LATVIJAS UNIVERSITĀTE

BAKALAURA DARBS

RĪGA 2018

UNIVERSITY OF LATVIA

FACULTY OF HUMANITIES

DEPARTMENT OF ENGLISH STUDIES

PARTS OF SPEECH IN THE MESSAGES OF THE MICROBLOG TWITTER.COM

VĀRDU ŠĶIRU LIETOJUMS MIKROBLOGA VIETNES TWITTER.COM ZIŅOJUMOS

BACHELOR THESIS

Elza Upmane

Matriculation card No. eu14008

Adviser: assoc.prof. Zigrīda Vinčela

RIGA 2018

Anotācija

Bakalaura darbs ir veltīts vārdšķiru absolūtā un relatīvā biežuma variācijām specializētajā tvītu korpusā *VerUn*, kas tika izveidots šī darba ietvaros. Bakalaura darba mērķis ir izpētīt lietvārdu un darbības vārdu lietojuma variācijas, kā arī absolūto un relatīvo biežumu politiskiem tematiem veltīto tvītu korpusa *VerUn* divos apakškorpusos: apstiprināto kontu tvītu apakškorpusā *Ver* un neapstiprināto kontu tvītu apakškorpusā *Un*. Lai sasniegtu mērķi, bakalaura darbā tika izmantota minētā korpusa kvantitatīvā analīze, izmantojot programmas CLAWS un *AntConc*. Rezultāti liecina, ka lietvārdu lietojums ir biežāks gan apstiprinātajos, gan neatpstiprinātajos kontos (63% un 55%), taču neapstiprinātajos kontos lietvārdu biežums ir vismaz 2 reizes lielāks un darbības vārdu – 3 reizes. Lielākās atšķirības tika pamanītas visbiežāk lietotajos īpašvārdos.

Atslēgvārdi: lietvārdi, darbības vārdi, korpusa analīze, vārdšķiru absolūtais un relatīvais biežums, Tvitera ieraksti, vārdšķiru marķēšana

Abstract

The present paper is devoted to the part of speech absolute and relative frequency variations in the specialised tweet corpus *VerUn* that has been compiled for the specific purposes of the present research. The goal of the paper is to investigate the noun and verb variations, as well as relative and absolute frequencies in the corpus *VerUn*, which is devoted to political themes and consists of two sub-corpora: verified account tweet sub-corpus *Ver* and unverified account tweet sub-corpus *Un*. In order to reach the goal, the method of the mentioned corpus quantitative analysis was employed, and the part-of-speech tagger CLAWS was used, as well as the corpus analysis toolkit *AntConc*. The results show that nouns are predominant in both verified and unverified Twitter tweet sub-corpora (63% and 55%), but the frequency of nouns is at least 2 times higher and that of verbs - 3 times higher in unverified account sub-corpus. The largest differences can be seen in the most frequent proper nouns used.

Key words: nouns, verbs, corpus analysis, part of speech absolute and relative frequency, Twitter posts, part of speech tagging

Table of Contents

Introduction

Along with the modern 21st-century advancements in technology, the communication has also become more modernised with various Internet sites aiming to make it more accessible, faster and dynamic. Since the speed of sending and receiving information has improved drastically, the messages, too, have become more compact and concise. This unlocks the possibility for linguistic studies in the communicative context of social media, one of them being the social networking and microblogging site Twitter.

The emergence of Twitter has piqued the interest of linguists, but because it is so recent (Twitter was founded in March of 2006 (Online 1), it is still underinvestigated. There are several approaches of conducting linguistic research of Twitter tweet data and it might be arranged into four trends according to what linguistic characteristics of tweets is in the focus of attention of researches: geolocation and language diversity in different locations, sentiment analysis, research of a mini corpus and multidimensional analysis of Twitter tweets. Researchers Zhao & Chao, 2017; Graham, Hale & Gaffney, 2014 use the function of geolocation to study language diversity (e.g. Chinese, English etc.) in different parts of Hong Kong (Zhao & Chao, 2017), as well as use the geolocation to study the language diversity in four areas – Cairo, Montreal, San Diego and Tokyo (Graham, Hale and Gaffney, 2014).

Sentiment analysis or opinion mining is the analysis of adjectives in tweet texts and is done by e.g. Öztürk, Ayvaz (2018), who used Twitter to conduct sentiment analysis in connection with Syrian refugee crisis, extracting those adjectives that convey a particular emotion. Wang and Fikis (2017) used the sentiment analysis to extract public's opinion regarding the Common Core State Standards and Ceron et al. (2015) employed sentiment analysis to monitor electoral campaigns and people's voting intentions.

David Crystal devoted a chapter of his book 'Internet Linguistics' (2011) to Twitter by compiling a mini-corpus of tweet texts and conducting a qualitative analysis. He predominantly analysed the functionality of Twitter, but did not focus on parts of speech in Twitter tweets.

The corpus-based, multidimensional analysis of web texts that also include the data from microblogging site Twitter is conducted by Titak and Roberson (2013), who investigated the variation of register in online web texts and concluded that there are four functional dimensions of online language: personal narrative focus versus descriptive, informational production; involved, interactive discourse; complex statement of opinion and past versus present orientation. Therefore, multidimensional analysis is obviously an

insightflul and detailed approach when analyzing Twitter tweet texts because it deals with the linguistic features and their co-occurrence in the texts.

Although Twitter has been in the focus of linguists because it is a versatile microblogging site that can be used for various research purposes, the research regarding the linguistic aspect of Twitter tweet texts is considerably underinvestigated and needs to be broadened.

Therefore, **the goal of the research** is to investigate the variations of nouns and verbs in tweets of political themes in the verified, as well as unverified Twitter accounts.

The research questions are as follows: What part of speech (noun or verb) is more prominent in the Twitter tweets of verified accounts that are devoted to political themes? What part of speech (noun or verb) is more prominent in the Twitter tweets of unverified accounts that are devoted to political themes?

The enabling objectives include:

1. to explore theories regarding parts of speech and their grammatical functions
2. to explore theories on corpora types
3. to explore the previous research devoted to various trends of linguistic research of Twitter tweet data
4. to explore the features of microblogging as a messaging system
5. to create a corpus of verified and Twitter account tweets devoted to political themes and issues
6. to conduct an analysis of the aforementioned corpus and sub-corpora of selected tweets
7. to draw relevant conclusions

The theoretical research methods include analysis of theories regarding part of speech division, previous research regarding different ways of Twitter tweet text analysis and analysis of the features of microblogging as a messaging system

The empirical research methods include the creation of specialised, annotated corpus (*VerUn*) of Twitter account tweets devoted to political themes, analysis of different corpus types and corpus creation process, as well as quantitative analysis of the occurrence of parts of speech in both sub-corpora: 1) the sub-corpora of tweets devoted to political themes in verified Twitter accounts (*Ver)* 2) the sub-corpora of devoted to political themes in unverified Twitter accounts (*Un*), where the hashtag of political theme has been used, for instance, #political, #whitehouse #elections, #potus etc.

Chapter 1
Parts of Speech in the English Language

This chapter of the bachelor thesis deals with the theories on the classification and functions of the parts of speech (Huddleston, Pullum, 2002; Downing and Locke, 2006) or word classes (Biber et al., 1993; Quirk et al., 1985) in the English language, as well as theory regarding their functions.

Matilal (1990) explains that the first attempts to classify parts of speech dated around 6th-5th century BC (Matilal, 1990: 7) with Sanskrit grammarian and philologists Pāṇini and Yāska. Yāska, who preceded Pāṇini, put forth four different word classes (Matilal, 1990:18). in his work *Nirukta – nāma* (noun), *ākhyāta* (verb), *upasarga* (prefix) and *nipāta* (particles). However, Pāṇini (Pāṇini, 1986: 173) divided all words into two classes – *subantas* and *tinantas*, which are, respectively, nouns and verbs.

The development of linguistics in Sanskrit is compared with the development of linguistics in the West by Matilal (1990), and he concludes that linguistics in the West developed rather late (Matilal, 1990:7). The earliest discussion regarding linguistics and philosophy in the West was few centuries after Yāskas *Nirukta* - approximately around 4th-3rd century BC – found in Plato's middle period dialogue *Cratylus* (Sedley, 2007: 162), but only in Plato's late period dialogue *Sophist* (ibid.) the idea of successful statement (*logos)* was polished - it was maintained that a minimal sentence consists of two parts – *onoma*, which is the name of the object, and *rhēma*, which is the thing said of it. Sedley (ibid.) concludes that the differentiation between *onoma* and *rhēma* would later become known as the distinction between the noun and verb.

In the 2nd century BC Dionysus Thrax (Dionysius, Bécares Botas and Dionysius, 2002: 22) maintained that grammatical analysis happens on three levels: letters, syllables and words and their classes. He split the words into eight groups (i.e. word classes): noun, verb, participle, article, pronoun, preposition, adverb and conjunction. Even if the division was not as accurate as proposed by the contemporary linguists, this was the first attempt to classify the word classes into broader, more distinct groups.

Upon observing the comparison of the two above discussed approaches in the classification of the world classes, two conclusions can be made. First, verbs and nouns have been the two central components of part of speech division even in the first ancient linguistic traditions. Second - Sanskrit and Indian ancient linguistic traditions began the discussion on how the parts of speech could be classified earlier than they did in the West and Ancient Greece.

The contemporary classification of parts of speech is similar to that previously proposed by Dionysus Thrax; according to Downing and Locke (Downing and Locke, 2006: 16), 'words are classified grammatically according to the traditional terminology, which includes noun, verb, adjective, adverb, preposition, pronoun, article and conjunction'. Downing and Locke (ibid.) and Biber et al. (1995:56) further divide the parts of speech into two main classes: open (noun, verb, adjective, adverb), i.e., those ' that freely admit new members into the vocabulary' (Downing and Locke, 2006: 16) and closed (preposition, pronoun and article), i.e., those that 'do not easily admit new members' (ibid.) The grammarians mentioned add (ibid., 57), however, that this distinction between open and closed classes is not etched in stone – it is possible for new prepositions to develop out of verb forms (for instance, the preposition *regarding*), but the process is lengthy, sometimes even taking centuries.

Researchers Biber et al., (1999: 55) and Huddleston and Pullum (2002: 16) maintain that the contemporary division of parts of speech can be further grouped into three main groups: lexical words, function words, and inserts or interjections. However, Huddleston and Pullum (ibid.) mention that interjections are a minor category, with examples like *oh, hello, ouch* etc., about 'which there isn't anything interesting to say'.

The aforementioned eight parts of speech (excluding the interjections) are discussed in the following sub-chapter due to the fact that it would be important to lay down the exact classification, functions and differences of parts of speech, necessary for further research. As nouns and verbs are connected with the empirical part of the present bachelor's thesis, they will be looked at on a deeper level.

1.1 Content words

Nouns

Nouns are the core part of speech of English and any language, therefore, grammarians like Downing and Locke (2006: 401) and Huddleston and Pullum (2002: 17) maintain that nouns are the core root of all the parts of speech, and refer to classes of entities that have experiential features, for instance, description, classification or identification, or denote all kinds of physical objects (classes of entities), for instance, persons, animals, inanimate objects, abstract ideas, phenomena and emotions.

Researchers (Huddleston and Pullum, 2002: 17, Downing and Locke 2006: 401, Hudson, 2010: 264) acknowledge three types of nouns: common nouns, proper nouns and pronouns. However, Quirk et al. (1985: 245) divide nouns into count, non-count and proper nouns and add numerals and interjections (ibid., 67, 73) to general classification of function words. Biber et al. (Biber et al., 1999: 62-63) maintain that pronouns should be divided as a

separate part of speech under function words, therefore, it can be concluded that there are different ways that noun types can be classified, and they vary from researcher to researcher.

Although Huddleston and Pullum (2002: 17) have put pronouns together with common and proper nouns in their classification, stating that 'in traditional grammar the pronoun is treated as a distinct part of speech', they argue that it would ignore the fact that there is, in fact, a syntactic similarity between common, proper nouns and pronouns – they occur as heads of noun phrases.

Common nouns are determined by their countability (Downing and Locke, 2006: 405), i.e., there are singular and plural common nouns, with the latter formed with the help of suffix /iz/ after a sibilant (cross- crosses), /s/ after a voiceless consonant (tick-ticks) or /z/ after a voiced consonant (eye-eyes). There also exist irregular plurals (ibid.) (cactus – cacti, man-men), as well as zero plurals, which retain their form even in the plural (sheep, trout).

Researchers Quirk et al. (1985: 246), Downing and Locke (2006: 405-406) and Huddleston and Pullum (2002: 85-86) note that there is a distinction between count and non-count nouns, which differentiate in a way that count nouns can take cardinal number (one flower, two apples), but non-count nouns cannot be counted (furniture, luggage). Quirk et al. (1985: 247) extend this notion by adding that there are concrete and abstract nouns, with concrete nouns describing nouns that are accessible to senses, observable and measurable, and abstract nouns being those that cannot be observed and measured.

Downing and Locke (2006: 410) put forth the differentiation between proper nouns and proper names – proper nouns (Anna, Latvia) 'have no definable meaning in the language' because there is no way one could define any features or entities by the name of *Anna*, whereas with common nouns like *apple* or *pear*, it is possible to do so. Proper names, however, have a more complex structure (ibid.) due to the fact that could include a proper noun, for instance, *University of Latvia* or *Pauls Stradins Clinical University Hospital.*

Grammarians Huddleston and Pullum (2002: 82-84), as well as Quirk et al. (1985: 256), maintain that a noun can form a noun phrase (a phrase of which a noun is a head of). They can range from simple to more complicated ones (Quirk et al, 1985: 356):

- The girl is my sister (definite article + noun head)
- The *blonde* girl is my sister (premodifying adjective)
- The blonde girl *in blue jeans* is my sister (prepositional phrase)
- The blonde girl *wearing blue jeans* is my sister (non-finite clause)
- The blonde girl *who is wearing blue jeans* is my sister (relative clause)
- *She* is my sister (noun phrase consists of one word, personal pronoun *she*)

In the noun phrase, the noun can act as a:

- subject (That dog is black),

- object (My mother will make us breakfast),
- complement of clauses (The old lady sitting on the porch),
- complement of prepositional phrases (My dress is in the *closet*).

Grammarians Biber et al. (1999: 292-311) and Huddleston and Pullum (2002: 67) mention that nouns also have the category of case – a formal category that defines the relation of a noun to other units. Biber et al. (1999: 292) state that although in Old English nouns were marked by four different cases, only one has survived until nowadays – the genitive case. However, if the noun is not marked by a genitive inflection, it is said to be in common case.

The two most important functions of the genitive case are to classify (her two children's clothes disappeared, her hair felt like a bird's nest) and specify (a/the/that/the girl's face). There are different kinds of genitives (Biber et al. 1999:293-299):

- genitives of time (last week's Observer, yesterday's job)
- genitives of measure (a minute's hesitation, at arm's length)
- elliptic genitives (If a car's dirty, it's a woman's [car], this isn't my handwriting, it's Selina's! [handwriting],)
- group genitives (mother-in-law's house, the father of the five's)
- double genitive (a good idea of Johnny's, a relative of Kupka's)

Noun also takes the category of gender, although Biber et al. (1999: 311-312) state that

> 'gender is a less important category in English than in many other languages. It is closely tied to the sex of the referent and is chiefly reflected in co-occurrence patterns with respect to singular personal pronouns (and corresponding possessive and reflexive forms).'

Nouns are divided into genders accordingly (ibid.):

Table 1.1 **Noun gender classes**

Personal/human	Example nouns	Pronouns
Masculine	Tom, a boy, the man	He
Feminine	Sue, a girl, the woman	She
Dual	A journalist, the doctor	He, she
Non-personal/neuter	A house, the bird	it

Verbs

Grammarians like Downing and Locke (2006: 317) and Biber et al. (1999: 368) both agree that verb is the most important word class in English because it has four functions: negation, inversion, code (substitution) and emphasis. Furthermore, the verb represents the experience of certain events, as well as all types of processes, states and activities.

Additionally, researchers Quirk et al. (1985: 96) join Downing and Locke (2016) and Biber et al. (1999) to state that verb can have two roles in a verb phrase – either the main verb

(that can stand alone in a sentence) or the auxiliary verb (a verb that needs to be together with a main verb).

Linguists Quirk et al. (1985:96), Biber et al. (1999:358) and Downing and Locke (2006: 317-318) state that verbs can be classified into three main groups, based on their function in a sentence: full or lexical verbs (*come, wait, dance, drink*), primary verbs (*be, am, is, are*) and modal auxiliary verbs (*must, would, should*). Full verbs can act only as the main verbs, modal auxiliaries can be only auxiliary verbs, and primary verbs can be either main verbs or auxiliary verbs.

Furthermore, Downing and Locke (2006: 318) divide the modal auxiliaries into three different categories: modal auxiliaries (*must, ought to, will, would*), semi-modals (modals in certain uses) (*need, dare, used to),* and lexical auxiliaries (*be able to, be about to, have to, have got to, had better, would rather*).

According to Quirk et al. (1985: 96), regular verbs can have up to five morphological forms. Additionally, regular verbs can be divided into regular and irregular ones (ibid.):

Table 1.2 **Morphological forms of regular and irregular verbs**

	Regular verbs	Irregular verbs
Base form	Call, want	Speak, cut, win
-s form	Calls, wants	Speaks, cuts, wins
-ing participle	Calling, wanting	Speaking, cutting, winning
Past form	Called, wanted	Spoke, cut, won
-ed participle	Called, wanted	Spoken, cut, won

Biber et al. (1999: 396) maintain that 'for many verbs, regular and irregular variants can be used both as past tense verbs and as past participles' (see Table 1.2), however, irregular verbs have more morphological forms, especially in past form and –ed participle. Quirk et al. (1985: 96) exemplify with the word *be* that has eight different morphological forms.

This leads to the discussion of finite and non-finite verb phrases (see Table 1.2) discussed by grammarians like Quirk et al. (1985:97), Downing and Locke (2006: 100-108), Biber et al. (1999: 193-201), Hudson (2010: 257-258) and Huddleston and Pullum (2002: 36-37). Finite form of the verb is 'limited', meaning that it is limited in terms of subject and tense. Non-finite, however, are not. According to Huddleston and Pullum (2002: 36) the only verb form that occurs in both finite and non-finite clause, is the plain form (or base form). Additionally, they state the relation between clause finiteness and verb inflection:

a. If the verb is a primary form, the clause is finite.
b. If the verb is a gerund-participle or a past participle, the clause is non- finite.
c. If the verb is a plain form, the clause may be finite or non- finite; specifically:

a. Imperative and subjunctive clauses are finite
b. Infinitival clauses are non-finite.

Table 1.3 **Relation between clause finiteness and verb inflection**

	VERB-FORM	CONSTRUCTION	EXAMPLE	FINITENESS
i	PRIMARY FORMS		*She brings her own food.*	FINITE
ii	PLAIN FORM	IMPERATIVE:	*Bring your own food.*	FINITE
iii	PLAIN FORM	SUBJUNCTIVE:	*We insist [that she bring her own food].*	FINITE
iv	PLAIN FORM	INFINITIVAL:	*It's rare [for her to bring her own food].*	NON-FINITE
v	GERUND-PARTICIPLE		*She regrets [bringing her own food].*	NON-FINITE
vi	PAST PARTICIPLE		*This is the food [brought by my sister].*	NON-FINITE

Huddleston and Pullum (2002: 36) provide a comprehensive table (See Table 1.3), where it can be seen that finiteness ends after the subjunctive construction due to the fact that 'the structure of non-finite subordinate clauses differs more radically from that of main clauses than does that of finite subordinate clauses' (ibid.)

Adjective

Researchers like Downing and Locke (2006: 475) and Huddleston and Pullum (2002: 112) agree that adjectives can express a state (*sad*), a quality (*thin*), a sub-class (*southern*), or property (*intelligent*). They can also express attitude (*lovely*) and a judgement (*false*).

Additionally, Biber et al. (1999: 504) state that adjectives are more common in written texts, especially academic prose, and they modify nouns, therefore complementing the informational density of descriptive nature of registers, for instance, news and academic prose.

Quirk et al. (1985:402) mention the problem that it is not possible to identify the adjective just by looking at it, but certain suffixes can signal that the word could possibly be an adjective, for instance, *-able, -ful, -ish, -ous, -al, -ic, -less* and *–y*. However, it is not always possible to do so since there are adjectives with no identifying characteristics (*good, bad, evil).* Additionally, Huddleston and Pullum (2002: 226) propose that a word can easily be checked whether it is an adjective by adding the intensifier *very.*

Grammarians Downing and Locke (2006: 477) divide adjectives into three groups: simple, derived and compound adjectives. In the category of simple adjectives fall all adjectives that are usually monosyllabic or bisyllabic, for instance, *good, big, bad, small* and others. Compound adjectives are adjectives that are formed from nouns, verbs or other

adjectives by adding a suffix, for instance, *foremost*, and *handy*. Other adjectives are formed from either Latin/Greek base (*central, secondary*) or French base (*readable, marvellous*).

Quirk et al. (1985: 403) determine four criteria of adjectives: they can occur as an attribute and premodify a noun (*beautiful girl, fluffy dog*), they can occur in a predicative function and function as a subject or object complement (*The girl is beautiful, the dog is fluffy*), they can take comparative and superlative forms (*big – bigger – biggest, beautiful – more beautiful – the most beautiful*), as well as they can be premodified with the help of the intensifier *very* (*very beautiful girl, very fluffy dog*).

Adverb

Grammarians (Quirk et al. 1985: 438, Biber et al. 1999: 762, Huddleston and Pullum 2002: 122, Downing and Locke 2006: 503) agree that adverbs state how an action is done and that adverb modifies verbs, clauses, adjectives and other adverbs. However, Huddleston and Pullum (2002:122) and Quirk et al. (1985:438) emphasize the fact that adverb class is the most obscure due its great heterogeneity and the ability to readily accept new members.

Linguists Biber et al. (1999:762), Huddleston and Pullum (2002: 122) and Downing and Locke (2006: 502) maintain that adverb serves one of three major functions: add circumstantial information, to express a stance or to link a clause.

As stated by Quirk et al. (1985: 438), three different groups of adverbs can be distinguished, two of them being closed class with simple (*just, well*) and compound adverbs (*somehow, somewhere*) and lastly – open class, also called derivational, (*oddly, interestingly*).

1.2 Function words

Preposition

Prepositions, according to Downing and Locke (2006: 531-540), Huddleston and Pullum (2002: 127) and Quirk et al. (1985: 657) ‘express relation between two entities, one being that represented by the prepositional complement and the other by another part of sentence’. (***on*** *the table,* ***in*** *the* kitchen). Prepositions can consist of one, two, three and occasionally four words, but are considered as a single preposition.

Biber et al. (1999: 74) emphasize that a very important distinction in prepositions is the free prepositions versus the bound ones, where free prepositions can stand their own and their meaning is not dependent by any specific words in the context (he’ll go with one of the kids, late one morning in June), whereas the bound prepositions often have ‘little, independent meaning, and the choice of the preposition depends upon some other word’ (*They've got to be*

*willing to part **with** that bit of money*. *She confided **in** him above all others*.).

Lastly, as stated by Quirk et al. (1985: 657), prepositions can serve three different syntactical functions: as a postmodifier in a noun phrase (*the children in the school were learning*), as a conjunct (*on the other hand, the children were learning in school*) and as an adverbial, that can be rouped into adjunct (*the children were learning in the school*), subjunct (*the children just stopped learning*), disjunct (*without a doubt, the children were learning in the school*).

Pronoun

As stated by contemporary grammarians (Quirk et al. 1985: 335, Biber et al. 1999: 327-329, Huddleston and Pullum 2002: 100-102), 'most pronouns replace fully specified noun phrases and can be regarded as economy devices'. It is important to remember that pronouns are a subclass of nouns and contemporary linguists often choose to classify them as being together with nouns. However, they are 'distinguished syntactically by their inability to take determiners as a dependent' (Huddleston & Pullum, 2002: 100), which means that the proper statement will be *I am ill*, not *This I am ill.* Furthermore, Quirk et al. (1985:355) agree that it is 'best to see pronouns as comprising a varied class of closed-class words with nominal functions'.

Quirk et al. (ibid) also mention that pronoun can have three senses: as a substitute for a word or a phrase (*one*), signalling a reference to a person or anything in the linguistic context (*him, her*), and lastly – as a general concept (*somebody, something*).

Lastly, Hudson (2010: 266) provides a comprehensive overview of nine different pronoun types in the English language (See Table 1.4)

Table 1.4 **Types of pronouns**

Pronoun class	**Members**
Personal	Me (I), you, him (he), her (she), it, us (we), them (they)
Reflexive	Myself, yourself, himself, herself, itself, ourselves, yourselves, themselves
Reciprocal	Each other, one another
Possessive	Mine (my), yours (your), his, hers (her), its, ours (our), theirs (their)
Relative	Who (whom), which, whose, when, where
Interrogative	Who (whom), what, which, whose, when, where, how
Demonstrative	This (these), that (those)
Indefinite	One, some, any, each, every, none (no)
Compound	Everything, something, anything, nothing, everybody, etc, everyone, etc, everywhere, etc.

Articles and determiners

As stated by Downing and Locke (2006: 417-418) and Biber et al. (1999: 69-70), articles are words that determine or limit a noun or noun phrase. Downing and Locke (2006:417) add that '**definiteness** is marked by the definite article *the* and by the determinatives *this, that, these, those* or by the possessives *my, your*, etc.', whereas '**indefiniteness** is marked by *a(n), some, any* and *zero*, and **generic reference** by *zero*, and generic reference - by *a(n)* and by *the*.'

