaller number of instances. This finding suggests that compounding four non-abbreviated words is considerably common within the World of Warcraft chat. Conversely, none of the 4-grams present in this list consists solely of abbreviations. Similar to the findings in the previous *Chapter 9.2.*, the low frequency or complete absence of phrases composed of only abbreviations in both the 3G- and 4G-TOP50 indicates that players tend to incorporate at least one non-abbreviated word in both 4-word and 3-word phrases in their chat messages. In fact, the most frequent 4-gram containing only abbreviated words is *lfm dps sv n,* and it is on the 100th rank of most frequent 4-grams of *WoWChatlog*.

Conclusion

The primary aim of this Bachelor's thesis was to conduct a corpus-based analysis of *WoWChatlog*, a specially curated corpus consisting of in-game chat messages from World of Warcraft players, and to explore the use of abbreviations in such discourse. The analysis encompassed three key components: Wordlist analysis, Keywords analysis, and N-grams analysis. These analytical processes were facilitated through the utilization of the SketchEngine corpus analysis tool.

However, before drawing definitive conclusions from the findings presented in this thesis, it is crucial to note that more extensive and thorough research is required to establish concrete evidence about the language patterns, abbreviations, and customs for their use in World of Warcraft chat messages. Furthermore, while SketchEngine proves to be a valuable tool for corpus analysis, it is not ideally suited for handling the substantial volume of abbreviations present in the *WoWChatlog* corpus. None of the current Corpus Query Systems (CQS) possesses the capability to address this issue effectively. Consequently, certain functionalities of SketchEngine, such as Parts of Speech recognition, were not optimally operational. It is conceivable that the insights gained from this thesis may inspire the development of a new corpus analysis tool specifically designed to handle both abbreviated and non-abbreviated words proficiently.

The first research question aims to designate the differences between World of Warcraft chat communication and regular everyday speech represented by The *British National Corpus*. It was observed that the frequency of content words in *WoWChatlog* remains relatively similar to regular everyday speech, despite the choice of lexical words being completely different. However, it is important to acknowledge that some content words (lexical words) in *WoWChatlog* may not align with their original dictionary definitions but instead hold in-game mechanics or object-related significance. In addition, a higher occurrence of nouns and especially proper nouns in World of Warcraft chat messages was detected. Furthermore, players tend to omit some grammatically important components of sentences, presumably to make room for other sentence elements which carry the meaning of the message or abbreviations. On the contrary to lexical words, the choice of grammatical words is comparable to everyday English, regardless of the lower frequency of use. It can be assumed that the lower frequency of grammatical words in World of Warcraft chat is compensated by the presence of abbreviations.

Research question number two and three ask to what extent players of WoW use abbreviations in their message and how effective abbreviations are in reducing the number of letters in the original message. The *WoWChatlog* corpus exhibited a substantial occurrence of abbreviations, highlighting players' inclination to compress message content. Out of the 100 most frequent words, 41% were found to be abbreviations proving the significance of abbreviations and the tendency of players to use them in messages. Among the various types of abbreviation formation, initialism was identified as the most frequently employed, presumably due to its potential for message compression and because initialisms preserve only a single letter from the original non-abbreviated word, enabling the greatest degree of letter-count reduction among all formation styles of abbreviations. Furthermore, the analysis revealed that abbreviations have the capability to reduce the letter count of the original message by up to six times.

The fourth research question aims to find out whether or not can communication in World of Warcraft take place solely using abbreviations. The analysis of N-grams in *WoWChatlog* exposed the most prevalent type of phrases to be a combination of at least one non-abbreviated word and at least one abbreviation. While players frequently combine two abbreviations in two-word phrases, it is considerably uncommon to encounter longer phrases solely comprised of abbreviations. This observation indicates that players strive to preserve comprehensibility rather than inconsiderate reduction of letter counts to the highest possible degree, proving that communication based solely on abbreviations cannot take place in World of Warcraft chat. Conversely, longer phrases containing only non-abbreviated words were found to be noticeably common in *WoWChatlog*, signifying the importance of non-abbreviated words in WoW chat communication, despite the prevalent use of abbreviations. Additionally, in trigrams and 4-grams that combine the use of both abbreviations and non-abbreviated words, abbreviations tend to appear at the beginning of the phrase rather than at the end.

The examination of keywords in the corpus supported the findings obtained in wordlist analysis, additionally suggesting a possible higher concentration of Czech players or a greater willingness of Czech players to express their nationality in Czech compared to other nationalities because of the presence of Czech state code in the keywords list.