However, contemporary linguist Hudson (2010: 253) notes that modern linguists would typically delete the article from the list of parts of speech since 'determiner' is a broader class which is more efficient because 'it has more properties and members'.

Biber et al. (1999: 69-70) provide a comprehensive overview regarding articles and determiners:

- definite article (*the book*) – specifies that referent is known to the speaker and addressee
- indefinite article (*a book*) – narrows down the reference to a single member in a class
- demonstrative determiners (*this* book, *that* book) – establish the reference by proximity to the speaker and the addressee
- possessive determiners (*my* book, *your* book) – establish a connection with the participants in the speech
- quantifiers (*some* book, *many* books) – specify the number of entities referred to.

In conclusion, the contemporary part-of-speech division groups parts of speech into three groups: lexical words, function words and interjections. Lexical and function words are analysed further, specifically nouns and verbs.

Nouns are the core part of speech in English and they refer to classes of entities that have experiential features. There are three types of nouns common, proper and pronouns. Common nouns are determined by their countability – singular and plural nouns, as well as irregular and zero plurals. Nouns can form a noun phrase – a phrase of which a noun is a head of. They can range from simple to more complicated, but in the noun phrase, the noun can act either as a subject, object, complement of clauses or complement of prepositional phrases. Nouns can also have the category of case - the genitive or common case, and they can also take the category of gender.

Verbs are the most important part of speech in English because they have four functions: negation, inversion, code (substitution) and emphasis and verbs represent the experience of certain events. Verbs can have two roles in a sentence – main verb or auxiliary verb. They can be classified into three main groups, based on the function in the sentence: full verbs, primary verbs and modal auxiliary verbs. The modal auxiliary verbs can be divided

into modal auxiliaries, semi-modals and lexical auxiliaries. Verbs can have up to five morphological forms and they can be divided into regular and irregular verbs. Verbs are also a part of finite and non-finite verb phrases.

The lexical word group consists of nouns, verbs, adjectives and adverbs and function word group consists of prepositions, conjunctions, pronouns and articles. To discuss them, the term ‘parts of speech’ will be used throughout this bachelor’s thesis.

Chapter 2
Previous Research of Tweets as Texts

Chapter 2 deals with the analysis of previous research of tweets as texts, which includes four different approaches, as well as an overview of features of microblogging as a messaging system with the focus on the functions and the structure of Twitter.

Linguistic research connected with tweet texts displays several approaches when analysing tweet data. The approaches can be grouped according to what linguistic features are brought out in each study: (1) Twitter tweet analysis that is devoted to the determination of the geolocation and language diversity (e.g. Chinese, French, German), (2) sentiment analysis, (3) qualitative analysis of tweet texts, compiled in mini-corpora and (4) multidimensional analysis of Twitter tweet texts.

The investigation of the geolocation of the tweets sent is a research trend based on the compilation of tweet corpus where the main variable for analysis is the location from which a particular tweet has been sent (the longitude and latitude of the device the tweet is sent from (Online 2). For example, researchers like Graham, Hale and Gaffney (2014) and Zhao and Cao (2017) use geolocation to study language diversity (whether it is English, Chinese or other) used in different locations. Graham, Hale and Gaffney have compiled a corpus of 111 million tweets, using various programs such as *Compact Language detection kit, Alchemy API, Xerox open source, Twitter UI language, Google geocoding API* and *Yahoo! PlaceFinder*. After all the tweets were compiled, 1000 of them were chosen from four areas – Cairo, Montreal, San Diego and Tokyo, because of the language diversity that could uncover. The main conclusion of the research was that it is incredibly difficult to determine and analyse the language of the tweet in an automated way because tweets tend to be in multiple languages (bilingual people or polyglots tweeting). The mentioned researchers have asserted that

> The informal writing style, short length of tweets, use of multiple languages within a single tweet, and the presence of non-language-specific content such as Uniform Resource Locators (URLs) and emoticons complicate the identification and limit accuracy. (Graham, Hale and Gaffney (2014: 575)

Researchers Zhao and Cao (2017), however, used geolocation to study language diversity in tweet texts in different parts of Hong Kong, as it is presumed to be ‘an important indicator of city’s internationalization’ (Zhao & Chao, 2017: 2698). The results show that the highest value of spoken language diversity was spotted in Hong Kong International Airport and the University of Hong Kong. These results are self-evident since both locations normally harbour larger spoken language diversity – the airport because people from different countries

visit it on daily basis, and the university because of the exchange students and those who come to study from a different country.

Although researches connected with geolocation investigate the language diversity in tweet texts in different locations and regions, these studies are technology-based and mainly disregard the linguistic aspect of it (only a few of the studies (Zhao & Chao, 2017; Graham, Hale and Gaffney, 2014) that used geolocation in connection with languages were obtained).

The analysis of adjectives in tweet texts by applying sentiment analysis or opinion mining which 'deals with the computational treatment of opinion, sentiment, and subjectivity in text' (Pang and Lee, 2008: 2) has been experiencing sudden burst of activity due to the technology-based tools that can be applied to ease opinion mining. Taking into account that sentiment analysis research is based on the extraction and analysis of adjectives, the obtained results provide an insight into the linguistic characteristics of tweet texts.

Research regarding sentiment analysis has been conducted, for example, by researchers Öztürk and Ayvaz (2018), Wang and Fikis (2017) and Ceron, Curini and Iacus, (2014), who all have used the same methodology. First of all, a sentiment lexicon is compiled using a relevant opinion mining program that extracts adjectives expressing a particular emotion, then the tweet corpus is processed by assigning tags with a part-of-speech tagger (for instance, *Stanford POS* tagger or *CLAWS* POS tagger) to single out the adjectives from Twitter texts, cross-comparison is made and, finally, graphs and/or charts are created to arrange the obtained data.

Twitter text-based sentiment analysis was conducted by Öztürk and Ayvaz (2018), in the process of which the adjectives from tweets in the Turkish and English languages regarding Syrian refugee crisis were extracted and analysed. Before using the *RSentiment* program, which enables the user to extract the required adjectives out of the tweets, the researchers created a sentiment lexicon. The results revealed that the Turkish tweets showed a more positive sentiment towards the refugees, whereas English tweets expressed a more neutral and even negative sentiment towards the Syrian refugee crisis.

Wang and Fikis (2017) conducted a research on public's opinion regarding Common Core State Standards (CCSS). Their research procedure was similar to Öztürk and Ayvaz – they compiled a corpus of tweets containing hashtags #ccss and #commoncore, and then mined the data to obtain the sentiment analysis results. The researchers concluded that 'Twitter users expressed overwhelmingly negative sentiment towards the CCSS' (Wang & Fikis, 2017: 1)

Ceron et al. (2015) research was connected with using sentiment analysis to monitor electoral campaigns. The research procedure was similar to the procedure applied in the

previously discussed studies, whereas the main conclusion of Ceron et al. was that the sentiment analysis of Twitter tweets of political nature has the ability 'to 'nowcast' and even forecast election results' (Ceron et al., 2015: 3).

Although sentiment analysis deals with adjectives, such studies are fully technology-based and produce conclusions only within the framework of adjectives and their functions – the expression of a positive, neutral or negative sentiment, and the compilation and analysis of adjectives is the only linguistic feature these studies focus on.

Another insight into linguistic characteristics of Twitter texts is the compilation of mini corpora of Twitter texts aiming at their detailed linguistic analysis. Such an example can be found in David Crystal's *Internet Linguistics* (2011: 36-56), where a mini-research was conducted, using Twitter texts. A sample of 200 tweets was collected, using the keyword *language*. As he states, his aim was to 'draw conclusions about the linguistic character of Twitter using English as the medium of illustration.' (Crystal, 2011: 41). However, several problems were encountered during the process – as Crystal states (ibid., 41), the research process of retweeting is more complicated than simply forwarding an e-mail, where the original message is seen, accompanied with the text from the person forwarding it, to explain what has the forwarder done. In Twitter, however, only the original message is seen, and the phenomena called repetitiveness is a characteristic of Twitter, which was the first problem, and, to make the sample more representative, Crystal made the decision to exclude the retweets; the same was done to tweets in a foreign language, which brought the size of the sample down to 159 tweets. However, Crystal did not have any issues with the extraction of the tweets, because the process was done by manually entering the word 'language' in the search bar and copying the tweets that showed up in the results. Manually copying the tweet texts is a viable option if the corpus is really miniature (in Crystal's case, after removing the duplicates – 159 tweets), but as soon as the tweet count is larger, it is almost impossible to manually extract the tweets due to time constraints.

After extracting the Twitter tweet texts, David Crystal put forth several conclusions regarding Twitter by using only this small sample. One of the conclusions is connected with content issues, where Crystal notes that the tweets are quite short in linguistic content (ibid., 43), with only 100.9 characters per average, which gives no room for linguistic analysis of the tweets – Crystal notes that there was only one tweet that had the maximum character count – 140, and most of the tweets were very short. Also, the 'number of dots and ellipsis is erratic' (ibid.), where some tweeters reduce the dots in ellipsis to two, but some increase to four, and additionally, these dots usually do not have anything to do with truncation of the tweet, which, in turn, is very confusing for the one who is analyzing the tweets - sometimes the

tweet is seemingly finished, but the tweeter has chosen to end each sentence with ellipsis or four dots. Moreover, Crystal notes that there is the issue with tweeters using all kinds of contractions and acronyms (ibid.,44) (usually by omitting the vowels of the word or using just the first letter of each word), which makes it more difficult to identify the class of the part of the speech, for instance, with words like *u* (you), *fav* (favourite), *ppl* (people), *gf* (girlfriend) etc.

Crystal concludes (ibid., 48) that, due to the contractions and free use of ellipsis (either two or four dots), it is difficult to 'assign definite syntactic analysis to an utterance'. The acronyms used in the tweets mean something more, for instance, *btw* (by the way) and *smh* (shaking my head), therefore they need to be considered as a separate sentence inside the tweet. The main issue is that the tweets vary from grammatically correct to completely incomprehensible ones, filled with contractions, acronyms, which, in turn, makes any kind of linguistic analysis more complicated (ibid., 48).

The fourth approach that aims at detailed uncovering of linguistic characteristics of tweets is their multidimensional analysis (MDA). This corpus-based analysis is widely applied (e.g. Titak & Roberson, 2013; Biber and Kurijan, 2007; Biber, 2003; Grieve et al., 2010; Reyes, Rosso & Veale, 2012) in order to reveal the linguistic features that are typical for the texts of a particular genre and/or register.

Titak and Roberson (2013) employed multi-dimensional analysis to compare the variation of linguistic features in online texts of blogs, micro-blogs (Twitter), workplace e-mails, discussion posts and reader comments, basing their research on Biber's (1988) methodology regarding functional dimensions of web registers. The emphasis of their research lies in the fact that 'the rapid growth of online language makes it difficult for researchers to produce a theoretical model that defines online corpus design and provides focus for linguistic research across web texts' (Titak and Roberson, 2013: 236), as well as the fact that the linguistic characteristics in these various online text corpora have received minor attention. Therefore, the multidimensional analysis can help to identify sets of linguistics features that co-occur in various online texts (blogs, microblogging texts etc.).

The research procedure for Titak and Roberson's (2013) multi-dimensional analysis regarding texts that are available online was as follows (ibid., 238): firstly, an exploratory corpus of online texts (approx. 16,501,758 words (from which 3 504 762 were from Twitter) was compiled in the time period of years 2006 – 2013, using a combination of automated and manual compilation methods. Then the corpus was grammatically tagged, using the Biber tagger, and finally– the obtained results were interpreted in a more qualitative way. The focus of Titak and Roberson's study was

> to explore linguistic variation across comparable web registers in order to produce an initial set of functional linguistic dimensions that could be analysed further with a larger number of texts (Titak and Roberson, 2014: 238).

Accordingly, in multi-dimensional analysis, a dimension is a set of co-occurring characteristics in text or group of texts. (Online 3) and in Titak and Roberson's (ibid., 243-253) case, four dimensions were created:

1) *Personal Narrative Focus* versus *Descriptive, Informational Production* (28 linguistic features, positive scale includes features that suggest personal and subjective production of texts (private verbs, first-person pronouns, mental verbs), narrativity (demonstrative pronouns, emphatics) and syntactic features to provide more elaborated information (that- clauses with factive vers, all wh- words). Negative scale includes features that contribute to nominal style (passives, passives) and informational style of address (prepositions, be as the main verb, infinitive verbs))
2) *Involved, Interactive Discourse* (20 linguistic features, positive scale illustrates an involved and interactive style (private verbs, first and second person pronouns), whereas the negative side includes features of informational production, similarly as in Dimension 1.
3) *Complex statement of Opinion* (9 linguistic features, with only positive features (that-deletion, public verbs, that- clauses with factive verbs), there are no negative features)
4) *Past* versus *Present Orientation* (9 linguistic features, positive scale includes personal features (first-person pronouns), past tense narratives (past tense and third-person pronouns) and verbs (non-auxiliary verbs and activity verbs). Features on the negative scale include nominalisations, longer words and features that suggest a more analytical discourse, e.g process nouns)

The positive and negative features are assigned for each dimension so that it would be possible to put the six different online text types (blogs, e-mails, reader comments, opinion columns, newspaper articles and Facebook/Twitter updates) on a single scale within each dimension (Titak and Roberson, 2014: 246, 249, 251,253). The positive and negative features help to calculate a factor score – the more positive features a text type has, the higher the score.

For the purposes of the present research, the results of only the Twitter updates will be looked at. Accordingly, the results for Twitter updates in the first dimension reveal (ibid., 247) that Twitter updates 'focus more on the delivery of nominal information'. Moreover, Titak and Roberson (ibid.) state that

> The first dimension shows linguistic similarities between news articles and micro-blogs. Nouns (especially proper nouns) and prepositions are very common in these two groups

of texts.

Thus, it can be concluded Titak and Roberson (2014: 253), that Twitter tweet texts can also be regarded as being similar to small news updates, because, according to they have similar features.

When analyzing online texts, employing the multi-dimensional analysis is the best route to choose due to the fact that it deals only with the linguistic features and their co-occurrence in the texts, thus it is connected with linguistic aspects the most. Multi-dimensional analysis can be employed when detecting irony in Twitter texts (Reyes, Rosso & Veale, 2012), analyse register variations among blogs (Grieve et al., 2010) and others.

2.1. Features of microblogging as a messaging system

The specific features of Twitter as a messaging system determine the bulk and the structure of tweets and therefore, before conducting the empirical part of the research, it would be crucial to describe the features of microblogging, looking at it as a written text and analysing its' features, as well as describing the major structure of Twitter and changes that have been made in the microblogging system since its start on 2006.

According to Encyclopaedia Britannica (Online 4), Twitter was originally a small project of short messaging system (SMS) exchange service and was called *Twttr*, and primarily it was a free SMS exchange system. The social networking element became prominent in 2009 when celebrities and politicians started using Twitter to promote events and political campaigns. Accordingly, the biggest step in Twitter's evolution is its use as an outlet for information distribution, where amateur journalists can post pictures and update the information straight from the place of action. Such examples are the Iranian presidential election in 2009 and the earthquake in Haiti in 2010.

Linguist David Crystal (2011: 32) states that 'the language of the Internet cannot be identified with either spoken or written language, even though it shares some features with both', which puts the language of microblogging system Twitter somewhere in the middle between spoken and written language. Crystal states (2011: 20-21) that although microblogging has some features of written communication, where 'many of its functions (such as reference publishing and advertising) are no different from traditional situations that use writing', he emphasizes (ibid.) that 'email, chat, instant messaging, and texting, expressed through the medium of writing, display several of the core properties of speech'.

According to DeVoe (2009: 212), 'microblogging lets users share brief blasts of information to friends and followers'. She adds that although there are many microblogging

mediums, for instance, LinkedIn, Twitter, Facebook and Tumblr, Twitter is the one that has gained the most popularity over the years. However, according to research done in 2016 (Online 5), based on the comScore Digital Future Report (Online 6), Facebook has surpassed Twitter in its use and functionality – approximately 22% of the world use Facebook. However, approximately 83% of world leaders are active on Twitter (Online 7), therefore it could be possible that although nowadays Facebook has become the most popular social network site, Twitter is still used to promote events and political campaigns as a source for news.

Linguist David Crystal (2011:36-39) has laid out the foundations of the structure of Twitter, and, although some nuances have changed, the base of the system has stayed the same. First of all, he states that Twitter was the most rapidly growing microblogging medium in 2010, and often regarded as 'the SMS of the Internet' due to the fact that it is a platform that allows users to send and receive text-based posts that are up to 140 characters (20 of the characters are reserved for the username). He explains that the texts are regarded as *tweets* and the people who sent them –as *tweeters, twitterers, Twitter users* and other.

Twitter website clearly shows that it presents the opportunity to publish and share concise texts that contain the thoughts, opinion or any other news one would like to share without having to write and/or read blog posts that are noticeably larger than microblogs. Additionally, unlike blogging, the user of Twitter had the opportunity to 'follow' other users and vice versa, but, if a user does not want unapproved users seeing their posts, the Twitter user has the opportunity to 'lock' their account so that only approved users can access their tweets.

Crystal (2011: 37) states that as Twitter evolved, it added the functions of *retweet, reply, like, hashtag* and adding an external URI (uniform resource indicator) to the tweet.

The function of *retweet* allows a user to take another user's tweet and post it on their Twitter account, by either adding the letters **RT** in front of the tweet or pushing the *retweet* button.

The reply function allows the Twitter user to reply to a tweet, mentioning the original poster of the Tweet to reply to their tweet. The process is repetitive and several users can be mentioned in the tweet, thus replying to multiple users, not just one. An example:

- @marcorubio: When pride comes, disgrace comes; but with the humble is wisdom. Proverbs 11:2
- @_DECase replying to @marcorubio: The most prideful man I have ever seen is Trump. Why do you support him?

The like button allows the user to show their support of the original poster's opinion in their tweet, as well as save it in their personal Twitter folder, called *saved tweets.*

The hashtag function allows the Twitter user to add several keywords using the hash sign (#), which enables other users to find tweets that are related to similar content. An example:

- @replouiegohmert: Not enough of them [texts] is the problem. This is just more obfuscation on the part of people wanting to cover for the Obama Administration. #ReleaseTheTexts

This allows people who share similar interests to meet and read each other's tweets.

Lastly (Crystal, 2011:38), if a user wants to complement their tweet with additional information from a web source, they can add an external link in their Twitter post. However, the problem is that a normal URL (uniform resource locator) would be too long to fit in a tweet (for example, an address https://en.wikipedia.org/wiki/Turtle), therefore, each address is assigned a shorter, representative code in the form of bit.ly and the Wikipedia link is as follows - https://bit.ly/1R3g3Hy

Figure 2.1 **Example of a Twitter tweet**

Figure 2.1 presents a comprehensive example of present day Twitter tweet. The most important change that has been implemented since Crystal's (2011) analysis of Twitter as a messaging system, is the length of the tweet – instead of 140 characters, a tweet currently can be up to 280 characters long.

The structure of a tweet is as follows (see Figure 2.1) – it is comprised of the username (11), their account name and surname (1), the blue tick that represents that the account is 'verified' (Online 8) – shows to other users that the account of public interest is authentic (13), the button (10) that shows whether a person as a user if following the account of the person who tweeted the tweet, as well as the actual body of the tweet (2), the hashtags used

(3), other users mentioned in the tweet (4) and any external links (5). Additionally, the count of retweets (7) and likes (8) can be seen, as well as the @ replies (9), as well as the button which allows the user to directly message (DM) the author of the tweet (12).

Communication via Twitter, according to researchers Simpson (2002) and Crystal (2011) belongs to computer-mediated communication (CMC), which is a term to describe communication via computers and it has several distinctive characteristics, as proposed by Kaplan and Haenlein (2011). They maintain that this type of CMC is appealing to its users because of several reasons. First of all, it is the sense of ambient awareness (ibid., 107) – these short bursts of stream of consciousness 'can generate a feeling closeness and intimacy' (ibid.) in the person who is reading the tweets. Kaplan and Haenlein maintain that although reading one tweet from a user does not allow the reader to conclude anything about the user, all the tweets of a user, combined together, paint a general picture of what are the user's character traits, likes, dislikes and other characteristics.

The mentioned researchers explain in detail (ibid.) that one of the key characteristics of Twitter is the unique push-push-pull communication model. The first 'push' of information is achieved when a person follows someone else, thus reducing the amount of random information, but 'pushing the information they are interested in into their Twitter feed – this explains the celebrity Twitter accounts, with a significant amount of people 'pushing' their tweets.

The second 'push' happens, when a person decides that the information received is interesting enough for them to give it another 'push' by re-tweeting it. Lastly, the 'pull' part of the communication model happens when a user reads the information that has been 'pushed' and decides to give their input and add additional information to the already existing one by doing some research.

According to Kaplan and Haenlein (ibid., 108), although unpleasant, a characteristic of microblogging is voyeurism and virtual exhibitionism. They add that it is evident that the moment a person pushes the 'tweet' button, the tweet text the user has written becomes visible to everyone and a lot of people have access to it, except for cases when the user has specifically put restrictions for their tweets to be visible only to ones following them. Kaplan and Haenlein point out there are some users that enjoy the fact that their thoughts and actions are shared rapidly and are accessible to many people.

In addition to the Twitter functions, as proposed by Kaplan and Haenlein (2011), it is important to establish the communicative context of Twitter. Twitter is a very public domain and, although there exists the possibility to hide one's Twitter from public scrutiny, according to research done by Moore (Online 9, 2009), the number of protected accounts is dropping

and is less than 10% of the total number of Twitter users. Thus, taking into account that Twitter texts are published and retweeted at a rapid pace and the fact that approximately 500 million tweets are sent in one day (Online 10), Postmes and Brunstig (2002: 299) claim that Twitter tweet exchange can be regarded as a mass communication.

According to Matthiessen (1995, as cited in Zappavigna, 2015, 49), communicative context is 'functionally diversified' as a combination of field, tenor and mode. According to Zappavigna (ibid.), field is similar to the topic of the text, answering the questions 'what is happening' or 'what it is, whereas the tenor of the text deals with the participant relationships. Mode, according to Zappavigna, refers to the 'medium facilitating the communication'. However, according to researchers Crystal (2005: 1) and Kloučková (2013: 1), it is difficult to classify whether the medium of Twitter is distinctly spoken or written. They argue that computer-mediated communication is different from traditional spoken or written communication, and a comprehensive classification system to grasp all the modes of computer-mediated communication, has not yet been created.

Researcher Wikström (2017) mentions the *spokenness of CMC* (ibid., 43) in his doctoral thesis, stating that although computer-mediated communication is a written interaction, it is being pulled more towards the spoken communication. He adds that an assumption has been addressed that computer-mediated communication should be regarded as 'written speech' (Maynor, 1994, as cited in Wikström, 2017, 43).

To conclude, although Twitter tweet texts present a noticeable amount of variables that can be detected in each Twitter tweet (location, the use of emoticons, gender, the status of account (verified/unverified) etc.), not all of these researches put the linguistic aspects as the main element of the research. However, it is possible Twitter is still under-investigated since the platform was released only 12 years ago, and it is constantly upgraded, and the platform's functionality could be improved in the future.

Additionally, it can be observed that there are numerous researches that are concerned with sentiment analysis because it is relatively easier to extract the data and display it in tables and charts in the empirical part of the research with the help of relevant programs. However, the frequency of verbs and nouns in tweet texts have not been widely investigated, and therefore the present research could prove to be a valuable insight.

Several conclusions can also be made regarding the features and characteristics of microblogging as a messaging system:

Microblogging (Twitter) enables the users to share concise pieces of information for their friends and followers. Twitter has some basic functions, for instance, *retweets, mentions, likes* and *hashtags,* according to Crystal (2011), but the largest change that has occurred since

the publishing of Crystal's analysis of Twitter as a messaging system (2011), is the maximum characters allowed per tweet – if previously those were 140 characters, then now the maximum character count has been doubled to 280 characters.

According to Kaplan and Haenlein (2011), Twitter has several distinctive characteristics: the sense of ambient awareness, the unique push-push-pull communication model and lastly – the characteristic of voyeurism and virtual exhibitionism.

Researchers Wikström (2017), Crystal (2005, 2011) Kloučková (2013) state that the mode of computer-mediated communication (thus, including Twitter tweet texts) could be regarded as written speech – it is presented in a written form, but displays characteristics of spoken language – however, it is still predominantly a written language that is pulled strongly towards speech. Accordingly, the communication of Twitter tweet texts can be regarded as a mass communication.

Chapter 3

Methodology

Corpus Linguistics as a Quantitative Research Method

Chapter 3 deals with the definition of corpus linguistics as a quantitative research method, the process of corpus creation as well as the types of corpora.

The glossary, provided by Hardie, McEnery and Baker (2006: 48-49) states that corpus (Latin for *body*) is a 'collection of texts' (a body of language), that are large in body, are machine-readable and are stored in an electronic database. The researchers add that corpora can be used for both quantitative and qualitative types of research.

Researchers McEnery and Hardie (2014), Lüdeling and Kyto (2009), Biber, Conrad & Reppen (2012) and Kennedy (1998) explain that corpus linguistics is 'the study of language data on a large scale - the computer-aided analysis of very extensive collections of transcribed utterances or written texts' (McEnery and Hardie, 2014: ii). The researchers mentioned propose to view corpus linguistics as an area of procedures and methods to investigate language better, and, although they agree that the research procedures of corpus linguistics are still in development, they maintain that corpus linguistics has the opportunity to 'reorient our entire approach to the study of language' (McEnery & Hardie, 2014: 1).

According to McEnery and Wilson (2001: 103-129), corpus linguistics can be applied in various fields of linguistics, for instance, psycholinguistics, speech research, grammar, sociolinguistics and others.

Kennedy (1998: 1) and Biber, Conrad and Reppen (2012: 12) state that corpus is any kind of 'collection of texts in an electronic database' (Kennedy, 1998:1), which is a 'large and principled collection of natural texts' (Biber, Conrad & Reppen, 2012: 12).

3.1 Corpora Types

In order to correctly classify the corpus created for the purposes of the present research, the distinction between general and specialised corpora needs to be established. Researchers like Kennedy (1998: 3-20) and Gatto (2014: 15-16) propose a comprehensive distinction regarding the largest groups of different corpora typology.

Gatto (2014: 15-16) proposes three main groups of corpora typology: general vs. specialised, synchronic vs. diachronic and monolingual vs. multilingual. She explains that general corpus 'includes texts from various domains, genres and registers and can contain both written and spoken languages', whereas specialised corpus deals with 'representing only a given variety or domain of language in use, such as medical discourse or academic

discourse and restrictions may apply not only to domain, but also to genre, time and geographical variety'. Examples of general corpora include British National Corpora (Online 11) and the Bank of English, and examples of specialised corpora are any corpora built for a specific genre or study.

McEnery and Wilson (2011: 32-33) explain that both general and specialised corpora can be either unannotated or annotated – if the corpora are unannotated, the data is presented in its raw form and plain, whereas if the corpora is annotated, it is enhanced with various linguistic features, for instance, as in the case of the present study – part of speech tags. The researchers further add that, although unannotated corpora have been largely used in linguistic studies, the utility of annotated corpora is greater than the unannotated one.

The corpora type that the present research is based on is the specialised corpus, specifically – corpora of computer-mediated communication (Ludeling & Kyto, 2009: 292-306). The researchers add that for empirical studies, these 'corpora usually have to be individually acquired from the Internet', thus the CMC corpora can be divided into project-related or for general case. In the case of the present research, the CMC corpus is project-related.

The corpus type of the present research can be classified as specialised, annotated, project-related, computer-mediated communication corpus.

3.2 The Process of Corpus Creation

In order to conduct the corpus-based quantitative research, a specialised, annotated corpus was created for the specific purposes of the present research.

Nelson (cited in O'Keeffe & McCarthy, 2010: 53) states that although it might seem that compiling a written corpus is easier, one has to deal with the process of choosing the texts, gaining the access to them, as well as the storage and analysis of the texts.

Nelson (ibid.) also notes that the corpus size needs to be established so that it could be representative of the view of the language, meaning that the corpora are large enough to represent the language feature, genre etc. that are chosen to be studied. However, he states that 'any attempt at corpus creation is a compromise between the hoped for and the achievable', because no corpora can be indefinitely large in size, therefore, the researcher has to make constant decisions regarding each step of the process of corpus creation.

Sinclair (2005: 1-16), maintains that the corpus should be built according to the communicative function and community within which it arises and pays great attention to the fact that in the process of corpus creation, two target notions should be kept in mind: representativeness and balance. Sinclair adds that although those are not precise and

attainable goals, they are to be used as a guide in the process of creation of a corpus.

Thus, the present research is based on the methodology of Nelson (2010:53-63) and Sinclair (2005:1-16) in order to build a homogenous corpus that would accurately represent the verb and noun frequency and use in verified and unverified sub-corpora of Twitter tweet text corpus.

3.3 Absolute and Relative Frequencies within the Context of Corpus Analysis

Xiao (Online 12) maintains that there are two types of frequencies – absolute (raw) and relative (normalised), and both of them are types of descriptive statistics.

He states that absolute frequency is the arithmetic number of the instances of linguistic features within a corpus – it is the most direct result one can get out of corpus. However, the absolute frequency does not say anything in terms of validity for individual instances in generalised corpora – Xiao (ibid.) mentions the example of swear words in BNC corpus – the fact that there are swear words in BNC corpus, that does not mean that people swear a lot in general.

Regarding the relative frequency, Xiao (ibid.) explains that when different corpora (even sub-corpora) are compared, the relative frequencies need to be calculated in order to 'determine how often it could be assumed that the word *Y*/part of speech will be seen within the bulk of corpus text' (ibid.) The relative frequency is calculated accordingly: Rf = Absolute frequency / total count of words in a corpus and it is usually expressed in percentage.

When comparing the verified and unverified sub-corpora data of the corpus *VerUn* that was created for the specific purposes of the present research, it was concluded that both corpora are drastically different in size (11244 tweets for verified tweet sub-corpora and 3114 for unverified tweet sub-corpora), therefore, the decision to use relative frequencies when comparing the two sub-corpora was made.

The following formulas were used:

- Rf_1 = verified tweet noun or verb frequencies / 11244 (total count of verified account sub-corpora tweets)
- Rf_2 = unverified tweet noun or verb frequencies / 3114 (total count of unverified account sub-corpora tweets)

In conclusion, corpus linguistics as a method deals with language data on a large scale, and the analysis of it is computer-aided.

The corpus type of the present research can be classified as a specialised, annotated, project-related, computer-mediated communication corpus. Moreover, the present paper will

use the methodology regarding corpus creation proposed by Nelson (2010) and Sinclair (2005). Absolute frequency represents the direct data obtained by corpus, but relative frequency represents the percentage of possibility of the linguistic feature occurring within the bulk of the corpus text.

Chapter 4
The Analysis of Twitter Tweets

Chapter 4 describes the empirical part of the research. It presents the steps of creating a corpus, data extraction, the process of analysing the obtained data, as well as presenting results and relevant conclusions regarding the frequency of verbs and nouns both in verified and unverified Twitter account sub-corpora, and the comparison of different noun and verb tags in both sub-corpora.

The method used in the present study is corpus quantitative analysis, which allows to obtain and analyse the frequency of verbs and nouns in both sub-corpora created and draw relevant conclusions.

4.1 Research procedure

The research material selection and the arrangement in the specialised, annotated corpus of Twitter tweets *VerUn* was done according to the corpus creation theories proposed by Nelson (2010) and Sinclair (2005) and is comprised of six steps.

1) Creation of the *VerUn* corpus

The corpus creation comprised of six sub-steps:

Step 1: the preferable tweet count was set – approximately 10,000 tweets to be collected and analysed. Although some Twitter datasets can be obtained on the Internet (Online 13), there are none that fit the purposes of the present research, and therefore the author of the present bachelor thesis created their own corpus, which proved to be a considerably challenging process.

Step 2: the program TCDE (*Tweetcatcher Desktop Edition*) (Online 14) was selected to accumulate the tweets. Although the program TCDE comes together with the program TweetVis that creates graphs for sentiment analysis, for the purpose of the present paper, only the program TCDE was used. However, installing the program proved to be a time-consuming process because once the program was installed, it was necessary to go through solving several errors and troubleshooting before it could even be launched. After several attempts to fix the problems, the program finally complied, and it was possible to open it.

Step 3: the process of tweet aggregation, which comprised of three further sub-steps:

- firstly, it was determined that the main sub-corpus of tweets to be collected would be from the verified politician accounts (*Ver*), and the second sub-corpus (*Un*), for comparison, would be the tweets connected with political themes in the unverified Twitter accounts.In compilation of sub-corpus *Un*, hashtags #constitution, #potus,

#whitehouse, #elections, #government, #legislation, #political and #president were used, the tweets were accumulated in separate .xls files (8 hashtags – 8 separate .xls files). Together, 8246 tweets were accumulated before any cleaning and deletion.

- Secondly, in the process of accumulating the tweets connected with political themes in the verified accounts (sub-corpus *Ver*), a separate Twitter account was created just for the purpose of the corpus creation so that it would be possible to 'follow' only those Twitter accounts that are of interest in connection with the present paper.Accordingly, some of the most popular American politicians were followed from the specialised Twitter account, for instance, Donald Trump (@realDonaldTrump), Hillary Clinton (@HillaryClinton), Barack Obama (@BarackObama) and Bill Clinton (@BillClinton). American politician Twitter accounts were chosen for this research because those Twitter accounts were more frequent and accessible, which would accordingly make the corpus more unified and cohesive. Additionally, the microblogging site Twitter originates from America, which is the other reason why American political accounts were followed. Then, using the Twitter function of suggestions, called 'Who to Follow', forty-six more political, verified accounts were followed – altogether, fifty politician accounts were followed. Then, using the same tweet aggregation program used for unverified tweet subcorpus (TCDE), tweets from the specifically made Twitter account were extracted. Together, 600 tweets were extracted from each account (25962 tweets in total) and accumulated in one .xls file before any cleaning and deletion.
- Further, the author of the paper noticed that it was not possible to extract the 600 tweets from each account due to the fact that the program allows extracting tweets only from the past two months, therefore from Hillary Clinton's, Senator Roger Wicker's, Senator Thad Cochran's, Senator Elizabeth Warren's, Rep. Jim Jordan's, Trey Gowdy's it was possible to extract only 400 tweets, and from Barack Obama's, Bill Clinton's and Mike Pence's Twitter only 200 tweets were retrieved. This leads to believe that Twitter activity is closely connected with the particular situation – for instance, Barack Obama is no longer the president of the United States, and therefore he has no need to tweet so often. Similarly, Hillary Clinton campaigns for the position of the president of the United States, therefore, she has no need to tweet so much and be 'up to date'. A decision to remove the duplicates and *retweets* was made due to the fact that they could potentially present statistically wrong results (similarly as David Crystal had done in his mini-corpus Twitter analysis (Crystal, 2011: 36-56). For the present research the tweets have to be as original as possible to obtain objective and

linguistically accurate results. A major drawback was spotted in the raw (data that is not annotated and is presented in its raw, untouched form) data. Due to the fact that the microblogging site Twitter had recently been updated and at the present the maximum available character count for the Twitter message is 280 characters, whereas none of the programs have been updated to take into account that tweets can possibly be larger than 140 characters. It meant that all tweets extending the 140-character limit, are truncated and end with […] (See Figure 4.1), followed by the URL address that shows the location of the full tweet.

Step 4: The decision to manually 'complete' the tweet by looking it up on the Internet using the URL address to open the tweet directly and manually copy it into the raw data .xls file was made.

The process of completing the incomplete tweets took a considerably long time (approximately 77 hours). It was decided that it is crucial to finish the tweets and obtain the full spectrum of data, mostly as the present research dealt with verb and noun variations and frequency. It is more logical to create the tweet corpus and then look at the noun and verb frequencies when there is a bigger possibility for nouns and verbs to appear in the tweets, that is enabled by the recent update of 280 characters.

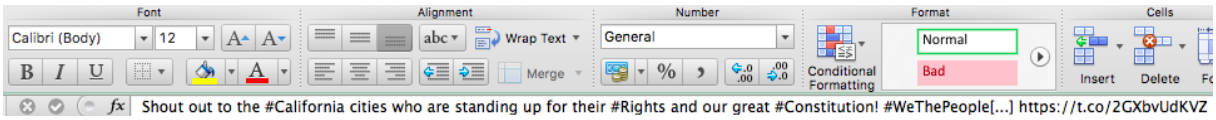

Figure 4.1 **Example of an unfinished tweet**

Step 5: During the process of completing the incomplete tweets, all the hyperlinks were removed, even if the tweet was already completed, in order to achieve data that was as clean as possible – no hyperlinks and emoticons. However, the decision to leave the hashtags was made as they also could potentially contain verbs or nouns. Additionally, during the process of tweet completion, the tweets that had no linguistic value, for instance, tweets that contained only external links, were removed.

After completing the tweets, as well as removing all the duplicate and *retweet* tweets, as well as deleting all the hyperlinks and emoticons, the final corpus consisted of 3 114 unverified account tweets and 11 244 verified account tweets, which makes the total tweet count of the corpora 14 358 tweets. A sample of raw corpus data can be seen in Appendix 5, Figure 5.1.

Step 6: the annotation process of the corpus *VerUn*

Once the data had been aggregated, it had to be tagged for the data processing program to understand and select only the necessary data units (nouns and verbs in this case). The part of speech tagger used for the present research was the CLAWS tagger (Online 15), with the C5

tagset (See Appendix 1), because it is 97% accurate (Online 16) and for the present research where noun and verb variations and their frequencies were researched, the results had to be as accurate as possible.

When observing the CLAWS5 tagset, it was concluded that it had 25 verb tags and 4 noun tags, but due to the fact that C5 tagset is the most basic and comprehensive, and, to answer the particular research questions of the present research and provide a more robust, bird's-eye view, the decision to use all the tags in the analysis was made.

In conclusion, the process of data aggregation, compilation and tweet completion proved to be the most time-consuming part of the study. The process of corpus creation was based on the methodology provided by Nelson (2010) and Sinclair (2005). The two core notions, proposed by Sinclair (2005: 15) have been respected as much as possible – the representativeness has been achieved by obtaining tweet texts from different verified politician accounts, with different political beliefs, therefore the sub-corpus *Ver* is representative of all verified political tweets. Accordingly – the sub-corpus *Un* was chosen using multiple hashtags that represent multiple themes, and they are also randomized, therefore the sub-corpus *Un* is also representative of all verified political tweets.

2) Data arrangement

The tweets were downloaded, completed and cleaned, and the nine separate .xls files were put into folder *Raw data* (representing data that is not annotated and is presented in its raw, untouched form). Then, the tweet texts, which were tagged using the CLAWS tagger C5 tagset (See Appendix 1) and were presented in a .txt format were put in a separate folder called *Tagged tweets.* After the tagged files from the folder *Tagged tweets* were put through the program *AntConc*, the *AntConc* results were converted from .txt file to .xls file in order to view them in a more comprehensive way and put in a folder called *AntCont results.* A sample of a tagged text can be seen in Appendix 5, Figure 5.2.

3) Corpus *VerUn* analysis

The corpus structure can be observed in Table 4.1. As it can be seen from the table, sub-corpus *Ver* is comprised of 11 244 tweets, whereas sub-corpus *Un* is comprised of 3114 tweets, and together the *VerUn* corpus is comprised of 14 358 tweets. Sub-corpus *Ver* is comprised of 191 287 words, whereas sub-corpus *Un* is comprised of 65 025 words, making the total amount of words for *VerUn* corpus 256 312.

Table 4.1 **Structure of the *VerUn* corpus**

	Sub-corpus *Ver* of tweet texts of verified accounts	Sub-corpus *Un* of tweet texts of unverified accounts
Number of tweets	11 244	3114
Total number of tweets in *VerUn* corpus	14 358	
Number of words	191 287	65 025
Total number of words in *VerUn* corpus	256 312	
Number of nouns	24 528	13 475
Number of verbs	14 379	10 965
Total number of nouns in *VerUn* corpus	38 003	
Total number of verbs in *VerUn* corpus	25 344	

The corpus analysis program *AntConc* (Online 17) was used to extract the nouns and verbs from the tagged .txt tweet files available in the folder *AntConc results.* A decision was made that all of the noun and verb tags from CLAWS5 tagset (See Appendix 1) were to be used, to maximise the possible nouns and verbs to be extracted from the tweet corpus. The *AntConc* KWIC (keyword in context) sort settings were set to "Level 1 – 1L, Level 2 – 2L and Level 3 – 3L" so that the tagged verbs and nouns could be displayed with the tag after them (*noun_NN0, running_VVG*) to make the process of distinguishing them easier. Then, when *AntCont* had finished pinpointing nouns and verbs separately, the results were saved as a .txt file, which was then converted to .xls file. Initially, the results in .xls file were presented in such a way that the according noun or verb was located in one column and the appropriate tag – in another, therefore Excel formulas were used to draw the according part of speech and their tag together:

108	109	N2 Advance_	VVB [_PUL -_	president_ta	6	1834
109	110	O0 advance_	VVI rare_AJ0	legislation_t	3	1292
110	111	)0 advance_	VVI women_	political_tag	4	846
111	112	&; advance_	VVB their_D	elections_tag	1	520
112	113	1 advances_	VVZ bill_NN	legislation_t	3	1350
113	114	C advances_	VVZ !_SENT -	legislation_t	3	1363
114	115	NN2 advise_	VVB to_TO0	whitehouse_	7	1602
115	116	V0 advised_	VVN not_XX(	potus_tagge	5	584
116	117	HZ advised_	VVN :_PUN #	whitehouse_	7	1279

Figure 4.2 **Example of .xls results before applying Microsoft Excel formulas**

As it can be seen in Figure 4.2, the word 'advise' in line 114 and the appropriate tag 'VVB' were both in separate columns and had to be drawn together in one column. Therefore, three columns were inserted between the columns containing the noun and the tag. In the first column, the formula =TRIM(RIGHT(SUBSTITUTE(B1;" ";REPT(" ";100));100)) was used to cut everything but the last word, which was, in this case, the according part of speech that was needed. In the second column, the formula =TRIM(LEFT(F1;3)) was used to cut everything but the first word, which was, in this case, the tag for the part of the speech. Then, in the third column, the formula =CONCATENATE(C1;D1) was used to draw the part of the

speech and its tag together in one cell. The result of the .xls file after the aforementioned formulas were applied can be seen in Figure 4.3.

108	109	N2 Advance_	Advance_	VVB	Advance_VVB	VVB [_PUL -_	president_ta	6	1834
109	110	O0 advance_	advance_	VVI	advance_VVI	VVI rare_AJ0	legislation_t	3	1292
110	111	)0 advance_	advance_	VVI	advance_VVI	VVI women_	political_tag	4	846
111	112	&; advance_	advance_	VVB	advance_VVB	VVB their_DI	elections_tag	1	520
112	113	1 advances_	advances_	VVZ	advances_VV	VVZ bill_NN	legislation_t	3	1350
113	114	C advances_	advances_	VVZ	advances_VV	VVZ !_SENT -	legislation_t	3	1363
114	115	NN2 advise_	advise_	VVB	advise_VVB	VVB to_TO0	whitehouse_	7	1602
115	116	AV0 advised_	advised_	VVN	advised_VVN	VVN not_XX(	potus_tagge	5	584
116	117	HZ advised_	advised_	VVN	advised_VVN	VVN :_PUN #	whitehouse_	7	1279

Figure 4.3 **Example of .xls results after applying Microsoft Excel formulas**

A sample of the extracted results can be seen in Appendix 5, Figure 5.3. Lastly, the results were saved as .xls files in a separate folder, called *Results.*

4) Frequency measurements

In order to measure the frequencies of nouns and verbs in both sub-corpora (*Ver* and *Un*), an online word frequency program was used (Online 18). The column of each .xls file from the folder *Results* that contained only the nouns and verbs was selected, then pasted into the frequency counter program. Then the results were copied and put in a separate .xls file, and the four separate .xls files called *unverified_nouns, unverified_verbs, verified_nouns* and *verified_verbs* were put in a separate folder called *Results_frequencies.*

5) Reviewing the results

To ensure that the data and corpus are as objective and as unique as possible (each tweet text was presented only once), the data was reviewed in all stages and the data that did not fit the purposes of the present research was deleted. For instance, in the stage of tweet completion, the tweets containing only emoticons or hyperlinks were deleted. Then, during the process of reviewing the *AntConc* results, the instances where the program has found the word that is mistakenly tagged as a part of speech that is other than noun or verb (for instance, *wonder_AJ0*), the instance was deleted.

4.2 Results of the corpus analysis

The research questions raised were as follows: What parts of speech are prominent in the Twitter tweets of verified accounts that are devoted to political themes? What parts of speech are more prominent in the Twitter tweets of unverified accounts that are devoted to political themes?

The results can be seen in the Appendices 2 and 3, which have been compiled into one table, with unverified account verb and noun count, their relative frequencies, and the actual words used, as well as verified account verb and noun count, their relative frequencies and words used. Furthermore, the frequencies of unverified and verified Twitter account nouns and verbs will be looked at, as well as the proper noun frequencies.

The corpus structure can be observed in Table 4.1 in the previous sub-chapter.

4.2.1 Frequency results of the verified Twitter accounts, devoted to political issues

As stated in the third chapter of the present paper, the verified Twitter accounts refer to those that have the blue verification tick symbol next to the Twitter username in their account. For in the sub-corpus *Ver*, 11 244 tweets were compiled.

To get an overview how the people of verified Twitter accounts organise and write their Tweets, several examples were examined, basing the analysis on the framework of Crystal's mini-research (Crystal, 2011: 36-51).

> **Example 1: (@JeffFlake)** Mitt Romney has shown the country what it means to lead with honor, integrity and civility. (Sentence 1) The people of #Utah and the nation need his strong voice, resolve and service now more than ever. (Sentence 2)
> **Example 2: (@SenatorCollins)** Patty's testimony will help us better understand what can be done to moderate the price of prescription drugs without discouraging innovation that helps us live healthier lives. (Sentence 3) I'll continue to advocate for our seniors. (Sentence 4)
> **Example 3: (@JohnBoozman)** Thanks for supporting my resolution to repeal FCC's midnight regulation on broadband providers. (Sentence 5)

As it can be seen from the examples, the Twitter tweet texts that are extracted from the verified accounts are finished sentences that are comprised of moderate difficulty lexicon, that are considerably larger in size than the unverified counterparts (see sub-chapter 4.2.2) – with 95 characters for the shortest example (Example 3) and 220 characters for the longest example (Example 2). Only one contracted form (*I'll – Sentence 4)* and one hashtag (*#Utah-Sentence 2*) are used. There are no grammatical issues in the present examples – all of them have been written in grammatically correct English with no spelling mistakes. Crystal mentions in his study that Twitter has the prompt question of 'What's happening?' and the present examples demonstrate that the politicians update their tweet readers with the information that is relevant to them – what is happening in their lives/campaigns/other political endeavours at the present moment.

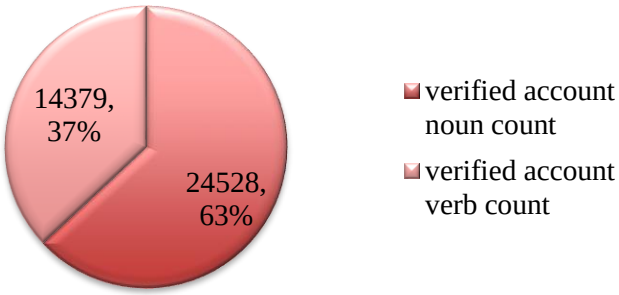

Figure 4.4 **Sub-corpus *Ver* noun and verb frequencies of**

As it can be seen from the Figure 4.4, overall, 23 528 nouns were counted, which makes up 63% of the total part of speech count. Accordingly, 14 379 verbs were counted, which is 37% of the total part of speech count.

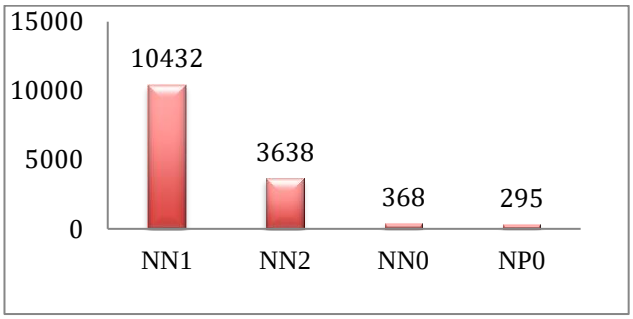

Figure 4.5 **Sub-corpus *Ver* noun tag frequency**

Each noun in the sub-corpus *Ver* was assigned a noun tag according to the CLAWS5 tagset (See Appendix 1), but, since the data had to be reviewed manually and also because a lot of verified account nouns occurred only once and therefore could be regarded as random words, a decision to narrow down the verified account nouns with the assigned tags and put them a separate table (See Appendix 4, Column 3) was made. The noun tag frequencies and corresponding figure (See Figure 4.5) was created based on the data that can be observed in Appendix 4. Together, 14 733 nouns with tags were analysed further.

According to Figure 4.5, the most frequent noun type was with the tag NN1 with the frequency of 10 432, which, according to CLAWS5 tagset (See Appendix 1), corresponds to a singular noun. The second most frequent noun type was with the tag NN2 with the frequency of 3638, and the tag denotes a noun in the plural. Accordingly, the tag NN0 that corresponds to the noun that is neutral in number with the frequency of 368. The tag NP0 that corresponds to the proper noun has the frequency of 295.

Each verb in the sub-corpora *Ver* was assigned a verb tag according to the CLAWS C5 tagset (See Appendix 1). Similarly as with the nouns, the decision to narrow down the verified account verbs was made (See Appendix 4, Column 4). The verb tag frequencies and

corresponding figure (See Figure 4.6) was created based on the data that can be observed in Appendix 4. Together, 11557 verbs with tags were analysed further.