The analysis results validated the initial assumptions underlying this study. It was assumed that World of Warcraft players would omit certain sentence parts to highlight the essential components and that the implementation of abbreviations would serve to compress and condense messages to a point while preserving the meaning of the message. Both of these

assumptions were confirmed. This analysis could be thereby considered successful. However, it is essential to recognize that this study primarily focused on the chat channels "LookingForGroup" and "World," where players primarily advertise their groups or activities. Consequently, this work could serve as a foundation for future profound analyses encompassing additional aspects and channels of *WoWChatlog*, thereby unraveling further information and promoting an enhanced understanding of World of Warcraft chat communication.

List of references

1. Adolphs, Svenja. 2006. *Introducing Electronic Text Analysis: A Practical Guide for Language and Literary Studies* (1st ed.). Routledge.

2. Andresen, Melanie and Zinsmeister, Heike. 2017. "The Benefit of Syntactic vs. Linear N-grams for Linguistic Description." *In Proceedings of the Fourth International Conference on Dependency Linguistics* (Depling 2017). Pages 4–14. Linköping University Electronic Press. Available at: https://aclanthology.org/W17-6503.pdf.

3. Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London / New York: Continuum.

4. Bennett, Gena R. 2010. *Using Corpora in the Language Learning Classroom: Corpus Linguistics for Teachers.* Ann Arbor, MI: University of Michigan Press.

5. Blizzard. 2022. "World of Warcraft." Available at:
   http://www.worldofwarcraft.com.

6. Brezina, Václav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

7. Brezina, Vaclav and Hawtin, Abi and McEnery, Tony. 2021. "The Written British National Corpus 2014 – design and comparability." Text & Talk 41, no. 5-6. Available at: https://doi.org/doi:10.1515/text-2020-0052.

8. The Editors of Encyclopaedia. 2016. "World of Warcraft". Encyclopaedia Britannica. Available at: https://www.britannica.com/topic/World-of-Warcraft.

9. Corver, Norbert and van Riemsdijk, Henk. 2001. *Semi-lexical categories: the function of content words and the content of function words.* Berlin: Walter de Gruyter

10. Crystal, David. 2006. *Language and the Internet.* 2nd ed. Cambridge: Cambridge University Press.

11. Crystal, David. 2008. *Txtng: The Gr8 Db8*. Oxford: Oxford University Press

12. Čermák, František. 2017. *Korpus a korpusová lingvistika.* Praha: Karlova Univerzita.

13. Dang, Thi Ngoc Yen. 2019. *Corpus-Based Word Lists in Second Language Vocabulary Research, Learning, and Teaching*. Routledge.

14. Cruz-Pavía, Irene De la and Marino, Caterina and Gervain, Judit. 2021. "Learning word order: early beginnings." *Trends in Cognitive Sciences*. Volume 25, Issue 9, Pages 802-812. Available at: https://doi.org/10.1016/j.tics.2021.04.011

15. Derks, Daantje and Bos, Arjan E. R. and von Grumbkow, Jasper. 2007. "Emoticons and social interaction on the Internet: the importance of soci al context." *Computers in Human Behavior.* Volume 23, Issue 1, Pages 842–849. Available at: https://doi.org/10.1016/j.chb.2004.11.013.

16. Dictionary, 2023. Cambridge English Dictionary. Available at:
   https://dictionary.cambridge.org/dictionary/english.

17. Gablasova, Dana and Brezina, Vaclav and McEnery, Tony. 2017. "Collocations in Corpus-Based Language Learning Research: Identifying, Comparing, and Interpreting the Evidence." *Language Learning*. Volume 67, Issue 1, Pages 155-179. Available at: https://doi.org/10.1111/lang.12225

18. Gentile, Douglas. 2009. "Pathological video-game use among youth ages 8-18: A national study." *Psychological Science*. Volume 20, Issue 5, Pages 594–602. Available at: https://doi.org/10.1111/j.1467-9280.2009.02340.x.

19. Hadziahmetovic Jurida, Sanel and Dzanic, Mirza and Pavlovic, Tanja and Jahić, Alma and Hanic, Jasmina. 2016. "Netspeak: Linguistic Properties and Aspects of Online Communication in Postponed Time." *Journal of Foreign Language Teaching and Applied Linguistics.* Volume 3. Available at: https://doi.org/10.14706/JFLTAL163115.

20. Hård af Segerstad, Ylva. 2003. "Use and adaptation of written language to the conditions of computer-mediated communication." PhD dissertation. Göteborg University. Department of Linguistics. Available at: https://nl.ijs.si/janes/wp-content/uploads/2014/09/segerstad02.pdf.