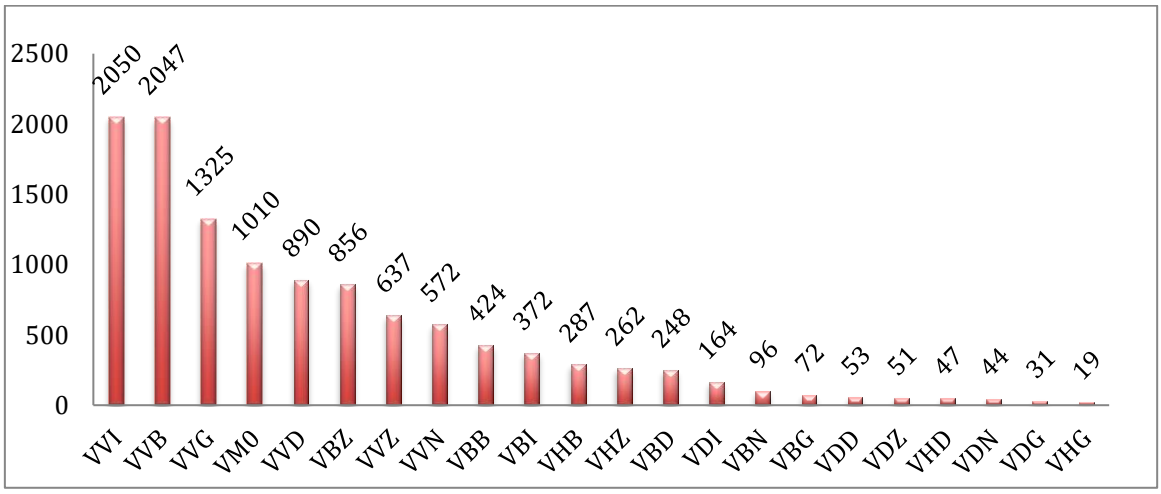

Figure 4.6 **Sub-corpus *Ver* verb tag frequencies**

According to Figure 4.6, the most frequent verb type was VVI with the frequency of 2 050, which, according to CLAWS5 (See Appendix 1), corresponds to infinitive of a lexical verb. The second most frequent verb type was VVB with a frequency of 2 047, which corresponds to the base form of a lexical verb. Third most frequent verb type was VVG with the frequency of 1325, which corresponds to the –ing form of a lexical verb.

4.2.2 Frequency results of the unverified Twitter accounts

As stated in the third chapter of the present bachelor thesis, the unverified Twitter accounts refer to those that do not have the verification blue tick right next to the username. For the sub-corpus *Ver* 3 114 tweets were aggregated, using the hashtags #whitehouse, #president, #elecions and others.

To get an overview how the people of unverified Twitter accounts organise and write their Tweets, several examples were examined further. Similarly as with the sub-corpus *Un*, the examination was done within the framework of Crystal's mini-research (Crystal, 2011: 36-51).

Example 1: (@incorect_p) Crooked gun grabbers! #2a #secondamendment #freedom #molonlabe #america #constitution #firearms #liberalLies #americaeveryday #constitutionalrights (Sentence 1)
Example 2: (@AliceMelott) The kids who created #Facebook had no way to foresee that a decade later it could become a tool for bad actors like #CambridgeAnalytica.. just like the kids who created the #Constitution couldn't imagine the #NRA hijacking their #2ndAmendment after two+ centuries. (Sentence 2)
Example 3: (@MEicEnt) Dont forget, we are not "Trump supporters" but American Patriots!!! (Sentence 3) When #POTUS is out in 6 YEARS what are you/we?! #MAGA #KAG #PATRIOT #QANON #NRA #CONSTITUTION #EducateNotHate (Sentence 4)

Looking at the tweet samples extracted, it can be seen that the length of the tweets is significantly diminished – if the hashtags are not counted, the pure text varies from 3 words (Example 1) to 42 (Example 2). There are several grammatical issues for the unverified tweets with incorrect use of punctuation - dots in second example (sentence 2 - should be three, there are two), lack of apostrophe in the third example (sentence 3) and use of three exclamation marks instead of one (sentence 3). Overall, the unverified account tweet texts contain significantly more hashtags, and it is possibly since the hashtags help to 'reach' the public and therefore, the more hashtags are used, the greater the possibility that the tweet will be noticed. A trend for unverified account users is to compile their whole tweet just with the help of hashtags or use the hash sign in front of words, thus incorporating the hashtag in the tweet text, but at the same time using the hashtag function. For verified account users it is not necessary to use as many hashtags because their follower base is already noticeably larger than that of unverified Twitter users.

The writing style of unverified account users is more colloquial – there are a lot of colloquial phrases (crooked gun grabbers (sentence 1), two+ centuries (the + indicates addition to the existing word 'two' (sentence 2)).

Overall, it can be observed that the unverified Twitter user tweets are more erratic, less thought out and give off the impression of hastiness. The process of composing a tweet is simple and does not take a lot of time, which could potentially be the reason for these erratic unverified account tweets – the tweeters do not think before they tweet.

As it can be seen in Figure 4.7, out of the 3 114 tweets, 13 475 or 55% of total words were nouns, and, accordingly, 10 965 or 45% were verbs.

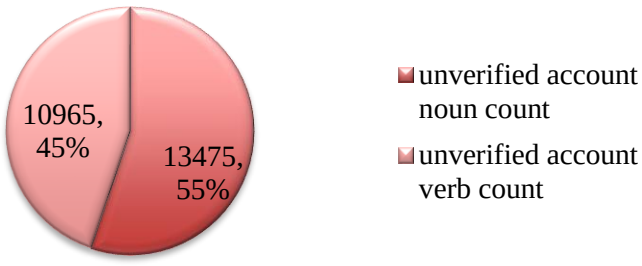

Figure 4.7 **Sub-corpus *Un* noun and verb frequencies**

Each noun in the sub-corpus *Un* was assigned a noun tag according to the CLAWS C5 tagset (See Appendix 1), but, because the data had to be reviewed manually and also because a lot of unverified account nouns occurred only once and therefore could be regarded as random words, the unverified account nouns with the assigned tags were narrowed down and put in a separate table (See Appendix 4, Column 1). The noun tag frequencies and figures

were created based on the data that can be seen in Appendix 4. Together, 7 981 nouns with tags were analysed further.

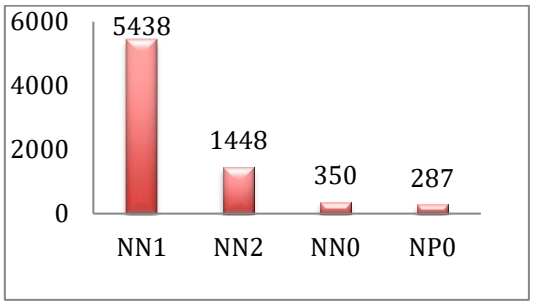

Figure 4.8 **Sub-corpus *Un* noun tag frequencies**

According to Figure 4.8, the most frequent noun type was with the tag NN1 with the frequency of 5 438, which, according to CLAWS5 tagset (See Appendix 1), corresponds to a singular noun. The second most frequent noun type was with the tag NN2 with the frequency of 1 448, which corresponds to a noun in plural. The noun type with the tag NN0 with the frequency of 350 corresponds to a noun that is neutral for the number and noun type with the tag NP0 with the frequency of 287 corresponds to a proper noun.

Each verb in the unverified account sub-corpus was assigned a verb tag according to the CLAWS5 tagset (See Appendix 1), but, due to the fact that the data had to be reviewed manually and also because a lot of the unverified account verbs occurred only once and therefore could be regarded as random words, the decision to narrow down the verbs (See Appendix 4, Column 2) was made. The verb tag frequencies and corresponding figure (See Figure 4.9) was created based on the data can be observed in Appendix 4. Together, 10090 verbs with tags were analysed further.

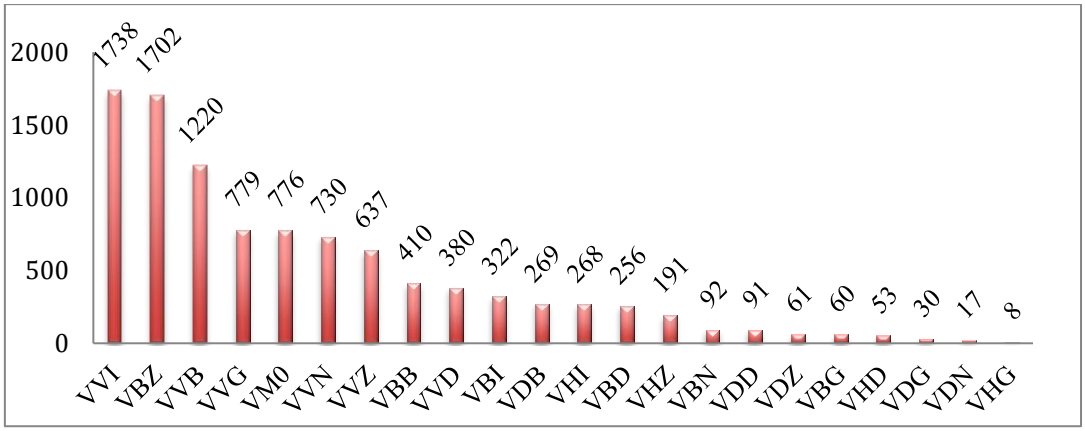

Figure 4.9 **Sub-corpus *Ver* verb tag frequencies**

According to Figure 4.9, the most frequent verb type was with the tag VVI with the frequency of 1 738, which, according to CLAWS5 (See Appendix 1), corresponds to infinitive of a lexical verb. Second most frequent verb type was with the tag VBZ, which corresponds to –s form of the verb 'be', with a frequency of 1702.

4.2.3. Comparison of frequency results of verified and unverified Twitter account proper nouns

Due to the fact that proper nouns are names used for individual persons, organisations, buildings and businesses, the author of the paper decided that it would be valuable to investigate them separately. The proper noun tags were detected, using the *AntConc* program, and then the results were put through word frequency count program. The fourteen proper nouns with the largest frequencies were compiled in two separate charts (See Figure 4.10 and 4.11) for each sub-corpus.

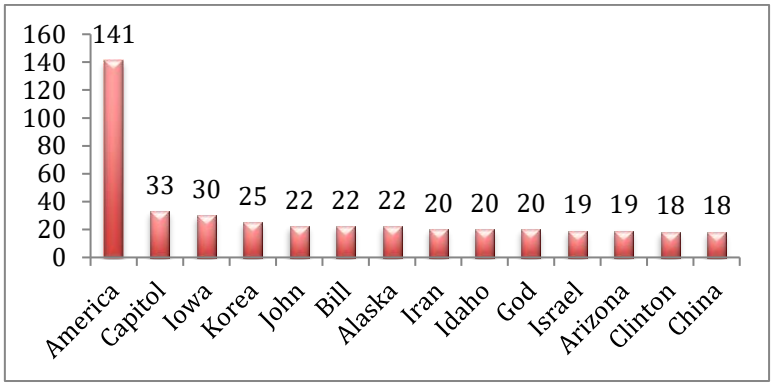

Figure 4.10 **Sub-corpus *Ver* proper noun frequencies**

Regarding the verified account proper noun frequencies (see Figure 4.10), it can be seen that America is the proper noun with the drastically largest frequency of 141. Then comes the proper noun 'Capitol' that refers to the United States Capitol in Washington D.C, the building, with the frequency of 33. In third place in terms of frequency is the United States state Iowa - frequency of 30. Then come Korea (frequency of 25), John (frequency of 22), Bill (frequency of 22), U.S. state of Alaska (frequency of 22), state of Iran (frequency of 20), U.S. state of Idaho (frequency of 20), God (frequency of 20), state of Israel (frequency of 19), U.S. state of Arizona (19), Clinton (frequency of 18), and the state of China (frequency of 18).

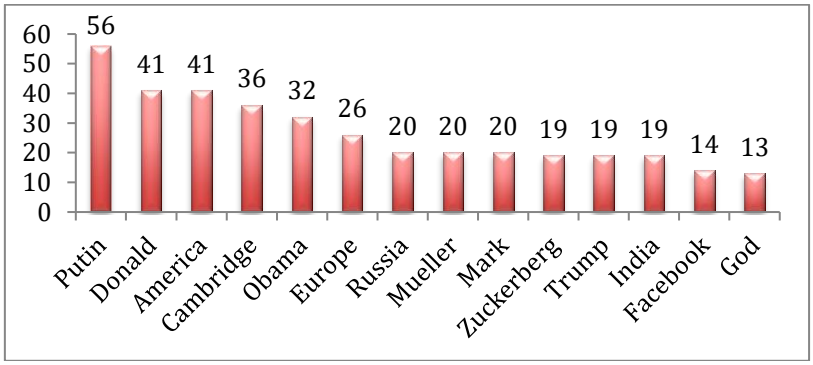

Figure 4.11 **Sub-corpus *Un* proper noun frequencies**

Regarding the unverified account proper noun frequency (See Figure 4.11), the difference between first and second most frequent proper nouns is not that noticeable – in the first place

with the frequency of 50 is the proper noun Putin, in the second and third places with the frequencies of 41 are the proper nouns Donald and America. Then come the proper nouns Cambridge (frequency of 36), Obama (frequency of 32), Europe (frequency of 26), Russia, Mueller, and Mark with the frequency of 20, Zuckerberg, Trump and India with the frequency of 19, Facebook (frequency of 14) and God with the frequency of 13.

4.3. Analysis and Summary of the Results

This sub-chapter deals with the analysis of the results that have been obtained and presented in the previous chapter regarding the corpus analysis and the frequency results.

First of all, the division between verified account nouns and verbs in terms of absolute frequency is 63% nouns (24 528) and 37% verbs (14 379). The unverified account noun and verb distribution is 55% nouns (13 475) and 45% verbs (10 965).

Referring once again to the Chapter 3 of the methodology of the present paper, it can be concluded that both corpora are noticeably different in size (11 244 tweets for verified tweet sub-corpora and 3 114 for unverified tweet sub-corpora), therefore, the decision to use relative frequencies when comparing the two sub-corpora was made.

Having calculated the corresponding noun/verb relative frequencies that displayed the relative frequency per tweet, they were put in separate tables (Appendix 2 and 3). However, the relative frequencies could appear similar for the two sub-corpora when compared, therefore, the decision to include also the absolute frequencies in both of the tables for comparison and clarity was made.

To determine whether the differentiations of frequency are significant, a statistical test of log-likelihood was applied, using the log-likelihood calculated provided by UCREL research centre of Lancaster University (Online 19).

Table 4.12 **Results of the log-likelihood test**

	Subcorpus *Ver*	Subcorpus *Un*	Log-likelihood
Noun relative frequency in the corpus	12.82 %	20.7 %	1898.35
Verb relative frequency in the corpus	7.52 %	16.86 %	3821.94

As it can be seen in Table 4.12, the log-likelihood for both the nouns and verbs is well over 6.63, which, according to University of Lancaster (Online 20), means that the chance of the results happening by chance is approximately 1%, which, in turn, means that the results are 99% significant. It can be observed, that the noun use is 1.16 times higher in frequency for unverified accounts, whereas verb use is 2.24 times higher in frequency.

When looking further into noun tag differentiations for verified (*Ver)* and unverified (*Un*) Twitter tweet sub-corpora (see Figure 4.5 and 4.8) that had been tagged using the

CLAWS5 tagger (See Appendix 1), it can be observed that the most frequent noun tag is the same for both verified and unverified Twitter accounts – NN1. However, it is not surprising that NN1 is the most frequent noun since the primary function of nouns is to function as a subject or object to a verb, as well as a complement of a verb etc. Nouns are primarily singular – it is their 'base' form, and any kind of modifications (plural, neutral or proper noun) are secondary, derived forms.

Nouns with the tag NN1 that, according to CLAWS5 tagset (See Appendix 1) correspond to a noun in singular are both the most frequent noun tags (absolute frequency of 10 432 (relative – 0.93) for verified accounts, absolute frequency of 5 438 (relative – 1.75) for unverified accounts). In the second place there are the nouns with the tag NN2, which correspond to a noun in plural (absolute frequency of 3 638 (relative – 0.32) for verified accounts and absolute frequency of 1448 (relative – 0.46) for unverified accounts). In the third place are the nouns with the tag NN0 that correspond to a noun that is neutral in number with the absolute frequency of 368 (relative – 0.03) for verified accounts and absolute frequency of 350 (relative (0.11) for unverified accounts. The last noun tag is NP0, which corresponds to proper nouns with the absolute frequency of 295 (relative – 0.03) for verified accounts and absolute frequency of 287 (relative – 0.09) for unverified accounts. The summary of verified and unverified Twitter account sub-corpora noun tag relative and absolute frequencies can be seen in Appendix 3.

Based on the frequency results, it can be concluded that the ratio for unverified account NN1 is 1.88 times more frequent and NN2 – 1.43 times. The ratio for nouns that are neutral in number is 3.67 in the unverified accounts and for proper nouns - 3. The ratio of 3 for more proper nouns use could be due to the fact that unverified account users are not cautious when talking about other people, organisations and possible scandals, mentioning specific names, whereas verified account users have to tread with caution due to the fact that they are public figures and are under public scrutiny.

When analysing the verb tag differentiations for verified and unverified Twitter tweet sub-corpora (see Figure 4.4 and 4.7) that had been tagged using the CLAWS5 tagger (See Appendix 1), it can be seen that compared with noun tag frequencies in verified and unverified sub-corpora, verb tag frequencies display more noticeable differences. Although the most frequent verb tag for both verified and unverified verbs is VVI that corresponds to infinitive of a lexical verb (frequency of 2 050 (Rf of 0.182) for verified accounts and 1 738 (Rf of 0.558) for unverified accounts), the frequencies for second, third and further verb tag frequencies differ within verified and unverified sub-corpora.

When comparing the verb frequencies for both verified and unverified sub-corpora, the relative frequency formula was used similarly as when comparing noun frequencies for both sub-corpora (See Appendix 2).

The complete comparison between verified and unverified Twitter tweet sub-corpora verb tag absolute and relative frequencies can be seen in Appendix 2. The tags are colour coded so it could ease seeing the tags that appear in both sub-corpora but are in different frequency positions. The tags that do not appear in one or the other sub-corpora have been left black. As it can be seen in Appendix 2, the verb tags that occur only in verified account sub-corpora are VHB (base form of the verb 'have', relative frequency of 0.026, absolute – 287) and VDI (infinitive of the verb 'do', relative frequency of 0.015, absolute – 164).

Regarding the verb tags that occur only in unverified account sub-corpora, one of them is VDB (base form of the verb 'do', relative frequency of 0.086, absolute – 269) and VHI (infinitive of the verb 'have', relative frequency of 0.096, absolute – 268).

The observation of Figure 4.13 reveals that verbs are used, on average, two times more frequently in unverified account sub-corpora than in the verified account sub-corpora. –s form of the verb 'be' (VBZ) has been used 7.20 times more frequently, and it could be due to the fact that unverified account users tend to talk about and scrutinise other people, organisations, governments or themes (*Open & frequent conversation on this topic is essential, This is how a #global #leader #POTUS should #tweet*). Past form of the verb 'do' (VDD) is 5.80 more frequent, and it could be so due to the fact that unverified account users talk about events that have already happened, for instance, news regarding politics or even what the politicians have done. (*Did #Potus hurt prosecution with comments? #malta #pilatus #not enough- the writing and evidence was all there to see but @MaltaGov @EdwardScicluna @JosephMuscat_JM did nothing).* VVN (past participle of a verb), that has been used 4.59 times more, has the function of conveying an action that has already been completed, possibly to express one's dismay over the action done (*So why hasn't #youtube & newspaper publishers been taken to court?*)

The significantly more frequent use of nouns and verbs in unverified accounts could be due to the fact that verified accounts use a lot of hashtags in order to be noticed, whereas verified account users compose sentences that are larger, therefore, the possibility of different parts of speech is higher. For shorter sentences, there still need to be at least one verb and noun, therefore, the noun-verb frequency is naturally higher in unverified account tweets.

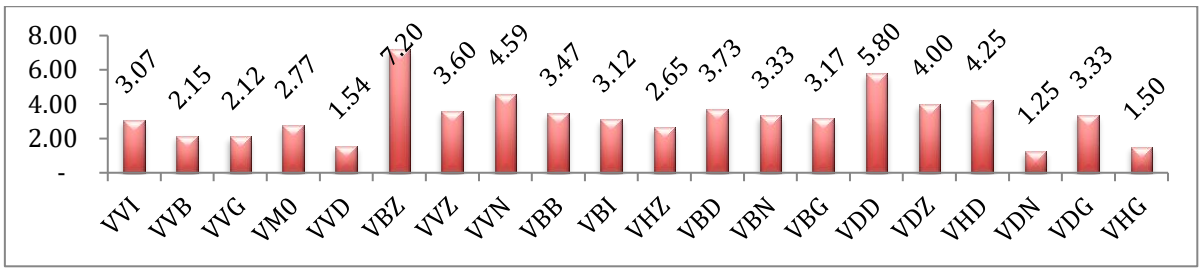

Figure 4.13 **Ratio of verbs in sub-corpus *Un*, comparing with sub-corpus *Ver***

The fourteen most frequent proper nouns of both verified and unverified Twitter sub-corpora were also extracted (See Figures 4.10 and 4.11) because proper nouns are unique and looking at the most frequently used proper nouns could provide an insight to what the most talked people/places/concepts for verified and unverified Twitter users are.

Comparing the two figures (4.10 and 4.11), the verified account proper nouns could potentially be labelled as "comprehensive" and "general" - America being the most frequent (frequency of 141), several U.S states – Iowa (frequency of 30), Alaska (frequency of 22), Idaho (frequency of 20) and Arizona (frequency of 19), some countries – Iran (frequency of 20), Israel (frequency of 20) and China (frequency of 18), as well as some names/surnames – John and Bill (frequency of 22) as well as Clinton and Obama (frequency of 18). Finally, there are some completely unique proper nouns like Capitol (frequency of 33) and God (frequency of 20).

Unverified account proper nouns are more interesting with the fact that they are completely different from the fourteen most frequent proper nouns in the verified account sub-corpora. By looking at the unverified account proper nouns, it can be seen that they reflect the real situation and the current events that the news are reporting.

For instance, the most frequent proper noun for the unverified account sub-corpora is Putin with the frequency of 56, which reflects how the current president of the Russian Federation – Vladimir Putin, is in the centre of public scrutiny – due to his decisions regarding foreign policy, the elections in Russian Federation and other issues. This is tied together with the proper noun Russia that has the frequency of 20.

The second most frequent proper noun within the 14 most frequent proper nouns examined, is the proper noun Donald with the frequency of 41, which is tied together with the proper noun Trump with the frequency of 19. Both of these proper nouns refer to the current president of the USA – Donald Trump, and he has been in the centre of attention due to the controversies connected with the process of his election as a president, as well as his unprofessional remarks and actions while fulfilling his duties as the president of the USA.

Comparing the verified account proper noun frequencies, where the proper noun America was distinctly in the first place with the frequency of 141, in the unverified account sub-corpora the proper noun *America* is only in the third place with the frequency of 41.

Then comes the proper noun Cambridge with the frequency of 36, which most likely is related to the *Cambridge Analytica* scandal (Online 21) where the political consulting firm *Cambridge Analytica* illegally obtained data from approximately 50 million Facebook users in order to influence the result of the U.S. elections in the result of which Donald Trump won. Moreover, it is speculated that Russia was also involved into this scandal, therefore, it seems only logical that *Cambridge* is one of the most frequent proper nouns used, but only within the sub-corpora of unverified Twitter accounts.

Closely tied with the *Cambridge Analytica* scandal are the proper nouns *Mark* (frequency of 20), *Zuckerberg* (frequency of 19) and *Facebook* (frequency of 14) purely because Mark Zuckerberg is the co-founder of the social media site Facebook, therefore it is evident that these three proper nouns appear in connection with the *Cambridge Analytica* data-leak scandal.

The proper noun that is also connected with elections, Russia and Donald Trump is the proper noun *Mueller* (frequency of 20), which is the surname of the American attorney Robert Mueller who is overseeing the ongoing investigation (Online 22) regarding Russia's interference in the 2016 U.S. elections.

The proper noun *Europe* (frequency of 26) is also among the most frequent proper nouns within the sub-corpora of unverified accounts, probably because European Union is a large power that is usually put opposite to the USA.

Among the most frequent proper nouns is also the surname of the previous president of the U.S. – Obama (frequency of 32), the country of India (frequency of 19) and in the last place – God with the frequency of 13.

In conclusion, it can be said that when analysing the frequencies of verified and unverified account sub-corpora, the ratio of nouns to verbs is 63%: 37% for verified accounts and 55%: 45% for unverified accounts.

The distribution of noun tag frequencies may appear similar between the two sub-corpora (See Appendix 3), but the relative frequency of nouns in unverified accounts ranges from being 1.42 to 3.67 times higher than in verified accounts.

The distribution of different verb tag frequency also ranges from 1.25-7.20 times higher than in verified accounts.

The fourteen most frequent proper nouns are completely different between the two sub-corpora.

Conclusions

The present research deals with the creation and analysis of the *VerUn* corpus, which comprises of two sub-corpora – *Ver* for verified account tweets that are devoted to political themes and *Un* for unverified account tweets that are devoted to political issues.

The goal of the thesis was to investigate the variations of nouns and verbs between tweets, devoted to political issues in the verified as well as unverified Twitter accounts.

The research questions are as follows: What parts of speech are prominent in the Twitter tweets of verified accounts that are devoted to political themes? What parts of speech are more prominent in the Twitter tweets of unverified accounts that are devoted to political themes?

The methodology of the present thesis includes literature review as well as corpus-based quantitative research using part-of-speech tagger CLAWS and the corpus analysis toolkit *AntConc*.

The literature review revealed that the contemporary division of parts of speech groups the words into lexical, functional and interjections. The analysis of previous research of tweets as texts revealed that linguistic research connected with tweet texts can be arranged into four trends: determining the geolocation, sentiment analysis, compiling a mini-corpora and the multidimensional analysis. Further, the overview of features of microblogging as a messaging system revealed three distinct characteristics of Twitter: ambient awareness, push-push-pull communication model, as well as voyeurism and virtual exhibitionism. Regarding the communicative context of Twitter texts, the mode can be regarded as 'written speech' and thus, tweeting can be regarded as mass communication. Lastly, the corpus-based quantitative research method deals with the absolute and relative frequency analysis in general or specialised corpora that can be either annotated or unannotated.

The analysis of the specialised corpus *VerUn* of verified and unverified Twitter account texts, created specifically for the present research revealed the following conclusions:

1. The verb-noun distribution between verified and unverified Twitter tweet text sub-corpora differs. For verified accounts, the division is 63% nouns and 37% verbs, whereas for unverified accounts, the division is 55% nouns and 45% verbs.
2. The most common noun tags are similar for verified and unverified Twitter tweet text sub-corpora frequency, but the frequency of noun tags proves to be at least two times larger in unverified accounts. The frequency of verb tags is at least three times larger in unverified accounts.

3. The verbs tags assigned for each sub-corpus was approximately 5 times larger than the number of noun tags assigned, therefore, there were minor verb tag frequency discrepancies between the two sub-corpora.
4. The most noticeable differences were revealed in the closer analysis of proper nouns – the fourteen most frequent verified account proper nouns do not mirror the topical questions of political themes and issues, whereas the fourteen most frequent unverified account proper nouns revealed the topical questions of political themes and issues at the present time.

To conclude, although for both verified and unverified accounts, the nouns were predominant when looking at the absolute frequencies of them, the relative frequencies revealed that nouns are at least 2 times more frequent and verbs are at least 3 times more frequent in unverified accounts of Twitter, which answers the research question raised.

In order to expand the study further, the sub-corpora of unverified Twitter accounts could be expanded to meet the volume of verified Twitter account sub-corpora. Additionally, other parts of speech could be included into the analysis, not only verbs and nouns.

To sum up, the present research has looked into an underinvestigated field of Twitter tweet corpus analysis. The corpus creation was a complicated process, however, the analysis presents relevant results regarding verb and noun frequency in the corpus created, as well as noun and verb tag variations across both verified and unverified Twitter accounts. Overall, the present work has been an interesting insight into processes of creating a corpus, annotating and analysing it and provides a relevant conclusion regarding the structure of the microblogging system Twitter.

Theses

1. The contemporary parts of speech can be grouped into lexical words that include verbs, nouns, adjectives and adverbs, function words that include pronouns, prepositions, conjunctions and articles, as well as interjections.
2. Linguistic research of tweets as texts displays several approaches, such as determining the geolocation and language usage variations, sentiment analysis, the compilation and qualitative analysis of a mini-corpora, as well as multidimensional analysis.
3. Twitter is a microblogging system that allows its users to share brief blasts of information to friends and followers using 280 characters.
4. Corpus-based quantitative analysis deals with the frequency of various parts of speech in the chosen corpus.
5. Specialised corpus is a corpus compiled for specific research purposes, and it can be either unannotated or annotated.
6. Corpus *VerUn* can be defined as a specialised, POS-annotated corpus.
7. The tweet extraction programmes have not been updated to aggregate more than 140 characters in tweets, therefore the texts have to be manually completed to answer the particular research questions.
8. CLAWS is the part of speech tagger that has 97% accuracy rate when tagging corpora and C5 is a reliable and comparatively smaller tagset, which allows focusing on POS discussion in this study.
9. *AntConc* is the concordance toolkit that is used for corpus analysis.
10. The verb-noun distribution between verified (*Ver*) and unverified (*Un*) Twitter tweet text sub-corpora differs - for verified accounts, the division is 63% nouns and 37% verbs, whereas for unverified accounts, the division is 55% nouns and 45% verbs.
11. The most frequent noun type for both sub-corpora was NN1 (noun in singular), and the most frequent verb type for both sub-corpora was VVI (infinitive of a lexical verb), but the frequency of noun tags in unverified sub-corpus *Un* was approximately 2 times higher and the frequency of verbs – 3 times.
12. The largest differences were distinguished when comparing the most frequent proper nouns within the two sub-corpora.

References

1. Biber, D. (1988) *Variation across Speech and Writing*. Cambridge: Cambridge University Press.
2. Biber, D. (2003). Variation among University Spoken and Written Registers: A New Multidimensional Analysis. *Language and Computers, 46,* 47-70.
3. Biber, D., & Kurjian, J. (2007). Towards a taxonomy of web registers and text types: a multi-dimensional analysis. 109-131.
4. Biber, D., Conrad, S., & Reppen, R. (2012). *Corpus linguistics: Investigating language structure and use*. Cambridge [u.a.: Cambridge Univ. Press.
5. Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E. (1999) *Longman Grammar of Spoken and Written English.* London: Pearson Education Limited.
6. Ceron, A., Curini, L. and Iacus, S. (2014). Using Sentiment Analysis to Monitor Electoral Campaigns. *Social Science Computer Review*, 33(1), pp.3-20.
7. Crystal, D. (2011). *Internet Linguistics: A Student Guide*. Internet Linguistics: Routledge
8. DeVoe, K. (2009). Burst of information: Microblogging. *The Reference Librarian*, 50, pp.212-214.
9. Dionysius, Bécares Botas, V. and Dionysius (2002). *Gramática ; Comentarios antiguos*. Madrid: Editorial Gredos.
10. Downing, A., Locke, P. (2006) *English Grammar: A University Course Second Edition.* NY: Routledge.
11. Gatto, M. (2014). *The web as corpus: Theory and practice*. London.: Bloomsbury.
12. Graham, M., Hale, S. and Gaffney, D. (2014). Where in the World Are You? Geolocation and Language Identification in Twitter. *The Professional Geographer*, 66(4), pp.568-578.
13. Grieve J., Biber D., Friginal E., Nekrasova T. (2010) Variation Among Blogs: A Multi-dimensional Analysis. In: Mehler A., Sharoff S., Santini M. (eds) Genres on the Web. Text, Speech and Language Technology, vol 42. Springer, Dordrecht
14. Hardie, A., McEnery, T., & Baker, Paul. (2006). *Glossary of Corpus Linguistics, A. Glossaries in Linguistics*. Edinburgh University Press.
15. Huddleston, R., & Pullum, G. K. (2002). *A student's introduction to English grammar*. Cambridge: Cambridge University Press.
16. Hudson, R. A. (2010). *An introduction to word grammar*. Cambridge: Cambridge University Press.

17. Kaplan, A. M., & Haenlein, M. (2011). The early bird catches the news: Nine things you should know about micro-blogging. *Business Horizons,* 54(2), 105-113.
18. Kennedy, G. (2014). *Introduction to Corpus Linguistics*. New York: Routledge.
19. Lüdeling, A., & Kytö, M. (2009). *Corpus linguistics: An international handbook.* Berlin: W. de Gruyter.
20. MacEnery, T., & Wilson, A. (2011). *Corpus linguistics: An introduction*. Edinburgh: University Press.
21. Matilal, B. (1990). *The word and the world*. New Delhi: Oxford University Press.
22. McEnery, T., Hardie, A. (2014). *Corpus linguistics: Method, theory and practice*. Cambridge: Cambridge University Press.
23. O'Keeffe, A., McCarthy, M. (2010) *Handbook of Corpus Linguistics.* Routledge Ltd.
24. Öztürk, N. and Ayvaz, S. (2018). Sentiment analysis on Twitter: A text mining approach to the Syrian refugee crisis. *Telematics and Informatics*, 35(1), pp.136-147.
25. Pang, B., & Lee, L. (2008). *Opinion mining and sentiment analysis*. Boston, Mass: Now Publishers.
26. Panini. (1986). *Sanskrit grammar (Ashtadhyayj) in Sanskrit & English.* New York: Orientalia.
27. Postmes, T., & Brunsting, S. (2002). Collective Action in the Age of the Internet: Mass Communication and Online Mobilization. *Social Science Computer Review, 20,* 3, 290-301.
28. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985) *A Comprehensive Grammar of the English Language.* London: Longman Group Limited.
29. Reyes, A., Rosso, P., & Veale, T. (March 01, 2013). A multidimensional approach for detecting irony in Twitter. *Language Resources and Evaluation, 47,* 1, 239-268.
30. Sedley, D. N. (2007). *Plato's Cratylus*. Cambridge, U.K: Cambridge University Press.
31. Simpson, J. (2002). Computer-Mediated Communication. *Elt Journal, 56,* 4, 414-15.
32. Sinclair, J. (2005) Corpus and Text - Basic Principles. In M. Wynne (ed.) *Developing Linguistic Corpora: a Guide to Good Practice* (pp. 1-16). Oxbow Books.
33. Titak, A. and Roberson, A. (2013). Dimensions of web registers: an exploratory multi-dimensional comparison. *Corpora*, 8(2), pp.235-260.
34. Wang, Y. and Fikis, D. (2017). Common Core State Standards on Twitter: Public Sentiment and Opinion Leaders. *Educational Policy*, pp.1-34.
35. Wikström, P. (2017). *I tweet like I talk: Aspects of speech and writing on Twitter*. Karlstad: Faculty of Arts and Social Sciences, English, Karlstads universitet.

36. Zappavigna, M. (2015). *Discourse of twitter and social media: [how we use language to create affiliation on the web]*. London: Bloomsbury.
37. Zhao, N. and Cao, G. (2017). Quantifying and visualizing language diversity of Hong Kong using Twitter. *Environment and Planning A*, 49(12), pp.2698-2701.

Internet resources:

1. Carlson, N. (2018). *The Real History of Twitter*. Business Insider. Available from http://www.businessinsider.com/how-twitter-was-founded-2011-4 [Accessed March 16, 2018]
2. Tweet location FAQs. Available from https://help.twitter.com/en/safety-and-security/tweet-location-settings [Accessed March 21, 2018]
3. Multi-dimensional analysis. Available from http://idrd-bham.info/idrd/?page_id=454 [Accessed March 18, 2018]
4. Twitter | History, Description, & Uses. (2018). Available from https://www.britannica.com/topic/Twitter [Accessed March 16, 2018].
5. Facebook vs. Twitter: Which is best for your brand? Available from https://sproutsocial.com/insights/facebook-vs-twitter/ [Accessed March 10, 2018].
6. 2016 Global Digital Future in Focus. Available from https://www.comscore.com/Insights/Presentations-and-Whitepapers/2016/2016-Global-Digital-Future-in-Focus [Accessed April 16, 2018]
7. World Leaders on Twitter - Adoption Stagnates Even as Follower Base Explodes. Available from https://www.prnewswire.com/news-releases/world-leaders-on-twitter--adoption-stagnates-even-as-follower-base-explodes-300208802.html [Accessed April 19, 2018]
8. Beginner's guide to Twitter. Available from https://michaelhyatt.com/the-beginners-guide-to-twitter/ [Accessed May 7, 2018]
9. Twitter Data Analysis: An investor's perspective. Available from https://techcrunch.com/2009/10/05/twitter-data-analysis-an-investors-perspective-2/ [Accessed May 2, 2018]
10. Twitter statistics. Available from http://www.internetlivestats.com/twitter-statistics/ [Accessed May 2, 2018]
11. The British National Corpus. Available from https://corpus.byu.edu/bnc/ [Accessed April 15, 2018]

12. Richard Xiao. *Making statistical claims.* Available from http://www.lancaster.ac.uk/fass/projects/corpus/ZJU/xpresentations/session%205.ppt [Accessed May 2, 2018]
13. Twitter tweet data archive. Available from https://archive.org/search.php?query=collection%3Atwitterstream&sort=-publicdate [Accessed May 5, 2018]
14. The Chorus Project Downloads. Available from http://chorusanalytics.co.uk/downloads/ [Accessed March 10, 2018]
15. CLAWS POS tagger. Available from http://ucrel.lancs.ac.uk/claws/trial.html [Accessed March 2, 2018]
16. Schneider, G., Hundt, M., & Oppliger, R. (2016). *Part-Of-Speech in Historical Corpora: Tagger Evaluation and Ensemble Systems on ARCHER.* Available from https://www.linguistics.rub.de/konvens16/pub/33_konvensproc.pdf [Accessed May 7, 2018]
17. Laurence Anthony's AncConc. Available from http://www.laurenceanthony.net/software/antconc/ [Accessed March 20, 2018]
18. Word frequency counter – WriteWords. Available from http://www.writewords.org.uk/word_count.asp [Accessed March 10, 2018]
19. Log-likelihood and effect size calculator. Available from http://ucrel.lancs.ac.uk/llwizard.html [Accessed May 15, 2018]
20. Testing for Significance: log-likelihood. Available from https://www.lancaster.ac.uk/fss/courses/ling/corpus/blue/l08_4.htm [Accessed May 15, 2018]
21. Nytimes.com. (2018). *Cambridge Analytica and Facebook: The Scandal and the Fallout So Far*. Available from: https://www.nytimes.com/2018/04/04/us/politics/cambridge-analytica-scandal-fallout.html [Accessed April 19, 2018].
22. Drury, C. (2018). *Trump's chilling suggestion over 'illegal' Mueller investigation: 'Does everybody know what that means?'*. The Independent. Available at: https://www.independent.co.uk/news/world/americas/donald-trump-robert-mueller-special-counsel-russia-investigation-illegal-twitter-a8315636.html [Accessed April 19, 2018].

Appendix 1

CLAWS5 tagset

***tags used**

Table 1.1 **CLAWS5 tagset**

AJ0	adjective (unmarked) (e.g. GOOD, OLD)
AJC	comparative adjective (e.g. BETTER, OLDER)
AJS	superlative adjective (e.g. BEST, OLDEST)
AT0	article (e.g. THE, A, AN)
AV0	adverb (unmarked) (e.g. OFTEN, WELL, LONGER, FURTHEST)
AVP	adverb particle (e.g. UP, OFF, OUT)
AVQ	wh-adverb (e.g. WHEN, HOW, WHY)
CJC	coordinating conjunction (e.g. AND, OR)
CJS	subordinating conjunction (e.g. ALTHOUGH, WHEN)
CJT	the conjunction THAT
CRD	cardinal numeral (e.g. 3, FIFTY-FIVE, 6609) (excl ONE)
DPS	possessive determiner form (e.g. YOUR, THEIR)
DT0	general determiner (e.g. THESE, SOME)
DTQ	wh-determiner (e.g. WHOSE, WHICH)
EX0	existential THERE
ITJ	interjection or other isolate (e.g. OH, YES, MHM)
NN0	**noun (neutral for number) (e.g. AIRCRAFT, DATA) ***
NN1	**singular noun (e.g. PENCIL, GOOSE) ***
NN2	**plural noun (e.g. PENCILS, GEESE) ***
NP0	**proper noun (e.g. LONDON, MICHAEL, MARS) ***
NULL	the null tag (for items not to be tagged)
ORD	ordinal (e.g. SIXTH, 77TH, LAST)
PNI	indefinite pronoun (e.g. NONE, EVERYTHING)
PNP	personal pronoun (e.g. YOU, THEM, OURS)
PNQ	wh-pronoun (e.g. WHO, WHOEVER)
PNX	reflexive pronoun (e.g. ITSELF, OURSELVES)
POS	the possessive (or genitive morpheme) 'S or '
PRF	the preposition OF
PRP	preposition (except for OF) (e.g. FOR, ABOVE, TO)
PUL	punctuation - left bracket (i.e. (or [ )
PUN	punctuation - general mark (i.e. . ! , : ; - ? ...)
PUQ	punctuation - quotation mark (i.e. ` ' ")
PUR	punctuation - right bracket (i.e.) or])
TO0	infinitive marker TO
UNC	"unclassified" items which are not words of the English lexicon
VBB	**the "base forms" of the verb "BE" (except the infinitive), i.e. AM, ARE ***
VBD	**past form of the verb "BE", i.e. WAS, WERE ***

VBG	**-ing form of the verb "BE", i.e. BEING ***
VBI	**infinitive of the verb "BE" ***
VBN	**past participle of the verb "BE", i.e. BEEN ***
VBZ	**-s form of the verb "BE", i.e. IS, 'S ***
VDB	**base form of the verb "DO" (except the infinitive) ***
VDD	**past form of the verb "DO", i.e. DID ***
VDG	**-ing form of the verb "DO", i.e. DOING ***
VDI	**infinitive of the verb "DO" ***
VDN	**past participle of the verb "DO", i.e. DONE ***
VDZ	**-s form of the verb "DO", i.e. DOES ***
VHB	**base form of the verb "HAVE" (except the infinitive), i.e. HAVE ***
VHD	**past tense form of the verb "HAVE", i.e. HAD, 'D ***
VHG	**-ing form of the verb "HAVE", i.e. HAVING ***
VHI	**infinitive of the verb "HAVE" ***
VHN	**past participle of the verb "HAVE", i.e. HAD ***
VHZ	**-s form of the verb "HAVE", i.e. HAS, 'S ***
VM0	**modal auxiliary verb (e.g. CAN, COULD, WILL, 'LL) ***
VVB	**base form of lexical verb (except the infinitive)(e.g. TAKE, LIVE) ***
VVD	**past tense form of lexical verb (e.g. TOOK, LIVED) ***
VVG	**-ing form of lexical verb (e.g. TAKING, LIVING) ***
VVI	**infinitive of lexical verb ***
VVN	**past participle form of lex. verb (e.g. TAKEN, LIVED) ***
VVZ	-s form of lexical verb (e.g. TAKES, LIVES)
XX0	the negative NOT or N'T
ZZ0	alphabetical symbol (e.g. A, B, c, d)

Appendix 2

Verified and unverified verb tag relative and absolute frequencies

Table 2.1 **Verified and unverified verb tag relative and absolute frequencies**

Number by frequency (from most to least)	Verb tag (denomination of the tag)	Verified account verb tag relative frequency	Verified account verb tag absolute frequency	Verb tag (denomination of the tag)	Unverified account verb tag relative frequency	Unverified account verb tag absolute frequency
1.	**VVI (infinitive of a lexical verb)**	0.182	2 050	**VVI (infinitive of a lexical verb)**	0.558	1 738
2.	**VVB (base form of a lexical verb)**	0.182	2 047	**VBZ (-s form of the verb 'be')**	0.547	1 702
3.	**VVG (-ing form of a lexical verb)**	0.118	1 325	**VVB (base form of a lexical verb)**	0.392	1 220
4.	**VM0 (modal auxiliary verb)**	0.090	1 010	**VVG (-ing form of a lexical verb)**	0.250	779
5.	**VVD (past tense form of a lexical verb)**	0.079	890	**VM0 (modal auxiliary verb)**	0.249	776
6.	**VBZ (-s form of the verb 'be')**	0.076	856	**VVN (past participle of lexical verb)**	0.234	730
7.	**VVZ (-s form of lexical verb)**	0.057	637	**VVZ (-s form of lexical verb)**	0.205	637
8.	**VVN (past participle of lexical verb)**	0.051	572	**VBB (the 'base forms' of the verb 'be')**	0.132	410
9.	**VBB (the 'base forms' of the verb 'be')**	0.038	424	**VVD (past tense form of a lexical verb)**	0.122	380
10.	**VBI (infinitive of the verb 'be')**	0.033	372	**VBI (infinitive of the verb 'be')**	0.103	322
11.	**VHB (base form of the verb 'have')**	0.026	287	**VDB (base form of the verb 'do')**	0.086	269

12.	**VHZ (-s form of the verb 'have')**	0.023	262	**VHI (infinitive of the verb 'have')**	0086	268
13.	**VBD (past form of the verb 'be')**	0.022	248	**VBD (past form of the verb 'be')**	0,082	256
14.	**VDI (infinitive of the verb 'do')**	0.015	164	**VHZ (-s form of the verb 'have')**	0.061	191
15.	**VBN (past participle of the verb 'be')**	0.009	96	**VBN (past participle of the verb 'be')**	0.030	92
16.	**VBG (-ing form of the verb 'be')**	0.006	72	**VDD (past form of the verb 'do')**	0.029	91
17.	**VDD (past form of the verb 'do')**	0.005	53	**VDZ (-s form of the verb 'do')**	0.020	61
18.	**VDZ (-s form of the verb 'do')**	0.005	51	**VBG (-ing form of the verb 'be')**	0.019	60
19.	**VHD (past tense form od the verb 'have')**	0.004	47	**VHD (past tense form od the verb 'have')**	0.017	53
20.	**VDN (past participle of the verb 'do')**	0.004	44	**VDG (-ing form of the verb 'do')**	0.010	30
21.	**VDG (-ing form of the verb 'do')**	0.003	31	**VDN (past participle of the verb 'do')**	0.005	17
22.	**VHG (-ing form of the verb 'have')**	0.002	19	**VHG (-ing form of the verb 'have')**	0.003	8

Appendix 3

Verified and unverified noun tag relative and absolute frequencies

Table 3.1 **Verified and unverified noun tag relative and absolute frequencies**

Number by frequency (from most to least)	Noun tag (denomination of the tag)	Verified account noun tag relative frequency	Verified account noun tag absolute frequency	Noun tag (denomination of the tag)	Unverified account noun tag relative frequency	Unverified account noun tag absolute frequency
1.	NN1	0.93	10 432	NN1	1.75	5 438
2.	NN2	0.32	3 638	NN2	0.46	1 448
3.	NN0	0.03	368	NN0	0.11	350
4.	NP0	0.03	295	NP0	0.09	287

Appendix 4

The results of the corpus analysis

Table 4.1 **The results of corpus analysis**

Unverified account nouns		**Unverified account verbs**		**Verified account nouns**		**Verified account verbs**	
Total count : 13 475		**Total count: 10 965**		**Total count: 24 528**		**Total count: 14 379**	
trump_NP0	*171*	is_VBZ	*846*	tax_NN1	*230*	is_VBZ	*856*
president_NN1	*122*	are_VBB	*389*	bill_NN1	*211*	will_VM0	*488*
people_NN0	*121*	be_VBI	*322*	senate_NN1	*184*	are_VBB	*392*
bill_NN1	*81*	do_VDB	*269*	people_NN0	*170*	be_VBI	*372*
election_NN1	*73*	have_VHI	*268*	americans_NN2	*151*	have_VHB	*287*
world_NN1	*65*	will_VM0	*202*	america_NN1	*149*	has_VHZ	*262*
elections_NN2	*65*	has_VHZ	*191*	families_NN2	*144*	was_VBD	*185*
time_NN0	*64*	was_VBD	*190*	health_NN1	*135*	can_VM0	*174*
putin_NN1	*63*	can_VM0	*174*	congress_NN1	*125*	do_VDI	*164*
house_NN1	*63*	would_VM0	*138*	state_NN1	*121*	should_VM0	*117*
country_NN1	*62*	get_VVI	*112*	congratulations_NN2	*120*	working_VVG	*110*
government_NN1	*60*	should_VM0	*96*	trump_NP0	*118*	get_VVB	*105*
news_NN1	*57*	know_VVI	*92*	time_NN1	*114*	been_VBN	*96*
legislation_NN1	*57*	did_VDD	*91*	year_NN1	*112*	must_VM0	*95*
data_NN0	*54*	been_VBN	*90*	president_NN1	*110*	make_VVI	*95*
office_NN1	*50*	think_VVI	*79*	jobs_NN2	*106*	would_VM0	*91*
state_NN1	*48*	need_VVI	*76*	years_NN2	*105*	enjoyed_VVD	*86*
day_NN1	*46*	could_VM0	*70*	work_NN1	*95*	read_VVN	*81*
law_NN1	*45*	see_VVI	*67*	country_NN1	*95*	keep_VVB	*77*
days_NN2	*45*	were_VBD	*66*	act_NN1	*95*	continue_VVB	*77*
way_NN1	*43*	make_VVI	*62*	day_NN1	*93*	see_VVB	*75*
media_NN0	*43*	does_VDZ	*61*	care_NN1	*90*	hearing_VVG	*75*
thing_NN1	*42*	being_VBG	*60*	law_NN1	*88*	know_VVI	*73*
congress_NN1	*42*	read_VVN	*53*	congrats_NN0	*84*	need_VVI	*72*
year_NN1	*41*	had_VHD	*53*	economy_NN1	*83*	being_VBG	*72*
donald_NN1	*41*	going_VVG	*52*	students_NN2	*77*	take_VVI	*71*
america_NN1	*41*	want_VVB	*49*	week_NN1	*75*	work_VVI	*70*
security_NN1	*37*	take_VVB	*49*	reform_NN1	*75*	watch_VVI	*64*
party_NN1	*37*	go_VVI	*49*	family_NN1	*75*	support_VVI	*64*
cambridge_NN1	*36*	says_VVZ	*46*	nation_NN1	*70*	were_VBD	*63*
parties_NN2	*35*	say_VVB	*44*	states_NN2	*67*	thank_VVB	*62*
obama_NN1	*35*	help_VVI	*42*	world_NN1	*66*	look_VVI	*58*
facebook_NN1	*35*	made_VVN	*41*	school_NN1	*66*	passed_VVN	*57*
business_NN1	*33*	let_VVI	*39*	service_NN1	*65*	discuss_VVB	*57*
states_NN2	*32*	vote_VVI	*38*	community_NN1	*64*	did_VDD	*53*
years_NN2	*30*	said_VVD	*36*	workers_NN2	*62*	does_VDZ	*51*
money_NN1	*30*	got_VVN	*36*	veterans_NN2	*61*	let_VVI	*49*
gun_NN1	*30*	stop_VVI	*35*	job_NN1	*61*	webcast_VVI	*48*
call_NN1	*30*	must_VM0	*35*	security_NN1	*60*	provide_VVI	*48*
analytica_NN1	*30*	find_VVI	*35*	opportunity_NN1	*60*	want_VVI	*47*
school_NN1	*29*	may_VM0	*34*	thanks_NN2	*59*	had_VHD	*47*