21. Hrbáček, Josef. 1979. *Jazykové zkratky v češtině*. 1st ed. Praha: Karlova Univerzita.

22. Jurafsky, Dan and Martin, James H. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 1st ed. Prentice Hall.

23. Kameníková, Lucie. 2013. "Jazyk chatu (lexikální analýza)." Master's thesis. Palacký University Olomouc. Philosophical Faculty. Available at: https://theses.cz/id/s2xabi/00179097-742609960.pdf?lang=en.

24. Kerr, Elaine B. and Hiltz, Starr Roxanne. 2013. *Computer-mediated communication systems: Status and evaluation*. Academic Press.

25. Kilgarriff, Adam and Iztok Kosem. 2012. "Corpus Tools for Lexicographers." *Electronic Lexicography*. Available at: https://doi.org/10.1093/acprof:oso/9780199654864.003.0003.

26. Kilgarriff, Adam and Rychlý, Pavel and Smrz and Pavel and Tugwell, David. 2004. "The Sketch Engine." *Computational Lexicography and Lexicology*. Available at: https://euralex.org/publications/the-sketch-engine/.

27. Krejtz, Izabela and Szarkowska, Agnieszka and Łogińska, Maria. 2016. "Reading Function and Content Words in Subtitled Videos" *The Journal of Deaf Studies and Deaf Education*. Volume 21, Issue 2, Pages 222–232. Available at: https://doi.org/10.1093/deafed/env061.

28. Lindh, Simon. 2009. "Online computer game English: A study on the language found in World of Warcraft." Bachelor's thesis. Karlstad University. Faculty of Arts and Education. Available at: http://www.diva-portal.org/smash/get/diva2:221885/FULLTEXT02.pdf.

29. Manning, Christopher D. and Schütze, Hinrich. 1999. *Foundations of Statistical Natural Language Processing.* Cambridge: MIT Press

30. McEnery, Tony and Hardie, Andrew. 2012. *Corpus linguistics: method, theory and practice.* Cambridge: Cambridge University Press.

31. McEnery, Tony and Wilson, Andrew. 2001. *Corpus Linguistics: An Introduction.* Edinburgh University Press.

32. Nardi, Bonnie A. and Ly, Stella and Harris, Justin. 2007. "Learning Conversations in World of Warcraft." 2007. *40th Annual Hawaii International Conference on System Sciences.* Available at: http://hdl.handle.net/10524/1601.

33. Plag, Ingo. 2003. *Word-formation in English.* Cambridge Textbooks in Linguistics. Cambridge: Cambridge University Press.

34. Ray, Michael. 2022. *Activision Blizzard, Inc.*, Encyclopaedia Britannica. Available at: https://www.britannica.com/topic/Activision-Blizzard-Inc.

35. Rusaw, Erin. 2011. *Language and Social Interaction in the virtual space of World of Warcraft.* University of Illinois.

36. Sardinha, Tony Berber. 2000. "Comparing corpora with WordSmith Tools: How large must the reference corpus be?." *The Workshop on Comparing Corpora.* Hong Kong, China. Association for Computational Linguistics. Available at: https://doi.org/10.3115/1117729.1117731

37. Schmitt, Norbert. 2010. *An introduction to applied linguistics.* London: Routledge.

38. SketchEngine software. Available at: *https://www.sketchengine.eu/*

39. Statista. 2023. *Video Games – Europe.* Available at: https://www.statista.com/outlook/dmo/digital-media/video-games/europe.

40. Suznjevic, Mirko and Matijasevic, Maja. 2010. "Why MMORPG players do what they do: Relating motivations to action categories." *International Journal of Advanced Media and Communication.* Volume 4, Issue 4, Pages 405-424. Available at: https://doi.org/10.1504/IJAMC.2010.036838.

41. Szudarski, Pawel. 2018. *Corpus linguistics for vocabulary: A guide for research.* London and New York: Routledge.

42. Wowhead. 2021. *Wowhead Beginner Guides: How to Play World of Warcraft.* Available at: https://www.wowhead.com/guide/wowhead-beginner-guides-how-to-play-world-of-warcraft-5331

43. Wozniak, Jakub. 2017. "Analysis of Corpus Query Tools." Bachelor's thesis. Masaryk University in Brno, Faculty of Philosophy. Available at: https://theses.cz/id/5ol302.

44. YouGov. 2018. *Frequency of using group chat function of text messaging (SMS) or online messaging applications according to adults in the United States as of December 2017, by age group.* Statista. Available at: https://www.statista.com/statistics/800650/group-chat-functions-age-use-text-online-messaging-apps/.

45. Zerkina, Natalia and Kostina, Nikolaevna and Pitina, Svetlana Anatol'evna. 2015. "Abbreviation Semantics." *Procedia - Social and Behavioral Sciences.* Volume 199, Pages 137-142. Available at: https://doi.org/10.1016/j.sbspro.2015.07.497.