senate_NN1	*28*	wants_VVZ	*32*	house_NN1	*59*	protect_VVB	*45*
life_NN1	*26*	put_VVI	*32*	budget_NN1	*59*	tune_VVI	*44*
women_NN2	*26*	keep_VVB	*32*	news_NN1	*58*	needs_VVZ	*44*
wall_NN1	*25*	tell_VVB	*31*	government_NN1	*58*	done_VDN	*44*
person_NN1	*25*	protect_VVB	*31*	democrats_NN2	*58*	appreciate_VVB	*44*
rights_NN2	*25*	work_VVI	*30*	women_NN2	*56*	making_VVG	*43*
power_NN1	*23*	win_VVB	*30*	life_NN1	*56*	made_VVN	*43*
place_NN1	*23*	use_VVB	*30*	way_NN1	*54*	come_VVI	*43*
number_NN1	*23*	doing_VDG	*30*	issues_NN2	*52*	serve_VVB	*42*
job_NN1	*23*	congratulate_VVB	*30*	hearing_NN1	*52*	go_VVB	*42*
vote_NN1	*23*	thank_VVB	*29*	members_NN2	*51*	could_VM0	*42*
staff_NN0	*22*	support_VVI	*29*	businesses_NN2	*51*	said_VVD	*41*
politics_NN1	*22*	needs_VVZ	*28*	legislation_NN1	*50*	meet_VVB	*40*
leader_NN1	*22*	give_VVB	*28*	history_NN1	*49*	join_VVB	*40*
hours_NN2	*22*	come_VVB	*28*	support_NN1	*48*	fight_VVI	*40*
funding_NN1	*22*	making_VVG	*27*	morning_NN1	*47*	give_VVB	*39*
russia_NN1	*22*	call_VVI	*27*	money_NN1	*47*	coming_VVG	*39*
man_NN1	*21*	watch_VVI	*26*	taxes_NN2	*45*	going_VVG	*37*
democracy_NN1	*21*	trying_VVG	*25*	program_NN1	*45*	live_VVI	*36*
zuckerberg_NN1	*21*	sign_VVI	*25*	funding_NN1	*45*	meeting_VVG	*35*
work_NN1	*20*	join_VVB	*25*	education_NN1	*45*	helping_VVG	*35*
use_NN1	*20*	getting_VVG	*25*	leadership_NN1	*44*	check_VVI	*34*
problem_NN1	*20*	like_VVI	*24*	leaders_NN2	*44*	act_VVI	*34*
part_NN1	*20*	fix_VVI	*24*	freedom_NN1	*44*	talk_VVI	*33*
mueller_NN1	*20*	stand_VVI	*23*	statement_NN1	*43*	named_VVN	*33*
mark_NN1	*20*	working_VVG	*22*	plan_NN1	*43*	hear_VVB	*33*
information_NN1	*20*	coming_VVG	*22*	committee_NN1	*43*	bring_VVB	*33*
control_NN1	*20*	love_VVI	*21*	business_NN1	*43*	voted_VVN	*32*
campaign_NN1	*20*	hope_VVI	*21*	justice_NN1	*42*	fighting_VVG	*32*
act_NN1	*20*	enlist_VVB	*21*	employees_NN2	*42*	am_VBB	*32*
tax_NN1	*20*	elected_VVN	*21*	defense_NN1	*42*	learn_VVB	*31*
india_NN1	*19*	am_VBB	*21*	communities_NN2	*42*	doing_VDG	*31*
amendment_NN1	*19*	used_VVD	*20*	insurance_NN1	*40*	cuts_VVZ	*31*
action_NN1	*19*	look_VVI	*20*	system_NN1	*39*	put_VVB	*30*
week_NN1	*19*	learn_VVB	*20*	part_NN1	*39*	ensure_VVB	*30*
voters_NN2	*19*	check_VVI	*20*	children_NN2	*39*	create_VVB	*29*
leak_NN1	*19*	called_VVN	*20*	relief_NN1	*38*	pass_VVI	*28*
god_NN1	*18*	talking_VVG	*19*	office_NN1	*38*	allow_VVB	*28*
report_NN1	*18*	running_VVG	*19*	immigration_NN1	*38*	improve_VVB	*27*
lawmakers_NN2	*18*	run_VVI	*19*	benefits_NN2	*38*	got_VVN	*27*
kids_NN2	*18*	comes_VVZ	*19*	alexander_NP0	*38*	vote_VVI	*26*
health_NN1	*18*	save_VVI	*17*	infrastructure_NN1	*37*	stop_VVI	*26*
friends_NN2	*17*	might_VM0	*17*	honor_NN1	*37*	say_VVB	*26*
end_NN1	*17*	done_VDN	*17*	energy_NN1	*37*	lead_VVD	*26*
children_NN2	*17*	went_VVD	*16*	access_NN1	*37*	hope_VVI	*26*
ass_NN1	*17*	using_VVG	*16*	potus_NN1	*36*	getting_VVG	*26*
system_NN1	*17*	remember_VVB	*16*	lives_NN2	*36*	deserve_VVB	*26*
story_NN1	*17*	makes_VVZ	*16*	companies_NN2	*36*	celebrate_VVB	*26*
rules_NN2	*17*	goes_VVZ	*16*	secretary_NN1	*35*	says_VVZ	*25*

issues_NN2	*17*	gets_VVZ	*16*	ohio_NN1	*35*	means_VVZ	*25*
efforts_NN2	*16*	end_VVI	*16*	growth_NN1	*35*	introduced_VVD	*25*
army_NN1	*16*	become_VVB	*16*	efforts_NN2	*35*	cut_VVI	*25*
americans_NN2	*16*	stay_VVB	*15*	city_NN1	*35*	announced_VVD	*25*
things_NN2	*16*	pay_VVI	*15*	chairs_NN2	*35*	address_VVI	*25*
statement_NN1	*16*	meet_VVI	*15*	friend_NN1	*34*	grow_VVB	*24*
senators_NN2	*16*	knows_VVZ	*15*	enforcement_NN1	*34*	call_VVI	*24*
right_NN0	*16*	follow_VVB	*15*	cuts_NN2	*34*	stand_VVI	*23*
patriot_NN1	*15*	created_VVN	*15*	crisis_NN1	*34*	visit_VVI	*22*
officials_NN2	*15*	create_VVB	*15*	home_NN1	*33*	speak_VVB	*22*
minister_NN1	*15*	change_VVI	*15*	county_NN1	*33*	pay_VVI	*22*
leaders_NN2	*15*	believe_VVB	*15*	capitol_NN1	*33*	find_VVB	*22*
history_NN1	*15*	based_VVN	*15*	bonuses_NN2	*33*	breaking_VVG	*22*
committee_NN1	*15*	act_VVI	*15*	administration_NN1	*33*	strengthen_VVB	*21*
citizens_NN2	*15*	speak_VVB	*14*	rights_NN2	*32*	receive_VVB	*21*
cities_NN2	*15*	share_VVI	*14*	appropriations_NN2	*32*	makes_VVZ	*21*
budget_NN1	*15*	set_VVI	*14*	mississippi_NN2	*31*	leading_VVG	*21*
wife_NN1	*15*	seems_VVZ	*14*	russia_NN1	*30*	hold_VVI	*21*
war_NN1	*15*	leaked_VVN	*14*	policy_NN1	*30*	committed_VVD	*21*
thanks_NN2	*15*	hold_VVI	*14*	obamacare_NN1	*30*	benefit_VVI	*21*
team_NN1	*15*	happened_VVD	*14*	iowa_NN1	*30*	prevent_VVB	*20*
safety_NN1	*14*	ask_VVB	*14*	director_NN1	*30*	met_VVN	*20*
reason_NN1	*14*	affect_VVI	*14*	credit_NN1	*30*	lost_VVN	*20*
press_NN1	*14*	try_VVB	*13*	class_NN1	*30*	honor_VVI	*20*
position_NN1	*14*	start_VVI	*13*	border_NN1	*30*	believe_VVB	*20*
plan_NN1	*14*	passed_VVD	*13*	nations_NN2	*29*	using_VVG	*19*
nation_NN1	*14*	introduced_VVN	*13*	days_NN2	*29*	used_VVD	*19*
months_NN2	*14*	imagine_VVB	*13*	child_NN1	*29*	talking_VVG	*19*
leaker_NN1	*14*	holding_VVG	*13*	chairman_NN1	*29*	save_VVB	*19*
guns_NN2	*14*	happen_VVB	*13*	banks_NN2	*29*	remember_VVB	*19*
future_NN1	*14*	buy_VVB	*13*	amendment_NN1	*29*	held_VVD	*19*
fact_NN1	*14*	watching_VVG	*12*	action_NN1	*29*	having_VHG	*19*
crisis_NN1	*14*	wanted_VVN	*12*	wages_NN2	*28*	agree_VVB	*19*
countries_NN2	*14*	told_VVD	*12*	vote_NN1	*28*	welcome_VVI	*18*
corruption_NN1	*14*	taking_VVG	*12*	power_NN1	*28*	visiting_VVG	*18*
conference_NN1	*14*	pass_VVI	*12*	place_NN1	*28*	told_VVN	*18*
companies_NN2	*14*	moving_VVG	*12*	water_NN1	*27*	stay_VVI	*18*
changes_NN2	*14*	looking_VVG	*12*	washington_NP0	*27*	serving_VVG	*18*
candidates_NN2	*14*	heard_VVN	*12*	staff_NN0	*27*	raise_VVB	*18*
book_NN1	*14*	fight_VVI	*12*	nomination_NN1	*27*	giving_VVG	*18*
back_NN1	*14*	calls_VVZ	*12*	farmers_NN2	*27*	discussing_VVG	*18*
violence_NN1	*14*	calling_VVG	*12*	victims_NN2	*26*	discussed_VVN	*18*

usa_NN1	*14*	allowed_VVN	*12*	republicans_NN2	*26*	worked_VVD	*17*
tech_NN1	*14*	allow_VVB	*12*	obama_NN1	*26*	wish_VVI	*17*
shit_NN1	*14*	warns_VVZ	*11*	night_NN1	*26*	taking_VVG	*17*
services_NN2	*13*	wait_VVI	*11*	industry_NN1	*26*	missed_VVN	*17*
role_NN1	*13*	voted_VVD	*11*	group_NN1	*26*	looking_VVG	*17*
regulations_NN2	*13*	took_VVD	*11*	future_NN1	*26*	combat_VVI	*17*
policy_NN1	*13*	thought_VVN	*11*	confirmation_NN1	*26*	use_VVI	*16*
midterms_NN2	*13*	talk_VVI	*11*	college_NN1	*26*	talked_VVD	*16*
leakers_NN2	*13*	show_VVI	*11*	christmas_NN1	*26*	shows_VVZ	*16*
lawyer_NN1	*13*	required_VVN	*11*	abuse_NN1	*26*	seeing_VVG	*16*
issue_NN1	*13*	mean_VVB	*11*	trade_NN1	*25*	repeal_VVI	*16*
industry_NN1	*13*	live_VVI	*11*	success_NN1	*25*	listen_VVI	*16*
head_NN1	*13*	hear_VVB	*11*	men_NN2	*25*	leads_VVZ	*16*
europe_NN1	*13*	congratulating_VVG	*11*	korea_NN1	*25*	gave_VVD	*16*
effect_NN1	*13*	care_VVI	*11*	code_NN1	*25*	fix_VVI	*16*
drug_NN1	*13*	undermine_VVB	*10*	campaign_NN1	*25*	expand_VVB	*16*
deal_NN1	*13*	standing_VVG	*10*	alaska_NN1	*25*	confirmed_VVD	*16*
change_NN1	*13*	sent_VVN	*10*	air_NN1	*25*	comes_VVZ	*16*
button_NN1	*13*	remove_VVB	*10*	wall_NN1	*24*	amend_VVB	*16*
administration_NN1	*13*	passes_VVZ	*10*	senator_NN1	*24*	took_VVD	*15*
will_NN1	*13*	looks_VVZ	*10*	regime_NN1	*24*	speaking_VVG	*15*
votes_NN2	*13*	killing_VVG	*10*	recovery_NN1	*24*	signed_VVD	*15*
terrorism_NN1	*13*	killed_VVN	*10*	need_NN1	*24*	shut_VVN	*15*
show_NN1	*12*	hate_VVI	*10*	member_NN1	*24*	sent_VVD	*15*
sex_NN1	*12*	happens_VVZ	*10*	gun_NN1	*24*	providing_VVG	*15*
protection_NN1	*12*	following_VVG	*10*	fight_NN1	*24*	invest_VVB	*15*
process_NN1	*12*	fired_VVN	*10*	district_NN1	*24*	honored_VVD	*15*
politicians_NN2	*12*	discuss_VVB	*10*	decision_NN1	*24*	helped_VVD	*15*
police_NN1	*12*	continues_VVZ	*10*	choice_NN1	*24*	enjoy_VVI	*15*
laws_NN2	*12*	breaking_VVG	*10*	team_NN1	*23*	end_VVB	*15*
hell_NN1	*12*	works_VVZ	*9*	story_NN1	*23*	continuing_VVg	*15*
hands_NN2	*12*	won_VVD	*9*	review_NN1	*23*	works_VVZ	*14*
daniel_NN1	*12*	taken_VVN	*9*	north_NN1	*23*	won_VVN	*14*
court_NN1	*12*	state_VVI	*9*	murray_NN1	*23*	remain_VVB	*14*
case_NN1	*12*	signed_VVD	*9*	luck_NN1	*23*	received_VVD	*14*
briefing_NN1	*12*	showing_VVG	*9*	help_NN1	*23*	move_VVI	*14*
access_NN1	*12*	reading_VVG	*9*	court_NN1	*23*	love_VVI	*14*
washington_NN1	*12*	paying_VVG	*9*	west_NN1	*22*	led_VVN	*14*
threat_NN1	*12*	miss_VVB	*9*	texas_NN2	*22*	include_VVB	*14*
term_NN1	*12*	means_VVZ	*9*	step_NN1	*22*	helps_VVZ	*14*
technology_NN1	*12*	lost_VVD	*9*	resources_NN2	*22*	facing_VVG	*14*
snow_NN1	*11*	held_VVN	*9*	research_NN1	*22*	build_VVB	*14*
sir_NN1	*11*	guess_VVB	*9*	programs_NN2	*22*	advance_VVI	*14*
service_NN1	*11*	forget_VVB	*9*	meeting_NN1	*22*	served_VVD	*13*
risks_NN2	*11*	feel_VVB	*9*	john_NN1	*22*	protecting_VVG	*13*
piece_NN1	*11*	ensure_VVB	*9*	issue_NN1	*22*	passing_VVG	*13*
patriots_NN2	*11*	congratulates_VVZ	*9*	god_NN1	*22*	lower_VVI	*13*

men_NN2	*11*	claims_VVZ	*9*	food_NN1	*22*	learned_VVN	*13*
majority_NN1	*11*	caught_VVN	*9*	deal_NN1	*22*	keeping_VVG	*13*
line_NN1	*11*	bring_VVB	*9*	bills_NN2	*22*	joining_VVG	*13*
kind_NN1	*11*	added_VVD	*9*	visit_NN1	*21*	happen_VVB	*13*
joe_NN1	*11*	worry_VVI	*8*	speech_NN1	*21*	following_VVG	*13*
independence_NN1	*11*	supporting_VVG	*8*	services_NN2	*21*	encourage_VVB	*13*
force_NN1	*11*	suggests_VVZ	*8*	man_NN1	*21*	change_VVI	*13*
constitution_NN1	*11*	seen_VVN	*8*	event_NN1	*21*	care_VVI	*13*
china_NN1	*11*	seeks_VVZ	*8*	election_NN1	*21*	called_VVN	*13*
child_NN1	*11*	realize_VVB	*8*	effort_NN1	*21*	brings_VVZ	*13*
chief_NN1	*11*	quit_VVB	*8*	disaster_NN1	*21*	asked_VVN	*13*
car_NN1	*11*	putting_VVG	*8*	development_NN1	*21*	appreciated_VVN	*13*
businesses_NN2	*11*	meeting_VVG	*8*	colleagues_NN2	*21*	win_VVB	*12*
bills_NN2	*11*	lose_VVB	*8*	tour_NN1	*20*	succeed_VVB	*12*
woman_NN1	*11*	left_VVN	*8*	technology_NN1	*20*	stopped_VVD	*12*
warning_NN1	*11*	leave_VVI	*8*	student_NN1	*20*	spoke_VVD	*12*
video_NN1	*11*	leaking_VVG	*8*	role_NN1	*20*	show_VVI	*12*
truth_NN1	*10*	includes_VVZ	*8*	report_NN1	*20*	seen_VVN	*12*
sports_NN2	*10*	having_VHG	*8*	prayers_NN2	*20*	represents_VVZ	*12*
society_NN1	*10*	gain_VVI	*8*	month_NN1	*20*	plan_VVI	*12*
secretary_NN1	*10*	congratulated_VVN	*8*	iran_NP0	*20*	needed_VVD	*12*
schools_NN2	*10*	comply_VVB	*8*	idaho_NN1	*20*	moving_VVG	*12*
questions_NN2	*10*	click_VVI	*8*	example_NN1	*20*	mean_VVI	*12*
program_NN0	*10*	becomes_VVZ	*8*	drug_NN1	*20*	goes_VVZ	*12*
name_NN1	*10*	became_VVD	*8*	column_NN1	*20*	fund_VVI	*12*
lot_NN1	*10*	announce_VVB	*8*	center_NN1	*20*	failed_VVN	*12*
lies_NN2	*10*	agree_VVB	*8*	wyoming_NP0	*19*	deliver_VVB	*12*
james_NP0	*10*	achieve_VVB	*8*	schools_NN2	*19*	consider_VVB	*12*
idea_NN1	*10*	wish_VVI	*7*	response_NN1	*19*	came_VVD	*12*
hill_NN1	*10*	warned_VVD	*7*	resolution_NN1	*19*	calling_VVG	*12*
guy_NN1	*10*	updates_VVZ	*7*	questions_NN2	*19*	buy_VVB	*12*
family_NN1	*10*	tells_VVZ	*7*	process_NN1	*19*	boost_VVI	*12*
director_NN1	*10*	target_VVI	*7*	policies_NN2	*19*	begins_VVZ	*12*
congratulations_NN2	*10*	takes_VVZ	*7*	pay_NN1	*19*	wait_VVB	*11*
clinton_NN1	*10*	starting_VVG	*7*	others_NN2	*19*	urging_VVG	*11*
class_NN1	*10*	started_VVD	*7*	order_NN1	*19*	tell_VVB	*11*
boss_NN1	*10*	send_VVB	*7*	opportunities_NN2	*19*	taken_VVN	*11*
board_NN1	*10*	seem_VVB	*7*	israel_NP0	*19*	supporting_VVG	*11*
attention_NN1	*10*	saying_VVG	*7*	investment_NN1	*19*	spent_VVN	*11*
attacks_NN2	*10*	received_VVN	*7*	farm_NN1	*19*	running_VVD	*11*
word_NN1	*10*	protecting_VVG	*7*	fact_NN1	*19*	review_VVI	*11*
trade_NN1	*10*	promote_VVB	*7*	attack_NN1	*19*	praying_VVG	*11*
session_NN1	*10*	plans_VVZ	*7*	arizona_NN1	*19*	paying_VVG	*11*
ryan_NP0	*10*	move_VVB	*7*	agenda_NN1	*19*	open_VVI	*11*
risk_NN1	*9*	manipulating_VVG	*7*	trafficking_NN1	*18*	increase_VVI	*11*
research_NN1	*9*	improve_VVB	*7*	subcommittee_	*18*	includes_VVZ	*11*

				NN1			
reality_NN1	*9*	helping_VVG	*7*	shutdown_NN1	*18*	dedicated_VVD	*11*
project_NN1	*9*	grow_VVI	*7*	safety_NN1	*18*	treat_VVI	*10*
point_NN1	*9*	fighting_VVG	*7*	right_NN1	*18*	think_VVB	*10*
phone_NN1	*9*	face_VVI	*7*	proposal_NN1	*18*	testify_VVB	*10*
message_NN1	*9*	explain_VVB	*7*	presidents_NN2	*18*	takes_VVZ	*10*
members_NN2	*9*	covering_VVG	*7*	premiums_NN2	*18*	state_VVI	*10*
look_NN1	*9*	continue_VVB	*7*	paychecks_NN2	*18*	stands_VVZ	*10*
link_NN1	*9*	becoming_VVG	*7*	nominee_NN1	*18*	standing_VVG	*10*
level_NN1	*9*	approves_VVZ	*7*	months_NN2	*18*	share_VVI	*10*
john_NN1	*9*	written_VVN	*6*	leader_NN1	*18*	send_VVB	*10*
integrity_NN1	*9*	write_VVB	*6*	kids_NN2	*18*	run_VVI	*10*
home_NN1	*9*	wonder_VVI	*6*	info_NN1	*18*	restore_VVB	*10*
governor_NN1	*9*	winning_VVG	*6*	floor_NN1	*18*	puts_VVZ	*10*
governments_NN2	*9*	walk_VVI	*6*	clinton_NN1	*18*	pray_VVB	*10*
facts_NN2	*9*	visit_VVI	*6*	china_NN1	*18*	play_VVI	*10*
executive_NN1	*9*	updated_VVD	*6*	bank_NN1	*18*	impacted_VVD	*10*
episode_NN1	*9*	turned_VVN	*6*	back_NN1	*18*	hosting_VVG	*10*
door_NN1	*9*	tried_VVN	*6*	area_NN1	*18*	hire_VVI	*10*
dollars_NN2	*9*	suggest_VVB	*6*	weekend_NN1	*17*	growing_VVG	*10*
dems_NN2	*9*	spent_VVN	*6*	training_NN1	*17*	gives_VVZ	*10*
democrats_NN2	*9*	spend_VVB	*6*	times_NN2	*17*	found_VVN	*10*
death_NN1	*9*	speaking_VVG	*6*	tennesseans_NN2	*17*	fought_VVD	*10*
crime_NN1	*9*	signing_VVG	*6*	priorities_NN2	*17*	created_VVD	*10*
counsel_NN1	*9*	shut_VVN	*6*	party_NN1	*17*	celebrating_VVG	*10*
council_NN1	*9*	shall_VM0	*6*	number_NN1	*17*	brought_VVD	*10*
club_NN1	*9*	sell_VVI	*6*	laws_NN2	*17*	attending_VVG	*10*
city_NN1	*9*	risk_VVI	*6*	healthcare_NN1	*17*	voting_VVG	*9*
chaos_NN1	*9*	reported_VVD	*6*	graham_NN1	*17*	testifies_VVZ	*9*
california_NN1	*9*	released_VVN	*6*	game_NN1	*17*	start_VVI	*9*
agenda_NN1	*9*	receive_VVB	*6*	deadline_NN1	*17*	solve_VVB	*9*
advice_NN1	*9*	provide_VVB	*6*	attorney_NN1	*17*	set_VVI	*9*
words_NN2	*9*	prevent_VVB	*6*	americas_NN2	*17*	respond_VVB	*9*
voice_NN1	*9*	playing_VVG	*6*	afternoon_NN1	*17*	released_VVN	*9*
treason_NN1	*9*	moved_VVN	*6*	youth_NN1	*16*	rebuild_VVB	*9*
thursday_NN1	*9*	met_VVN	*6*	ways_NN2	*16*	promised_VVD	*9*
thoughts_NN2	*8*	meant_VVD	*6*	victory_NN1	*16*	paid_VVN	*9*
task_NN1	*8*	lying_VVG	*6*	utah_NN1	*16*	lose_VVB	*9*
speech_NN1	*8*	listening_VVG	*6*	things_NN2	*16*	knows_VVZ	*9*
side_NN1	*8*	listen_VVI	*6*	tennessee_NN1	*16*	kept_VVN	*9*
response_NN1	*8*	lets_VVZ	*6*	solution_NN1	*16*	included_VVD	*9*
republicans_NN2	*8*	leaks_VVZ	*6*	result_NN1	*16*	honoring_VVG	*9*
progress_NN1	*8*	lead_VVI	*6*	progress_NN1	*16*	holding_VVG	*9*
processes_NN2	*8*	kills_VVZ	*6*	parents_NN2	*16*	gets_VVZ	*9*
others_NN2	*8*	informed_VVN	*6*	ones_NN2	*16*	follow_VVB	*9*
opportunity_NN1	*8*	increase_VVI	*6*	letter_NN1	*16*	fed_VVN	*9*
oil_NN1	*8*	include_VVB	*6*	investigation_NN1	*16*	defend_VVB	*9*
north_NN1	*8*	hide_VVB	*6*	hope_NN1	*16*	continues_VVZ	*9*
management_NN1	*8*	giving_VVG	*6*	friends_NN2	*16*	checks_VVZ	*9*

lives_NN2	*8*	fire_VVI	*6*	field_NN1	*16*	bringing_VVG	*9*
kelly_NP0	*8*	exposed_VVD	*6*	discussion_NN1	*16*	born_VVN	*9*
innovation_NN1	*8*	expected_VVN	*6*	change_NN1	*16*	bless_VVB	*9*
infrastructure_NN1	*8*	expect_VVB	*6*	anniversary_NN1	*16*	become_VVB	*9*
info_NN1	*8*	destroy_VVB	*6*	admin_NN1	*16*	based_VVN	*9*
impact_NN1	*8*	defend_VVB	*6*	weeks_NN2	*15*	ask_VVB	*9*
husband_NN1	*8*	consider_VVB	*6*	war_NN1	*15*	apply_VVB	*9*
hillary_NP0	*8*	committed_VVN	*6*	union_NN1	*15*	announce_VVB	*9*
general_NN1	*8*	brought_VVD	*6*	thoughts_NN2	*15*	afford_VVB	*9*
freedom_NN1	*8*	bet_VVI	*6*	thing_NN1	*15*	add_VVB	*9*
fire_NN1	*8*	answer_VVI	*6*	south_NN1	*15*	achieve_VVB	*9*
example_NN1	*8*	add_VVB	*6*	sense_NN1	*15*	winning_VVG	*8*
enforcement_NN1	*8*	accept_VVB	*6*	senators_NN2	*15*	tax_VVI	*8*
email_NN1	*8*	worried_VVN	*5*	season_NN1	*15*	stopping_VVG	*8*
conservatives_NN2	*8*	wins_VVZ	*5*	rico_NN1	*15*	started_VVD	*8*
company_NN1	*8*	waiting_VVG	*5*	request_NN1	*15*	securing_VVG	*8*
care_NN1	*8*	understand_VVB	*5*	promise_NN1	*15*	saved_VVD	*8*
assembly_NN1	*8*	supports_VVZ	*5*	project_NN1	*15*	reform_VVI	*8*
approach_NN1	*8*	spending_VVG	*5*	problem_NN1	*15*	reduce_VVB	*8*
wednesday_NP0	*8*	seal_VVI	*5*	officials_NN2	*15*	rebuilding_VVG	*8*
webinar_NN0	*8*	rid_VVD	*5*	media_NN0	*15*	plans_VVZ	*8*
view_NN1	*8*	review_VVI	*5*	market_NN1	*15*	opening_VVG	*8*
victory_NN1	*8*	reveals_VVZ	*5*	hill_NN1	*15*	killed_VVD	*8*
twitter_NN1	*7*	resigns_VVZ	*5*	force_NN1	*15*	keeps_VVZ	*8*
tweets_NN2	*7*	require_VVB	*5*	faith_NN1	*15*	joined_VVD	*8*
top_NN0	*7*	remain_VVB	*5*	end_NN1	*15*	host_VVI	*8*
thought_NN1	*7*	regulate_VVB	*5*	drugs_NN2	*15*	hit_VVD	*8*
terrorist_NN1	*7*	register_VVI	*5*	debt_NN1	*15*	empower_VVI	*8*
tariffs_NN2	*7*	react_VVB	*5*	control_NN1	*15*	drive_VVI	*8*
systems_NN2	*7*	raises_VVZ	*5*	chance_NN1	*15*	doubling_VVG	*8*
stock_NN1	*7*	puts_VVZ	*5*	board_NN1	*15*	demand_VVD	*8*
sense_NN1	*7*	plan_VVI	*5*	ambassador_NN1	*15*	delivered_VVD	*8*
self_NN1	*7*	participate_VVB	*5*	weapons_NN2	*14*	covers_VVZ	*8*
seats_NN2	*7*	paid_VVN	*5*	vegas_NN2	*14*	calls_VVZ	*8*
robert_NP0	*7*	own_VVI	*5*	teachers_NN2	*14*	briefed_VVD	*8*
rise_NN1	*7*	opening_VVG	*5*	schumer_NN1	*14*	breaks_VVZ	*8*
respect_NN1	*7*	open_VVI	*5*	patients_NN2	*14*	ban_VVB	*8*
quality_NN1	*7*	offer_VVI	*5*	missile_NN1	*14*	awarded_VVN	*8*
privacy_NN1	*7*	matter_VVI	*5*	mccain_NN1	*14*	welcoming_VVG	*7*
post_NN1	*7*	leading_VVG	*5*	markup_NN1	*14*	try_VVI	*7*
peace_NN1	*7*	kill_VVB	*5*	lot_NN1	*14*	states_VVZ	*7*
payroll_NN1	*7*	keeping_VVG	*5*	internet_NN1	*14*	starts_VVZ	*7*
patients_NN2	*7*	introduce_VVB	*5*	intelligence_NN1	*14*	starting_VVG	*7*
paper_NN1	*7*	influenced_VVN	*5*	importance_NN1	*14*	spending_VVG	*7*

pablo_NN1	*7*	happening_VVG	*5*	impact_NN1	*14*	spend_VVB	*7*
opportunities_NN2	*7*	handle_VVI	*5*	folks_NN2	*14*	sanctions_VVZ	*7*
opinion_NN1	*7*	guide_VVI	*5*	emergency_NN1	*14*	represent_VVB	*7*
omnibus_NN1	*7*	form_VVI	*5*	dossier_NN1	*14*	receiving_VVG	*7*
need_NN1	*7*	fails_VVZ	*5*	courage_NN1	*14*	raising_VVG	*7*
month_NN1	*7*	failing_VVG	*5*	consumers_NN2	*14*	passes_VVZ	*7*
model_NN1	*7*	explained_VVN	*5*	conference_NN1	*14*	looks_VVZ	*7*
mind_NN1	*7*	established_VVN	*5*	company_NN1	*14*	lift_VVI	*7*
liberals_NN2	*7*	encourage_VVB	*5*	chair_NN1	*14*	inspired_VVD	*7*
legislators_NN2	*7*	died_VVN	*5*	capital_NN1	*14*	increased_VVD	*7*
leadership_NN1	*7*	deserves_VVZ	*5*	cancer_NN1	*14*	hurt_VVI	*7*
knowledge_NN1	*7*	demand_VVI	*5*	background_NN1	*14*	happened_VVD	*7*
investment_NN1	*7*	defends_VVZ	*5*	assistance_NN1	*14*	forget_VVB	*7*
investigation_NN1	*7*	contact_VVI	*5*	alex_NN1	*14*	focused_VVD	*7*
interference_NN1	*7*	choose_VVB	*5*	agreement_NN1	*14*	focus_VVI	*7*
intelligence_NN1	*7*	changed_VVN	*5*	troops_NN2	*13*	finding_VVG	*7*
homeland_NN1	*7*	celebrate_VVB	*5*	town_NN1	*13*	delivering_VVG	*7*
george_NN1	*7*	caused_VVN	*5*	spirit_NN1	*13*	cutting_VVG	*7*
fear_NN1	*7*	cause_VVI	*5*	rules_NN2	*13*	creating_VVG	*7*
father_NN1	*7*	build_VVI	*5*	rule_NN0	*13*	contact_VVI	*7*
experience_NN1	*7*	brings_VVZ	*5*	regulations_NN2	*13*	choose_VVB	*7*
events_NN2	*7*	bless_VVB	*5*	reforms_NN2	*13*	asks_VVZ	*7*
duty_NN1	*7*	betting_VVG	*5*	reason_NN1	*13*	allows_VVZ	*7*
democrat_NN1	*7*	asked_VVD	*5*	quality_NN1	*13*	allowing_VVG	*7*
definition_NN1	*7*	announced_VVN	*5*	problems_NN2	*13*	affected_VVD	*7*
deals_NN2	*7*	allowing_VVG	*5*	pres_NN2	*13*	access_VVI	*7*
computer_NN1	*7*	address_VVI	*5*	nafta_NN1	*13*	wants_VVZ	*6*
citizen_NN1	*7*	accepting_VVG	*5*	medal_NN1	*13*	waiting_VVG	*6*
capitol_NN1	*7*	welcome_VVI	*4*	interests_NN2	*13*	visited_VVD	*6*
canada_NN1	*7*	wearing_VVG	*4*	individuals_NN2	*13*	urge_VVI	*6*
boy_NN1	*7*	waste_VVI	*4*	increase_NN1	*13*	thought_VVI	*6*
bomber_NN1	*7*	uploaded_VVD	*4*	heroes_NN2	*13*	thinks_VVZ	*6*
blog_NN1	*7*	uphold_VVB	*4*	groups_NN2	*13*	strengthening_VVG	*6*
biden_NN1	*7*	update_VVI	*4*	games_NN2	*13*	stabilize_VVB	*6*
attack_NN1	*7*	understanding_VVG	*4*	epidemic_NN1	*13*	sit_VVB	*6*
ability_NN1	*7*	trust_VVI	*4*	cost_NN1	*13*	sharing_VVG	*6*
water_NN1	*7*	trending_VVG	*4*	commitment_NN1	*13*	scam_VVI	*6*
voting_NN1	*7*	thinking_VVG	*4*	chip_NN1	*13*	saw_VVD	*6*
vladimir_NN1	*7*	talks_VVZ	*4*	challenges_NN2	*13*	raised_VVN	*6*
visit_NN1	*7*	talked_VVN	*4*	areas_NN2	*13*	prepare_VVB	*6*
virginia_NN1	*6*	suck_VVB	*4*	aid_NN0	*13*	outlined_VVD	*6*
village_NN1	*6*	strengthen_VV	*4*	accountability_	*13*	miss_VVI	*6*

		B		NN1			
user_NN1	*6*	stated_VVN	*4*	words_NN2	*12*	left_VVN	*6*
trumps_NP0	*6*	sitting_VVG	*4*	wisconsin_NN1	*12*	involved_VVD	*6*
training_NN1	*6*	signs_VVZ	*4*	trip_NN1	*12*	impacting_VVG	*6*
ties_NN2	*6*	shown_VVN	*4*	sex_NN1	*12*	honors_VVI	*6*
threats_NN2	*6*	sharing_VVG	*4*	savings_NN2	*12*	harm_VVB	*6*
taxes_NN2	*6*	shared_VVD	*4*	record_NN1	*12*	happening_VVG	*6*
swamp_NN1	*6*	serving_VVG	*4*	plans_NN2	*12*	gone_VVN	*6*
surprise_NN1	*6*	selling_VVG	*4*	payments_NN2	*12*	eliminate_VVB	*6*
style_NN1	*6*	sees_VVZ	*4*	missourians_NN2	*12*	develop_VVB	*6*
street_NN1	*6*	seek_VVB	*4*	message_NN1	*12*	decide_VVB	*6*
stories_NN2	*6*	requires_VVZ	*4*	medicare_NN1	*12*	cosponsored_VVD	*6*
stigma_NN1	*6*	reporting_VVG	*4*	list_NN1	*12*	competes_VVZ	*6*
spaniel_NN1	*6*	realise_VVB	*4*	line_NN1	*12*	compete_VVB	*6*
south_NN1	*6*	ran_VVD	*4*	king_NN1	*12*	click_VVI	*6*
sitting_NN1	*6*	pushing_VVG	*4*	kind_NN1	*12*	cheering_VVG	*6*
sham_NN1	*6*	push_VVI	*4*	judge_NN1	*12*	chairing_VVG	*6*
senator_NN1	*6*	provides_VVZ	*4*	investments_NN2	*12*	celebrated_VVD	*6*
section_NN1	*6*	prove_VVB	*4*	income_NN1	*12*	built_VVN	*6*
scandal_NN1	*6*	prefer_VVB	*4*	hurricane_NN1	*12*	begin_VVB	*6*
sanctuary_NN1	*6*	praying_VVG	*4*	gold_NN1	*12*	becoming_VVG	*6*
results_NN2	*6*	pray_VVI	*4*	florida_NP0	*12*	became_VVD	*6*
restrictions_NN2	*6*	overturn_VVB	*4*	executive_NN1	*12*	assist_VVB	*6*
powers_NN2	*6*	missed_VVN	*4*	east_NN1	*12*	asking_VVG	*6*
polls_NN2	*6*	losing_VVG	*4*	dream_NN1	*12*	announces_VVZ	*6*
period_NN1	*6*	leak_VVI	*4*	difference_NN1	*12*	advocating_VVG	*6*
pedro_NN1	*6*	launches_VVZ	*4*	decades_NN2	*12*	advocate_VVI	*6*
paul_NN1	*6*	launched_VVD	*4*	crimes_NN2	*12*	acting_VVG	*6*
parents_NN2	*6*	keeps_VVZ	*4*	conversation_NN1	*12*	accept_VVB	*6*
page_NN1	*6*	interfered_VVD	*4*	call_NN1	*12*	wrote_VVD	*5*
order_NN1	*6*	informing_VVG	*4*	award_NN1	*12*	unite_VVB	*5*
opposition_NN1	*6*	impose_VVB	*4*	arkansas_NN2	*12*	understand_VVB	*5*
opioids_NN2	*6*	holds_VVZ	*4*	announcement_NN1	*12*	turn_VVI	*5*
nicotine_NN1	*6*	hit_VVD	*4*	agriculture_NN1	*12*	trust_VVI	*5*
mission_NN1	*6*	hired_VVD	*4*	address_NN1	*12*	treated_VVN	*5*
mexico_NN1	*6*	hire_VVI	*4*	academy_NN1	*12*	thrive_VVB	*5*
member_NN1	*6*	hiding_VVG	*4*	wishes_NN2	*11*	simplify_VVB	*5*
measures_NN2	*6*	helped_VVD	*4*	violence_NN1	*11*	sign_VVI	*5*
matters_NN2	*6*	forgot_VVD	*4*	treatment_NN1	*11*	serves_VVZ	*5*
market_NN1	*6*	force_VVI	*4*	threats_NN2	*11*	selling_VVG	*5*
love_NN1	*6*	focused_VVD	*4*	threat_NN1	*11*	saying_VVG	*5*
legalization_NN1	*6*	ensuring_VVG	*4*	term_NN1	*11*	rise_VVI	*5*
initiative_NN1	*6*	enact_VVB	*4*	summer_NN1	*11*	reverse_VVI	*5*
influence_NN1	*6*	drive_VVI	*4*	street_NN1	*11*	require_VVB	*5*
hollywood_NN1	*6*	discussing_VV	*4*	stories_NN2	*11*	request_VVI	*5*

		G					
healthcare_NN1	*6*	discover_VVB	*4*	society_NN1	*11*	replace_VVB	*5*
hand_NN1	*6*	deserve_VVB	*4*	retirement_NN1	*11*	renew_VVB	*5*
groups_NN2	*6*	declare_VVB	*4*	results_NN2	*11*	remains_VVZ	*5*
fun_NN1	*6*	dare_VM0	*4*	republican_NN1	*11*	reflect_VVB	*5*
fraud_NN1	*6*	cut_VVI	*4*	rate_NN1	*11*	reducing_VVG	*5*
face_NN1	*6*	crying_VVG	*4*	protection_NN1	*11*	recover_VVB	*5*
expansion_NN1	*6*	crosses_VVZ	*4*	price_NN1	*11*	recognize_VVB	*5*
evidence_NN1	*6*	creates_VVZ	*4*	peace_NN1	*11*	reach_VVI	*5*
energy_NN1	*6*	covered_VVD	*4*	nominees_NN2	*11*	ranked_VVN	*5*
employees_NN2	*6*	compromised_VVD	*4*	moment_NN1	*11*	raises_VVZ	*5*
education_NN1	*6*	combat_VVI	*4*	mandate_NN1	*11*	putting_VVG	*5*
economy_NN1	*6*	close_VVI	*4*	maine_NN1	*11*	purchase_VVI	*5*
disease_NN1	*6*	charged_VVN	*4*	levels_NN2	*11*	provides_VVZ	*5*
discussion_NN1	*6*	causing_VVG	*4*	legacy_NN1	*11*	produce_VVI	*5*
development_NN1	*6*	cares_VVZ	*4*	jerusalem_NN1	*11*	preserve_VVB	*5*
details_NN2	*6*	bought_VVD	*4*	intel_NN1	*11*	playing_VVG	*5*
decisions_NN2	*6*	believed_VVN	*4*	information_NN1	*11*	offer_VVI	*5*
decision_NN1	*6*	begin_VVB	*4*	facility_NN1	*11*	moves_VVZ	*5*
decades_NN2	*6*	beat_VVI	*4*	elections_NN2	*11*	moved_VVD	*5*
decade_NN1	*6*	bans_VVZ	*4*	deduction_NN1	*11*	losing_VVG	*5*
debate_NN1	*6*	awaits_VVZ	*4*	david_NN1	*11*	learning_VVG	*5*
cryptocurrency_NN1	*6*	avoid_VVB	*4*	data_NN0	*11*	knew_VVD	*5*
crown_NN1	*6*	audit_VVI	*4*	countries_NN2	*11*	introduce_VVB	*5*
convention_NN1	*6*	attacks_VVZ	*4*	costs_NN2	*11*	increasing_VVG	*5*
contracts_NN2	*6*	attacked_VVN	*4*	consumer_NN1	*11*	increases_VVZ	*5*
box_NN1	*6*	attach_VVB	*4*	competition_NN1	*11*	incentivize_VVB	*5*
border_NN1	*6*	asking_VVG	*4*	breakfast_NN1	*11*	implemented_VVD	*5*
benefit_NN1	*6*	approved_VVD	*4*	birthday_NN1	*11*	head_VVI	*5*
authorities_NN2	*6*	apply_VVB	*4*	ban_NN1	*11*	given_VVN	*5*
association_NN1	*6*	announces_VVZ	*4*	army_NN1	*11*	forgotten_VVN	*5*
area_NN1	*6*	allows_VVZ	*4*	actions_NN2	*11*	fired_VVD	*5*
analytics_NN1	*6*	affecting_VVG	*4*	wind_NN1	*10*	finds_VVZ	*5*
agreement_NN1	*6*	advance_VVI	*4*	votes_NN2	*10*	fill_VVB	*5*
addiction_NN1	*6*	addressing_VVG	*4*	valley_NN1	*10*	feel_VVB	*5*
abuse_NN1	*6*	acts_VVZ	*4*	tom_NN1	*10*	feed_VVI	*5*
whitehouse_NN1	*6*	wouldnt_VVB	*3*	theyre_NN1	*10*	face_VVI	*5*
west_NN1	*6*	wondering_VVG	*3*	terror_NN1	*10*	experience_VVI	*5*
walter_NN1	*6*	warn_VVB	*3*	taxpayers_NN2	*10*	expected_VVN	*5*
uncle_NN1	*5*	wagering_VVG	*3*	strategy_NN1	*10*	expect_VVB	*5*
tuesday_NN1	*5*	vows_VVZ	*3*	solutions_NN2	*10*	ensuring_VVG	*5*
topic_NN1	*5*	voting_VVG	*3*	skills_NN2	*10*	enforce_VVB	*5*
text_NN1	*5*	view_VVI	*3*	side_NN1	*10*	ending_VVG	*5*
test_NN1	*5*	verify_VVB	*3*	reminder_NN1	*10*	encouraged_VVD	*5*

talks_NN2	5	uses_VVZ	3	ranchers_NN2	10	earned_VVN	5
talk_NN1	5	unveiled_VVD	3	public_NN0	10	earn_VVB	5
strategy_NN1	5	tweets_VVZ	3	proof_NN1	10	doubles_VVZ	5
status_NN1	5	tweet_VVB	3	products_NN2	10	died_VVD	5
standards_NN2	5	turning_VVG	3	production_NN1	10	deal_VVI	5
stage_NN1	5	turn_VVI	3	police_NN2	10	curb_VVI	5
staffer_NN1	5	trailblazing_VVG	3	point_NN1	10	confirms_VVZ	5
sources_NN2	5	ties_VVZ	3	pleasure_NN1	10	competing_VVG	5
silence_NN1	5	threatens_VVZ	3	percent_NN0	10	combating_VVG	5
signs_NN2	5	threatened_VVD	3	paycheck_NN1	10	claims_VVZ	5
shutdown_NN1	5	testing_VVG	3	page_NN1	10	changed_VVD	5
seminar_NN1	5	testify_VVB	3	missouri_NN2	10	caused_VVN	5
seat_NN1	5	telling_VVG	3	mayor_NN1	10	buying_VVG	5
scum_NN1	5	suggesting_VVG	3	kansas_NN2	10	avoid_VVB	5
russians_NN2	5	succeed_VVB	3	ideas_NN2	10	attended_VVN	5
roads_NN2	5	subscribe_VVB	3	heart_NN1	10	attacked_VVD	5
road_NN1	5	strikes_VVZ	3	flexibility_NN1	10	attack_VVI	5
rest_NN1	5	strengthening_VVG	3	fighter_NN1	10	answered_VVN	5
reasons_NN2	5	stopped_VVD	3	experience_NN1	10	answer_VVB	5
reagan_NN1	5	states_VVZ	3	details_NN2	10	alert_VVI	5
reaction_NN1	5	spoke_VVD	3	department_NN1	10	worry_VVI	4
race_NN1	5	speaks_VVZ	3	democrat_NN1	10	wishes_VVZ	4
question_NN1	5	sounds_VVZ	3	democracy_NN1	10	watching_VVG	4
propaganda_NN1	5	shows_VVZ	3	cruz_NN1	10	wanted_VVD	4
product_NN1	5	shot_VVI	3	constituents_NN2	10	update_VVI	4
priorities_NN2	5	shooting_VVG	3	confidence_NN1	10	understands_VVZ	4
prince_NN1	5	shares_VVZ	3	colleges_NN2	10	undermine_VVB	4
price_NN1	5	sets_VVZ	3	cities_NN2	10	turns_VVZ	4
pressure_NN1	5	serve_VVB	3	christopher_NP0	10	turned_VVD	4
potential_NN1	5	sends_VVZ	3	ceremony_NN1	10	trying_VVG	4
points_NN2	5	seeking_VVG	3	century_NN1	10	tries_VVZ	4
play_NN1	5	seeing_VVG	3	centers_NN2	10	teach_VVB	4
picture_NN1	5	screw_VVI	3	book_NN1	10	struggling_VVG	4
photos_NN2	5	saw_VVI	3	billy_NN1	10	stepping_VVG	4
peru_NN1	5	runs_VVZ	3	agencies_NN2	10	sold_VVD	4
past_NN1	5	roll_VVI	3	addiction_NN1	10	shutting_VVG	4
parsons_NN2	5	rigged_VVD	3	acts_NN2	10	showing_VVG	4
parliament_NN1	5	revising_VVG	3	workforce_NN1	9	shooting_VVG	4
parenthood_NN1	5	return_VVI	3	values_NN2	9	shared_VVD	4
papers_NN2	5	respect_VVI	3	use_NN1	9	sending_VVG	4
package_NN1	5	represent_VVB	3	uniform_NN1	9	seem_VVB	4

owners_NN2	5	report_VVI	3	tools_NN2	9	seeking_VVG	4
option_NN1	5	related_VVD	3	steps_NN2	9	seek_VVB	4
nixon_NN1	5	reject_VVI	3	space_NN1	9	secure_VVI	4
network_NN1	5	regards_VVZ	3	show_NN1	9	saving_VVG	4
nations_NN2	5	refuses_VVZ	3	sessions_NN2	9	sacrificed_VVD	4
narrative_NN1	5	refused_VVD	3	seniors_NN2	9	sacrifice_VVI	4
movie_NN1	5	refuse_VVI	3	sector_NN1	9	rooting_VVG	4
mothers_NN2	5	reflect_VVB	3	room_NN1	9	rising_VVG	4
morning_NN1	5	reduce_VVB	3	retailers_NN2	9	reveals_VVZ	4
monday_NN1	5	receiving_VVG	3	respect_NN1	9	resulted_VVN	4
mom_NN1	5	reaching_VVG	3	potential_NN1	9	responding_VVG	4
minutes_NN2	5	raised_VVD	3	photos_NN2	9	requiring_VVG	4
minds_NN2	5	raise_VVI	3	partners_NN2	9	reports_VVZ	4
matter_NN0	5	quitting_VVG	3	park_NN1	9	report_VVB	4
maryland_NN1	5	pull_VVI	3	owners_NN2	9	repealing_VVG	4
loyalty_NN1	5	project_VVI	3	officers_NN2	9	rely_VVB	4
lawrence_NN1	5	prohibit_VVB	3	nebraska_NN1	9	reject_VVB	4
land_NN1	5	process_VVI	3	mueller_NN1	9	register_VVI	4
kremlin_NP0	5	posting_VVG	3	mexico_NN1	9	refuses_VVZ	4
joke_NN1	5	pose_VVI	3	markets_NN2	9	reads_VVZ	4
jobs_NN2	5	played_VVD	3	mark_NP0	9	reading_VVG	4
jeff_NP0	5	play_VVI	3	love_NN1	9	reaching_VVG	4
involvement_NN1	5	pick_VVI	3	look_NN1	9	reached_VVN	4
interview_NN1	5	pays_VVZ	3	liberty_NN1	9	ran_VVD	4
inaction_NN1	5	owned_VVD	3	interview_NN1	9	pushing_VVG	4
importance_NN1	5	opens_VVZ	3	homeland_NN1	9	pursuing_VVG	4
impeachment_NN1	5	obtain_VVB	3	guns_NN2	9	pursue_VVB	4
ideas_NN2	5	notes_VVZ	3	generation_NN1	9	provided_VVD	4
hope_NP0	5	negotiate_VVB	3	funds_NN2	9	promote_VVB	4
help_NN1	5	needed_VVN	3	fire_NN1	9	promise_VVI	4
harm_NN1	5	navigate_VVB	3	ethanol_NN1	9	prioritize_VVB	4
group_NN1	5	moves_VVZ	3	dreamers_NN2	9	prevented_VVN	4
germany_NN1	5	monitor_VVB	3	desk_NN1	9	prepared_VVD	4
georgia_NN1	5	mitigate_VVB	3	deputy_NN1	9	plays_VVZ	4
games_NN2	5	missing_VVG	3	coverage_NN1	9	pick_VVB	4
game_NN1	5	mining_VVG	3	climate_NN1	9	participating_VVG	4
friend_NN1	5	mention_VVI	3	cleveland_NN1	9	participate_VVB	4
frank_NP0	5	meets_VVZ	3	challenge_NN1	9	owned_VVD	4
founder_NN1	5	mark_VVI	3	career_NN1	9	oversight_VVI	4
football_NN1	5	manipulate_VVB	3	cara_NN1	9	oppose_VVB	4
food_NN1	5	managed_VVN	3	attention_NN1	9	offers_VVZ	4
folks_NN2	5	lived_VVD	3	application_NN1	9	offering_VVG	4
florida_NP0	5	link_VVI	3	advocates_NN2	9	occurred_VVD	4
firm_NN1	5	lie_VVI	3	accounts_NN2	9	lowering_VVG	4
film_NN1	5	led_VVN	3	woman_NN1	8	limits_VVZ	4
fan_NN1	5	leads_VVZ	3	webcast_NN1	8	letting_VVG	4
experts_NN2	5	launch_VVI	3	virginia_NN1	8	judged_VVD	4

effort_NN1	5	knew_VVD	3	usda_NN1	8	investigating_VVG	4
drama_NN1	5	kept_VVD	3	tribute_NN1	8	inspiring_VVG	4
dog_NN1	5	joins_VVZ	3	ted_NN1	8	injured_VVD	4
difference_NN1	5	issues_VVZ	3	taxpayer_NN1	8	ignore_VVB	4
defense_NN1	5	isnt_VVB	3	sunday_NN1	8	hosted_VVN	4
debt_NN1	5	install_VVB	3	stop_NN1	8	heard_VVN	4
damage_NN1	5	influencing_VVG	3	shooting_NN1	8	headed_VVN	4
cyber_NN1	5	influence_VVI	3	science_NN1	8	grew_VVD	4
crypto_NN1	5	increases_VVZ	3	repeal_NN1	8	grants_VVZ	4
county_NN1	5	ignoring_VVG	3	region_NN1	8	forced_VVD	4
connection_NN1	5	ignored_VVN	3	reduction_NN1	8	fear_VVI	4
congressman_NN1	5	ignore_VVB	3	rates_NN2	8	fail_VVB	4
compliance_NN1	5	hurry_VVB	3	prosperity_NN1	8	facilitate_VVB	4
community_NN1	5	hosting_VVG	3	prevention_NN1	8	express_VVB	4
communities_NN2	5	hoping_VVG	3	press_NN1	8	expanding_VVG	4
communications_NN2	5	hopes_VVZ	3	pockets_NN2	8	enter_VVB	4
code_NN1	5	highlights_VVZ	3	person_NN1	8	employs_VVZ	4
challenges_NN2	5	harvesting_VVG	3	oversight_NN1	8	eliminating_VVG	4
century_NN1	5	hacking_VVG	3	optimism_NN1	8	elected_VVN	4
captains_NN2	5	gone_VVN	3	olympics_NN1	8	eat_VVB	4
cannabis_NN1	5	gave_VVD	3	needs_NN2	8	deserved_VVD	4
candidate_NN1	5	funding_VVG	3	missiles_NN2	8	covered_VVD	4
campaigns_NN2	5	funded_VVN	3	memorial_NN1	8	cover_VVI	4
britain_NP0	5	fought_VVN	3	manufacturing_NN1	8	count_VVI	4
borders_NN2	5	focus_VVZ	3	loss_NN1	8	contribute_VVB	4
bombs_NN2	5	fits_VVZ	3	land_NN1	8	continued_VVD	4
bombings_NN2	5	fit_VVI	3	james_NP0	8	contains_VVZ	4
benefits_NN2	5	felt_VVD	3	isis_NN1	8	confirming_VVG	4
base_NN1	5	feature_VVI	3	innovation_NN1	8	confirm_VVB	4
banking_NN1	5	failed_VVN	3	hours_NN2	8	chosen_VVN	4
award_NN1	5	fail_VVB	3	hero_NN1	8	chose_VVD	4
art_NN1	5	expose_VVB	3	harvey_NN1	8	cheated_VVD	4
argument_NN1	5	explains_VVZ	3	hand_NN1	8	challenge_VVI	4
animals_NN2	5	expands_VVZ	3	hall_NN1	8	celebrates_VVZ	4
analysis_NN1	5	exists_VVZ	3	governors_NN2	8	caught_VVN	4
affairs_NN2	5	exist_VVB	3	george_NP0	8	catching_VVG	4
affair_NN1	5	enter_VVB	3	fusion_NN1	8	cancel_VVB	4
advisers_NN2	5	enjoying_VVG	3	forces_NN2	8	broken_VVN	4
additions_NN2	5	engage_VVB	3	fix_NN1	8	break_VVI	4
actors_NN2	5	enforce_VVB	3	facts_NN2	8	block_VVI	4
actions_NN2	5	ends_VVZ	3	equipment_NN1	8	blessed_VVD	4
account_NN1	5	enacts_VVZ	3	editorial_NN1	8	benefitting_VVG	4
accidents_NN2	5	educate_VVB	3	donald_NN1	8	attend_VVB	4
york_NP0	5	ease_VVI	3	dollars_NN2	8	approaches_VVZ	4
workers_NN2	5	earn_VVB	3	dept_NN1	8	agreed_VVD	4

window_NN1	*5*	dropped_VVD	*3*	delegation_NN1	*8*	adding_VVG	*4*
welfare_NN1	*5*	doubt_VVI	*3*	debate_NN1	*8*	acknowledge_VVB	*4*
web_NN1	*4*	dig_VVB	*3*	cut_NN1	*8*	abandon_VVB	*4*
warnings_NN2	*4*	die_VVB	*3*	creation_NN1	*8*	wishing_VVG	*3*
victoria_NN1	*4*	destroyed_VVN	*3*	counsel_NN1	*8*	went_VVD	*3*
venue_NN1	*4*	denying_VVG	*3*	corporations_NN2	*8*	walking_VVG	*3*
values_NN2	*4*	delve_VVB	*3*	constitution_NN1	*8*	usher_VVI	*3*
users_NN2	*4*	delivered_VVN	*3*	constituent_NN1	*8*	uses_VVZ	*3*
updates_NN2	*4*	delayed_VVD	*3*	citizens_NN2	*8*	urges_VVZ	*3*
university_NN1	*4*	defeat_VVI	*3*	changes_NN2	*8*	urged_VVD	*3*
type_NN1	*4*	decided_VVD	*3*	bonus_NN1	*8*	tried_VVD	*3*
trends_NN2	*4*	dealing_VVG	*3*	bob_NN0	*8*	travelling_VVG	*3*
treatment_NN1	*4*	criticized_VVN	*3*	approach_NN1	*8*	travel_VVI	*3*
transparency_NN1	*4*	criticize_VVB	*3*	advocate_NN1	*8*	trained_VVN	*3*
testimony_NN1	*4*	covers_VVZ	*3*	word_NN1	*7*	trafficking_VVG	*3*
terror_NN1	*4*	cost_VVB	*3*	wolf_NN1	*7*	trade_VVI	*3*
summer_NN1	*4*	control_VVI	*3*	winners_NN2	*7*	track_VVI	*3*
student_NN1	*4*	considers_VVZ	*3*	wing_NN1	*7*	touring_VVG	*3*
statements_NN2	*4*	considered_VVN	*3*	waivers_NN2	*7*	threatens_VVZ	*3*
staffers_NN2	*4*	conned_VVD	*3*	wage_NN1	*7*	threaten_VVB	*3*
squatter_NN1	*4*	confirms_VVZ	*3*	vision_NN1	*7*	testified_VVN	*3*
speaker_NN1	*4*	compromise_VVI	*3*	vietnam_NN1	*7*	tear_VVI	*3*
space_NN1	*4*	compares_VVZ	*3*	video_NN1	*7*	sworn_VVN	*3*
son_NN1	*4*	collected_VVN	*3*	university_NN1	*7*	supports_VVZ	*3*
solutions_NN2	*4*	collect_VVB	*3*	understanding_NN1	*7*	supported_VVN	*3*
solution_NN1	*4*	claiming_VVG	*3*	truth_NN1	*7*	suggest_VVB	*3*
shooting_NN1	*4*	cheating_VVG	*3*	tournament_NN1	*7*	suffered_VVN	*3*
shirt_NN1	*4*	catch_VVI	*3*	testimony_NN1	*7*	submit_VVB	*3*
shares_NN2	*4*	carrying_VVG	*3*	terrorism_NN1	*7*	streamlining_VVG	*3*
share_NN1	*4*	carry_VVB	*3*	tech_NN1	*7*	stops_VVZ	*3*
sectors_NN2	*4*	campaign_VVI	*3*	steel_NN1	*7*	stood_VVD	*3*
sector_NN1	*4*	buying_VVG	*3*	spring_NN1	*7*	stick_VVI	*3*
science_NN1	*4*	butt_VVI	*3*	son_NN1	*7*	stepped_VVD	*3*
sacramento_NP0	*4*	broke_VVD	*3*	soldiers_NN2	*7*	step_VVI	*3*
rule_NN1	*4*	bringing_VVG	*3*	sides_NN2	*7*	spy_VVI	*3*
room_NN1	*4*	breaks_VVZ	*3*	shooter_NN1	*7*	sponsored_VVD	*3*
review_NN1	*4*	boycott_VVB	*3*	session_NN1	*7*	sitting_VVG	*3*
responsibility_NN1	*4*	block_VVI	*3*	servant_NN1	*7*	shares_VVZ	*3*
resignation_NN1	*4*	blew_VVD	*3*	scout_NN1	*7*	sends_VVZ	*3*
republican_NN1	*4*	blaming_VVG	*3*	saturday_NN1	*7*	sell_VVB	*3*
republic_NN1	*4*	begun_VVN	*3*	rock_NN1	*7*	selected_VVD	*3*
reporting_NN1	*4*	banning_VVG	*3*	ribbon_NN1	*7*	secured_VVD	*3*
release_NN1	*4*	ban_VVI	*3*	responsibility_N	*7*	saves_VVZ	*3*

				N1			
regulation_NN1	*4*	backing_VVG	*3*	relationship_NN1	*7*	saddened_VVN	*3*
reelection_NN1	*4*	attended_VVD	*3*	reasons_NN2	*7*	runs_VVZ	*3*
provisions_NN2	*4*	approaching_VVG	*3*	reagan_NN1	*7*	roll_VVB	*3*
properties_NN2	*4*	appreciate_VVB	*3*	race_NN1	*7*	ride_VVI	*3*
products_NN2	*4*	applied_VVD	*3*	question_NN1	*7*	reviews_VVZ	*3*
problems_NN2	*4*	aimed_VVD	*3*	psalms_NN2	*7*	revealed_VVD	*3*
poll_NN1	*4*	aim_VVI	*3*	protections_NN2	*7*	return_VVI	*3*
policies_NN2	*4*	affected_VVN	*3*	projects_NN2	*7*	resolved_VVN	*3*
places_NN2	*4*	advocate_VVI	*3*	producers_NN2	*7*	resist_VVB	*3*
photographer_NN1	*4*	admitted_VVN	*3*	pressure_NN1	*7*	requires_VVZ	*3*
petition_NN1	*4*	admit_VVB	*3*	politics_NN1	*7*	represented_VVN	*3*
pete_NN1	*4*	addressed_VVD	*3*	partnership_NN1	*7*	reported_VVD	*3*
pennsylvania_NP0	*4*	acting_VVG	*3*	organization_NN1	*7*	repeals_VVZ	*3*
paywall_NN1	*4*	account_VVI	*3*	opposition_NN1	*7*	release_VVB	*3*
pay_NN1	*4*	accepted_VVN	*3*	opening_NN1	*7*	rein_VVI	*3*
partner_NN1	*4*	wrote_VVD	*2*	northey_NN1	*7*	recorded_VVN	*3*
pakistan_NN1	*4*	writing_VVG	*2*	modernization_NN1	*7*	recognizes_VVZ	*3*
overview_NN1	*4*	writes_VVZ	*2*	minutes_NN2	*7*	recognized_VVN	*3*
needs_NN1	*4*	worked_VVN	*2*	military_NN1	*7*	reauthorize_VVB	*3*
movement_NN1	*4*	withdraw_VVB	*2*	miles_NN2	*7*	questioned_VVD	*3*
mother_NN1	*4*	wishes_VVZ	*2*	mike_NN1	*7*	qualified_VVN	*3*
misuse_NN1	*4*	was_VBN	*2*	mcconnell_NN1	*7*	push_VVB	*3*
mississippi_NN2	*4*	wake_VVI	*2*	management_NN1	*7*	protected_VVD	*3*
minute_NN1	*4*	votes_VVZ	*2*	majority_NN1	*7*	proposed_VVD	*3*
meeting_NN1	*4*	voicing_VVG	*2*	mainers_NN2	*7*	projects_VVZ	*3*
mcconnell_NP0	*4*	violated_VVD	*2*	loans_NN2	*7*	prohibits_VVZ	*3*
marijuana_NN1	*4*	veto_VVI	*2*	liberals_NN2	*7*	prioritized_VVD	*3*
map_NN1	*4*	urges_VVZ	*2*	level_NN0	*7*	preview_VVI	*3*
manager_NN1	*4*	urged_VVD	*2*	leave_NN1	*7*	preventing_VVG	*3*
local_NN1	*4*	upset_VVI	*2*	lawyer_NN1	*7*	praise_VVI	*3*
list_NN1	*4*	uproar_VVI	*2*	lawmakers_NN2	*7*	pleased_VVD	*3*
links_NN2	*4*	understands_VVZ	*2*	judges_NN2	*7*	placed_VVN	*3*
lines_NN2	*4*	uncover_VVB	*2*	joe_NN1	*7*	partner_VVI	*3*
liberty_NN1	*4*	tweeting_VVG	*2*	intro_NN1	*7*	owe_VVB	*3*
legislature_NN1	*4*	tweak_VVI	*2*	integrity_NN1	*7*	overturn_VVB	*3*
lady_NN1	*4*	turns_VVZ	*2*	input_NN1	*7*	overcome_VVB	*3*
lack_NN1	*4*	treated_VVD	*2*	industries_NN2	*7*	ought_VM0	*3*
justice_NN1	*4*	treat_VVI	*2*	hospital_NN1	*7*	opened_VVD	*3*

jokes_NN2	*4*	trading_VVG	*2*	holiday_NN1	*7*	notes_VVZ	*3*
jimmy_NP0	*4*	touching_VVG	*2*	harassment_NN1	*7*	nominated_VVD	*3*
jail_NN1	*4*	tossed_VVN	*2*	guard_NN1	*7*	modernizing_VVG	*3*
iowa_NN1	*4*	top_VVI	*2*	grant_NN1	*7*	matter_VVI	*3*
incompetence_NN1	*4*	tolerate_VVB	*2*	google_NN1	*7*	marking_VVG	*3*
illinois_NN1	*4*	tighten_VVB	*2*	goal_NN1	*7*	maintain_VVB	*3*
ignorance_NN1	*4*	threaten_VVB	*2*	gift_NN1	*7*	loved_VVD	*3*
housing_NN1	*4*	terrified_VVD	*2*	friday_NP0	*7*	lies_VVZ	*3*
hour_NN1	*4*	teach_VVB	*2*	fraud_NN1	*7*	leaves_VVZ	*3*
hosts_NN2	*4*	targeting_VVG	*2*	forest_NN1	*7*	leave_VVB	*3*
harassment_NN1	*4*	tackle_VVI	*2*	failure_NN1	*7*	leaks_VVZ	*3*
guys_NN2	*4*	sworn_VVN	*2*	expenses_NN2	*7*	landed_VVD	*3*
guest_NN1	*4*	sway_VVI	*2*	evidence_NN1	*7*	laid_VVN	*3*
governance_NN1	*4*	survive_VVB	*2*	events_NN2	*7*	kill_VVB	*3*
gandhi_NN2	*4*	surrounding_VVG	*2*	enemies_NN2	*7*	jump_VVI	*3*
fund_NN1	*4*	surprised_VVD	*2*	employers_NN2	*7*	joins_VVZ	*3*
francisco_NN1	*4*	supported_VVD	*2*	doctors_NN2	*7*	invite_VVB	*3*
france_NN1	*4*	suffering_VVG	*2*	cybersecurity_NN1	*7*	investigate_VVB	*3*
fight_NN1	*4*	sucking_VVG	*2*	cyber_NN1	*7*	intercept_VVB	*3*
eyes_NN2	*4*	strive_VVB	*2*	crime_NN1	*7*	informed_VVD	*3*
event_NN1	*4*	stood_VVD	*2*	creators_NN2	*7*	improving_VVG	*3*
environment_NN1	*4*	stolen_VVN	*2*	construction_NN1	*7*	imagine_VVB	*3*
enemies_NN2	*4*	stole_VVD	*2*	comey_NN1	*7*	identify_VVB	*3*
employers_NN2	*4*	starts_VVZ	*2*	collins_NN2	*7*	housing_VVG	*3*
east_NN1	*4*	stands_VVZ	*2*	colleague_NN1	*7*	highlighted_VVD	*3*
drugs_NN2	*4*	spreading_VVG	*2*	chamber_NN1	*7*	highlight_VVI	*3*
donny_NP0	*4*	spread_VVI	*2*	cause_NN1	*7*	hidden_VVN	*3*
dinner_NN1	*4*	spot_VVI	*2*	cabinet_NN1	*7*	heading_VVG	*3*
dictators_NN2	*4*	sponsor_VVI	*2*	bruce_NN1	*7*	happens_VVZ	*3*

Appendix 5

Samples

Figure 5.1 **Sample of raw corpus data**

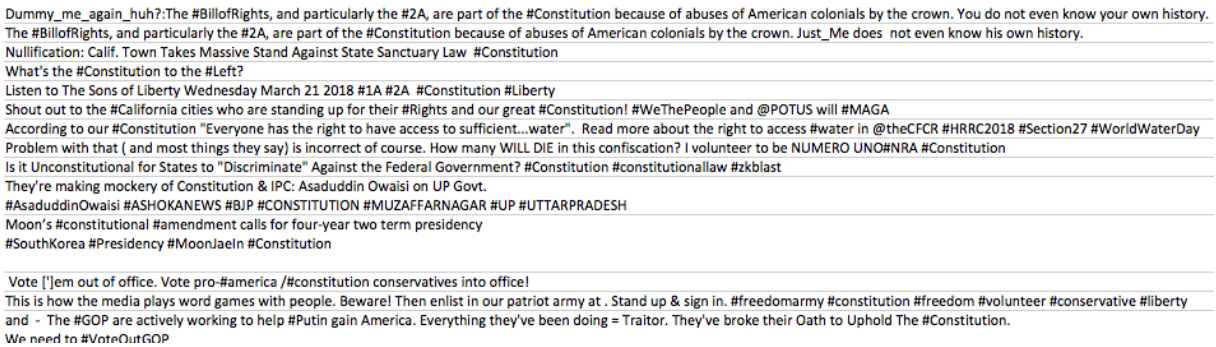

Figure 5.2 **Sample of tagged corpus data**

```
-----_PUN  Sadly_AV0 ,_PUN we_PNP see_VVB evidence_NN1 every_AT0 Day_NN1 that_CJT the_AT0  #GOP_UNC is_VBZ
n't_XX0 representing_VVG the_AT0 American_AJ0 people_NN0  or_CJC the_AT0 #Constitution_UNC either_AV0 !
_SENT -----_PUN  We_PNP need_VVB politicians_NN2 to_TO0 sign_VVI contracts_NN2 !_SENT -----_PUN  They_PNP
must_VM0 be_VBI held_VVN accountable_AJ0 ._SENT -----_PUN  Americans_NN2 are_VBB tired_AJ0 of_PRF
#Lies_UNC and_CJC #LipService_UNC  during_PRP campaigns_NN2 ._SENT -----_PUN  We_PNP must_VM0 also_AV0
Impose_VVB #TermLimits_UNC I_ZZ0 try_VVB not_XX0  to_TO0 get_VVI political_AJ0 here_AV0 but_CJC I_PNP
dare_VM0 say_VVI that_CJT  anyone_PNI that_CJT comes_VVZ to_PRP my_DPS home_NN1 to_TO0 take_VVI away_AV0
my_DPS guns_NN2 will_VM0 quickly_AV0 find_VVI out_AVP why_AVQ the_AT0  #2ndAmendment_NN0 was_VBD
written_VVN into_PRP the_AT0 #Constitution_UNC  Dummymeagainhuh_NN1 ?_PUN :_SENT -----_PUN  The_AT0
#BillofRights_UNC ,_PUN and_CJC particularly_AV0 the_AT0 #2A_NN0 ,_PUN  are_VBB part_NN1 of_PRF the_AT0
#Constitution_UNC because_PRP of_PRP  abuses_NN2 of_PRF American_AJ0 colonials_NN2 by_PRP the_AT0
crown_NN1 ._SENT  -----_PUN  You_PNP do_VDB not_XX0 even_AV0 know_VVI your_DPS own_DT0 history_NN1 ._SENT
-----_PUN  The_AT0 #BillofRights_UNC ,_PUN and_CJC particularly_AV0 the_AT0 #2A_NN0 ,_PUN  are_VBB
part_NN1 of_PRF the_AT0 #Constitution_UNC because_PRP of_PRP  abuses_NN2 of_PRF American_AJ0 colonials_NN2
by_PRP the_AT0 crown_NN1 ._SENT  -----_PUN  JustMe_NN1 does_VDZ not_XX0 even_AV0 know_VVI his_DPS own_DT0
history_NN1  ._SENT -----_PUN  Nullification_NP0 :_PUN Calif_NP0 ._SENT -----_PUN  Town_NN1 Takes_VVZ
Massive_AJ0 Stand_NN1 Against_PRP State_NN1 Sanctuary_NN1  Law_NN1 #Constitution_UNC What_DTQ 's_VBZ
the_AT0 #Constitution_UNC to_PRP  the_AT0 #Left_UNC ?_SENT -----_PUN  Listen_VVB to_PRP The_AT0 Sons_NN2
of_PRF Liberty_NN1 Wednesday_NP0 March_NP0  21_CRD 2018_CRD #1A_NN0 #2A_NN0 #Constitution_UNC #Liberty_UNC
Shout_VVB  out_AVP to_PRP the_AT0 #California_UNC cities_NN2 who_PNQ are_VBB standing_VVG  up_AVP for_PRP
their_DPS #Rights_UNC and_CJC our_DPS great_AJ0  #Constitution_UNC !_PUN #WeThePeople_UNC and_CJC
@POTUS_UNC will_VM0 #MAGA_UNC  According_PRP to_PRP our_DPS #Constitution_UNC "_PUQ Everyone_PNI has_VHZ
the_AT0 right_NN1 to_TO0 have_VHI access_NN1 to_PRP sufficient_AJ0 ..._PUN  water_NN1 "_PUQ ._SENT -----
_PUN  Read_VVB more_AV0 about_PRP the_AT0 right_NN1 to_TO0 access_VVI #water_UNC  in_PRP @theCFCR_UNC
#HRRC2018_NN0 #Section27_NN0 #WorldWaterDay_UNC  Problem_NN1 with_PRP that_DT0 (_PUL and_CJC most_DT0
things_NN2 they_PNP  say_VVB )_PUR is_VBZ incorrect_AJ0 of_AV0 course_AV0 ._SENT -----_PUN  How_AVQ
```

Figure 5.3 **Sample of extracted results**

64	es_NN2 Act_	Act_	VVB	Act_VVB	VVB abuse_N	verified_tagg
65	es_NN2 Act_	Act_	VVB	Act_VVB	VVB with_PR	verified_tagg
66	ss_NN1 act_	act_	VVB	act_VVB	VVB all_DT0	verified_tagg
67	os_NN2 Act_	Act_	VVB	Act_VVB	VVB at_PRP t	verified_tagg
68	os_NN2 Act_	Act_	VVB	Act_VVB	VVB passes_	verified_tagg
69	os_NN2 Act_	Act_	VVB	Act_VVB	VVB deliveri	verified_tagg
70	os_NN2 Act_	Act_	VVB	Act_VVB	VVB modern	verified_tagg
71	os_NN2 Act_	Act_	VVB	Act_VVB	VVB to_TO0	verified_tagg
72	s_NN2 Act_	Act_	VVB	Act_VVB	VVB which_D	verified_tagg
73	os_NN2 Act_	Act_	VVB	Act_VVB	VVB Another	verified_tagg
74	os_NN2 Act_	Act_	VVB	Act_VVB	VVB ,_PUN I_	verified_tagg
75	ns_NN2 Act_	Act_	VVB	Act_VVB	VVB Check_'	verified_tagg
76	ty_NP0 Act_	Act_	VVB	Act_VVB	VVB &; the_A	verified_tagg
77	we_PNP act_	act_	VVB	act_VVB	VVB now_AV	verified_tagg
78	-_PUN ACT_	ACT_	VVB	ACT_VVB	VVB NOW_A	verified_tagg
79	)_PUR Act_	Act_	VVB	Act_VVB	VVB "_PUQ t	verified_tagg
80	to_TO0 act_	act_	VVI	act_VVI	VVI ._SENT -	verified_tagg

Dokumentārā lapa

Bakalaura darbs „Parts of Speech in the Messages of the Microblog ‘Twitter.com’ (Vārdu šķiru lietojums mikrobloga vietnes twitter.com ziņojumos)” izstrādāts LU Humanitāro zinātņu fakultātē.

Ar savu parakstu apliecinu, ka pētījums veikts patstāvīgi, izmantoti tikai tajā norādītie informācijas avoti un iesniegtā darba elektroniskā kopija atbilst izdrukai.

Autors: Elza Upmane …. …. 2018.

Rekomendēju/nerekomendēju darbu aizstāvēšanai

Vadītāja: asoc. prof. Zigrīda Vinčela …. …. 2018.

Recenzents:

Studiju metodiķe: …. …. 2018.

Darbs iesniegts Anglistikas nodaļā …. …. 2018.

Darbu pieņēma:

Darbs aizstāvēts bakalaura gala pārbaudījuma komisijas sēdē

2018. gada…… jūnijā, prot. Nr. ….., vērtējums ……..

Komisijas sekretāre: