FAKULTA
PŘÍRODOVĚDNĚ-HUMANITNÍ
A PEDAGOGICKÁ TUL

Bakalářská práce

Corpus analysis of selected lexemes connected with modern technology

Studijní program:	B0114A300068 Anglický jazyk se zaměřením na vzdělávání
Studijní obory:	Anglický jazyk se zaměřením na vzdělávání Německý jazyk se zaměřením na vzdělávání
Autor práce:	**Rudolf Hýrek**
Vedoucí práce:	Mgr. Jaromír Haupt, Ph.D. Katedra anglického jazyka

Liberec 2023

FAKULTA
PŘÍRODOVĚDNĚ-HUMANITNÍ
A PEDAGOGICKÁ TUL

Zadání bakalářské práce

Corpus analysis of selected lexemes connected with modern technology

Jméno a příjmení:	**Rudolf Hýrek**
Osobní číslo:	P19001022
Studijní program:	B0114A300068 Anglický jazyk se zaměřením na vzdělávání
Specializace:	Anglický jazyk se zaměřením na vzdělávání Německý jazyk se zaměřením na vzdělávání
Zadávající katedra:	Katedra anglického jazyka
Akademický rok:	2020/2021

Zásady pro vypracování:

Cílem práce je prozkoumat kolokace výrazů, které označují moderní technologie nebo s nimi souvisejí. Zvláštní pozornost bude věnována problematickému výskytu předložek ve spojeních s podstatnými jmény. Student použije korpusový nástroj SketchEngine a porovná četnosti výskytů vybraných alternujících variant. Dále prozkoumá kontexty výskytů z hlediska syntaktického, sémantického a stylistického.

Rozsah grafických prací:	
Rozsah pracovní zprávy:	
Forma zpracování práce:	tištěná/elektronická
Jazyk práce:	Angličtina

Seznam odborné literatury:

Crystal, David. 2006. *Language and the Internet*. Cambridge: Cambridge University Press.
Dušková, Libuše. 1988. *Mluvnice současné angličtiny na pozadí češtiny.* Praha: Academia.
James, Kate. 2010. ”A Corpus-Based Case Study in Prepositional AT/TO-Infinitive Alternation using the Lemma ’AIM’.” *Lexis – Journal in English Lexicology*, no. 5, https://doi.org/10.4000/lexis.445.
Stewart, Dominic. 2009. *Semantic Prosody: A Critical Evaluation*. New York/London: Routledge.
Zufferey, Sandrine. 2020. *Introduction to Corpus Linguistics*. London: ISTE.

Vedoucí práce:	Mgr. Jaromír Haupt, Ph.D. Katedra anglického jazyka
Datum zadání práce:	30. června 2021
Předpokládaný termín odevzdání:	15. července 2022

L.S.

prof. RNDr. Jan Picek, CSc.
děkan

Mgr. Zénó Vernyik, Ph.D.
vedoucí katedry

V Liberci dne 30. června 2021

Prohlášení

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

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

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

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

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

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

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

27. dubna 2023 Rudolf Hýrek

Annotation

The subject of this thesis is examination of prepositional phrases from the field of modern technologies. According to the presence of more than one variant of using prepositions in connection with a particular modern technology such as *in / on your screen* etc., there is doubt which of these variants is more frequently used, therefore the analysis answers the question, which one of the existing options is more naturally used by users of English. In the theoretical part, basic morphological, syntactical and corpus linguistics concepts were introduced. Not only the concepts but also information and data that the practical part cannot do without were introduced and well explained. In the practical part there are two main segments - methodology and data. In the methodology there are above all the procedure and concrete steps that were done to make the whole corpus based research possible. In the data section the outputs in a form of numbers were analyzed and concrete outcomes of the research were introduced. The study is based on corpus data provided by the Sketch Engine and not only employs pieces of knowledge from the area of corpus linguistics, but it also applies traditional linguistics sciences such as morphology, syntax. At the end the research question was fulfilled and it was shown which of the given variants of phrases in / your screen, in / on / at Wikipedia are more frequently used. Not only the results in the form of order from the point of view of the frequency particular varieties are used but also the types of invalid results were provided to readers.

Keywords

corpus linguistics, syntaxis, semantics, analysis, prepositional phrases, modern technologies, Sketch Engine, noun phrases, prepositions

Anotace

Předmětem této práce je analýza předložkových frází z oblasti moderních technologií. Vzhledem k přítomnosti více než jedné předložky ve spojení s určitou moderní technologií jako například *in / on your screen* atd. se objevují nejasnosti, která z těchto variant je více a též přirozeněji používaná mluvčími anglického jazyka, a proto se tato analýza zabývá touto otázkou, a i na ní odpovídá.

V teoretické části byly představeny základní morfologické, syntaktické a korpusovo-lingvistické koncepty. Nejen koncepty, bez nichž se tato práce neobejde, ale i data byly představeny a pečlivě vysvětleny.

Tato analýza je založena na korpusových datech zprostředkovaných programem Sketch Engine a nejen využívá znalosti z pole korpusové lingvistiky, ale také aplikuje poznatky z tradičních lingvistických věd jako je morfologie či syntax. Vzhledem k tomuto faktu je tato analýza příkladem komplexní aplikace lingvistických poznatků. Na závěr byla zodpovězena vytyčená otázka výzkumu a bylo ukázáno, která z daných variant *in / on your screen, in / on / at Wikipedia* je častěji používaná. Nejen výsledky ve formě pořadí dle frekvence užití, ale také typy nevalidních výsledků byly zprostředkovány čtenáři.

Klíčová slova

korpusová lingvistika, syntax, sémantika, analýza, předložkové vazby, moderní technologie, Sketch Engine, jmenné skupiny, předložky

Acknowledgement

I would like to extend my sincere thanks to my supervisor Dr. Jaromír Haupt, without whom would not be even possible to start writing this thesis, for his guidance, asistance, effort and endless patience. I also owe many thanks to my English teacher MA. Alena Halamová thanks to whom I started my long journey of studying English. I am also very thankful to Dr. Peldová for her corpus linguistic education and Dr. Madson for his seminars of syntax and valuable materials.

Table of contents

List of tables

Theoretical part

1 Introduction

The thesis "Corpus analysis of selected lexemes connected with modern technology" is focused on selected pieces of modern technologies in connection with particular prepositions.

The reason to do this whole analysis is to find out which of the alternatives – e.g. *on / in / at + Wikipedia* or *on / in + screen* etc. is more frequently used or not and provide well sustained evidence in the form of data analyzed in the way as it is stated below, because in many cases there are doubts, which preposition is convenient to use when talking about particular technology in a sense of an adverbial of place. Sometimes there are even more than two prepositions that are used when referring to a particular piece of technology in a sense of an adverbial of place.

The irrelevant results, those which do not fit the demanded function: e.g. *on / in / at + Wikipedia* in a sense of adverbial of place, will also be dealt with in the thesis because undoubtedly to aim at the difficulties of this corpus analysis and also show the key used for sorting out the results provided by Sketch Engine appears to be greatly important. The key will be, as it was outlined earlier, the relevance of a particular sample from the point of view of its syntactic structure or semantics. At the end of this section, the data will be statistically analyzed. They will be sorted according to their relevance into particular subtypes. This will be done because of emphasis on difficulties while doing the corpus analysis and what is needed to be cautious about if one wants to obtain the demanded data. After subtraction of all irrelevant results of each version of a structure e.g. *on / in/ at Wikipedia* there should be relevant statistical outcomes and also the aim of the thesis should be reached.

2 Noun phrases

In the case of studying prepositions in connection with modern technologies, the knowledge of particular prepositions is not sufficient for the analysis of samples provided by the Sketch Engine. To analyze the data well, the knowledge of noun phrases is needed in order to see the relations between individual elements inside a noun phrase, e.g. *in your screen brightness.* In this example it is apparent that not every phrase given has a meaning of an adverbial of place referring to the surface of a screen as stated before. The phrase mentioned has the following structure: preposition + determiner (possessive pronoun) + descriptor + head (noun). In comparison to an example of a phrase whose meaning is different from the demanded (screen as an adverbial of place) examples such as *I struggled a bit to get the same view and setup as shown in your screen* meet criteria.

To provide a well done analysis of the prepositions connected with particular modern technologies and their alternatives and find out which of these alternatives is a more frequent and more relevant one, a decent knowledge of noun phrases cannot be omitted. Without having a good understanding, what noun phrases are, the chance of doing an analysis of an evidential value seems to be slightly unfeasible.

According to Madsen, a noun phrase is a grouping of words that consists of a head, premodifiers and postmodifiers. The head of a noun phrase can be a noun, pronoun *a new house* (noun), *someone to adore* (pronoun). The head is a nucleus of a noun phrase, therefore it cannot be omitted. A phrase would not be a phrase if it loses its nucleus, which is the central component. The elements preceding the head are called premodifiers. Those which are located behind the head are called the postmodifiers.

The noun phrase can be composed of both premodifier and postmodifier, more of these both elements, or even one of these. Premodifiers can be further divided into predeterminers (*all, both, many, half, what, such multipliers*), determiners (articles - indefinite, definite, demonstratives pronouns - *this, these, that*, quantifiers - noun phrases, numerals, descriptors - *adjective phrases, noun phrases, noun phrase in genitive*). The postmodifiers can be further divided into noun phrases, adverbial phrases, adjective phrases, relative clauses, appositive clauses or complement clauses. If there is a preposition before a noun phrase, it will be marked as a prepositional phrase (Madsen 2022, 56).

In contrary to the division of function words specifying the reference of a noun introduced above, Biber et al. claim that there are three groups of determiners: *predeterminers (all, both, half and multipliers)*, *central determiners (articles, demonstrative determiners and possessive determiners)* and *postdeterminers* including two subgroups (ordinal numerals a semideterminers and the semideterminers *same, other, former, latter, last, and next)*; cardinal numerals and quantifying determiners (Biber et al. 1999, 258).

The division of function words by Biber et al. seems to be more general at the beginning, because of only three major groups: *predeterminers, central determiners and postdeterminers*, which are further subdivided into more detailed groups. The principle of both divisions by Madsen; Biber et al. seems to be basically the same, but only described and subdivided in a different way.

3 Prepositional phrases

For the analysis of prepositions connected with modern technologies the knowledge of prepositional phrases appears to be absolutely essential, because in principle it could be said that prepositional phrases are noun phrases expanded by prepositions. The knowledge of noun phrases is a basis for the determining of the syntactic function of words referring to a particular piece of technology inside a sentence, whereas the prepositional phrases are crucial to mark a particular piece of technology in connection with a particular preposition as a whole.

According to Biber, prepositional phrases are composed of a preposition and a complement. The complement is typically in the form of a noun phrase. The characteristic prepositional phrase might be perceived as a noun phrase enlarged of a link that illustrates the relationship to neighboring structures. In bold there is complement in the following examples: *to **town,** in **the morning**, to **him**, on **the night [of the first day]**, in **a street [with no name]*** (Biber et al. 1944, 103), In a sentence, they can occupy different functions. Dušková claims that the most common function is a function of an adverbial a) of place: J*ohn lives on the third floor. Sue is from England.* b) of time: *I get up at seven o clock. I will be at home by midnight.* c) of manner: *The exact value can be found in the following way*. *You can remove dirt by using a brush.* d) of cause: *She asked only out of curiosity.* e) adverbial of attendant circumstances: *The picture fell with a crash.* f) adverbial of affirmation: *She is tall for her* age (Dušková 1994, 277).

Prepositional phrases can also serve as an object of a verb: *The situation calls for prompt action, as an object of an adjective: I am sorry for his parents.,* or subject or object complement: *In this country, Jones passes for a clever scientist.* Occurrence of this prepositional type is really limited. These are partly fixed phrases, such as: *I take it for granted*. (Dušková 1994, 277). Most of the verbs that require a complement as a compulsory completion is a complement without a preposition or using *as*: *He died a rich man*. *We consider these measures unnecessary* x *We regard these measures as unnecessary* (Dušková 1994, 278).

3.1 Prepositions

For the analysis of particular modern technology connected with prepositions, the fundamental knowledge of prepositions is necessary to understand their basic meaning. There are many prepositions in English and also many criteria for their classification. The difficulties connected with a large number of exceptions in use of prepositions has to be taken into account. As Tunaz, Muyan and Muratoglu claim, "The proper use of prepositions in English is of perennial concern to the linguistics field in general. Sometimes the semantics of prepositions is rather a frame than a strict rule of usage" (Tunaz, Muyan and Muratoglu, 2016, 1).

Slightly different is a definition of Biber, who adds his own definition of prepositions: "Prepositions are links which introduce prepositional phrases. As the most typical complement in a prepositional phrase is a noun phrase, they can be regarded as a device which connects noun phrases with other structures. Many prepositions in English correspond to case inflections in other languages" (Biber et al.1944, 74).

Dušková claims that as a word class, prepositions are counted as empty words (also marked as grammatical or function words). They do not form an independent sentence constituent, however they form one sentence constituent together with a syntactical / compositional noun. They express relations between a syntactical /

compositional noun and other words - between two nouns, verb and noun and adjective and noun, such *as the fight for peace, look at it, he was proud of his son.*

The prepositions can be divided according to their form into one syllable prepositions or simple e.g. *on, in, at, from, of, to, up* and more syllable prepositions or complex e.g. *except for, with regard to, in connection with.* From a point of view of origin, prepositions can be marked as proper or derived. Proper prepositions e.g. *at, by, for, from, of, on, to, with* stand on their own and have only one component whereas derived prepositions e.g. *about, aboard, across, after, alongside, apart, around, before, beneath, besides, below, beyond, down, notwithstanding* are composed of components that are derived from another word class.

According to their semantics, they can be divided into prepositions of place, movement and direction. Some of the prepositions can be included in more than one semantic group. Their usage is also universal, and their meaning can differ in various contexts, therefore a membership of different groups might also be different. Prepositions together with noun phrases can form prepositional phrases: *on the first floor, in your house, in my opinion, on your computer screen.* The number of prepositions is not stable. Some new prepositions arise, whereas the other prepositions vanish. In most cases, derived prepositions come into existence by combining a preposition with a noun and another preposition (Dušková 1994, 273-278).

3.1.1 Prepositions of place

To be able to compare two or more varieties of a particular prepositional phrase: *in your screen x on your screen, on Wikipedia x in Wikipedia x at Wikipedia* referring to a particular piece of technology in a sense of an adverbial of place, the knowledge of basic prepositions of place *on, in and at* is crucial to know to which are these prepositions referring to. Dušková and Strnadová introduce more theoretical rather than practical information than for example Swan. Only basic principles rather than concrete examples are presented by Dušková (273-302).

Dušková and Strnadová mention that static localization and dynamical localization are mostly distinguished in the field of prepositions. For determining a type of localization, auxiliary questions can be used. The static localization can be asked using the adverb *where? Mrs. Black stayed at home. (Where did he stay?) She was in a pub*. (Where was she?) We stayed at the entrance. (Where were we?). In case of dynamical localization, there are more possibilities to ask. We can use where*...from: We returned from the theater*. (Where did we return from?)*, which way: We went alongside the river. (which way did we go?) and where (to): We went to Stratford. (Where did we go to?).* Because of a significant semantic difference in case of the same preposition, such as: *He is at school. X He arrived.* (Dušková 1994, 279-280).

Dušková's and Strnadová's conception of prepositions is more abstract and philosophical than concrete and practical. In contrast to the Czech authors Dušková and Strnadová the British author Michael Swan describes the prepositions *on, in, at* from the usage point of view. Swan states, that"preposition *on, in* and *at* belong to frequently used prepositions of place. *On* is used to talk about position on a line (for example a road or a river): *His house is on the way from Aberdeen to Dundeen. Stratford is on the river Avon.* It is also used for position on a surface: *Hurry up - supper is on the table! The picture would look better on the other wall. There is a big spider on the ceiling*. It can also mean attracted to: *Why do you wear that ring on your first finger? There are not many apples on the tree this year*. To talk about position by a lake or a sea *on* is also used: *Bowness is on Lake Windermere. Southend-on-sea.*

In is used for positions inside large areas, and in three-dimensional space (when something is surrounded on all sides): I don't think he’s in his office. Let’s go for a walk in the woods. She grew up in Swaziland. I last saw her in the car park. He lived in the desert for three years. *In* is also used for the position of things which form part of the line: T*here’s a misprint in line 6 on page 22. Who’s the good-looking boy in the sixth row?* When talking about public transport we use on if talking about travel using public transport (buses, trains, planes and boats), as well as (motor) cycles and horses: *There’s no room on the bus. He’s arriving on the 3.15 train. We’re booked on flight 604. It took five days to cross the Atlantic on the Queen Elizabeth. I’ll go down the shop on my bike.* In connection with cars and small private planes and boats we use *in:* She came in a taxi. He fell into the river when he was getting out of his canoe. The preposition *at* is usually used after a verb arrive, whereas *in* is used before very large places: *He arrives at the airport at 15.30. What time do we arrive in New York?* In case of referring to addresses, all the three chosen prepositions (*on, in, at)* are used. To talk generally about addresses *at* is used: *Are you still at the same address? She lives at 73 Albert Street.* In American English when giving the name of the street it is possible to use *in* instead of *on*: *She lives in Albert Street.* To talk about the number of a floor *on* is used: She lives in a flat on the third floor. *At* can be used with a possessive to mean “at somebody's house or shop”: *Where’s Jane? She’s round at Pat’s. You’re always at the hairdresser's. “*(Swan 2016, 72-74).

4 Corpus and its significance

“In the language sciences, a corpus is a body of written text or transcribed speech which can serve as a basis for linguistic analysis and description. Over the last three decades, the compilation and analysis of corpora stored in computerized databases has led to a new scholarly enterprise known as corpus linguistics” (Kennedy 1998, 3). Kennedy also defines a corpus to be a collection of texts in an electronic database and also says that electronic databases beg many questions, therefore there are many different kinds of corpora.

By some dictionaries, it is suggested that corpora necessarily consists of a structured collection of text specifically compiled for linguistic analysis. These collections are large, and their attempt is to be representative of a language as a whole. He also adds that it is not necessarily so. Even if today’s norm, historically it was not even a norm, the necessity of electronic storage of corpora. From Kennedy’s point of view, electronic corpora can be composed of whole texts or collections of whole texts. There is also a possibility to create a corpus out of text samples taken from whole text or even citations might be used to establish a corpus (Kenndedy 1998,3).

Čermák does not oppose the idea of Kennedy when he confirms that the corpora nowadays are mostly in an electronic form and mentions the fact that most of the texts are not only created by computers but also analyzed using them (Čermák 2017, 13). He also adds a fact, that the biggest interest of users is concentrated in monolingual synchronous written, alternatively spoken corpora. Even though, there are far more types of corpora that might be used for various purposes according to the objective set by a particular user. According to the number of languages included, there are monolingual, multilingual or parallel corpora. From a perspective of a topic, general or specialized corpora are distinguished. From the perspective of the modus or also a form of a corpora, written or spoken corpora are defined. Synchronic and diachronic corpora are representing a time perspective. From the perspective of age there are synchronic and archive corpora. According to the purpose, various corpora could be used (ad hoc) (Čermák 2017, 74). The opinion on corpora has evolved a lot from the traditionalist to modern application.

Even if having been rejected by many linguists, nowadays, corpora serves not only for descriptive and theoretical studies of a language (Meyer 2004, 1). "Even though there are numerous functional theories of language, all have a similar objective: to demonstrate how speakers and writers use language to achieve various communicative goals. Because functionalists are interested in language as a communicative tool, they approach the study of language from a markedly different perspective than the generative grammarians. As "formalists", generative grammarians are primarily interested in describing the form of linguistic constructions and using these descriptions to make more general claims about Universal Grammar. For instance, in describing the relationship between *I made mistakes,* a sentence in the active voice, and its passive equivalent, *Mistakes were made by me,* a generative grammarian would be interested not just in the structural changes in word order between actives and passives in English but in making more general claims about the movement of constituents in natural language. Consequently, the movement of noun phrases in English actives and passives is part of a more general process termed "NP [noun phrase] - movement" (Haegeman 1991:270-3). A functionalist, on the other hand, would be more interested in the communicative potential of actives and passives in English. And to study this potential, the functionalist would investigate the linguistic and social context favoring or disfavoring the use of, say, a passive rather than an active construction" (Meyer 2022, 5-6).

When conducting a corpus based research, the necessity of choosing a suitable corpus is indisputable. First it is needed to be aware of what the corpus is about, how it functions and what and the basic terms in the field of a corpus linguistic are. The phenomenon that is going to be examined has to be taken into account from a semantic and sometimes also morphological-syntactic perspective. Logically, any corpus will not fit the purpose of a particular corpus research to an extent of 100 percent. The main point is to select a corpus that can provide as much suitable material as possible for the theses. Without any doubt, the significance of corpus may differ according to the user and his affiliation with a particular linguistic group. Also, the use can vary a lot in relation to what phenomenon one wants to examine or not. The same corpora or its part may be used for different purposes, analysed from different perspectives and for diverse objectives. Therefore, a great deal of caution is needed when stating the aim of a corpus analysis.

5 Basic terms of corpus linguistics

When working with and analyzing a corpus a cautious acquaintance with basic corpus linguistic terms is needed to name the processes, data and results in the right way. There are five basic terms needed to know well to conduct a corpus analysis correctly: *token, type, lemma, concordance, collocation and colligation.*

Čermák mentions that some of the corpus linguistics units are slightly abstract, and their abstractness rises from the most concrete to the most abstract one: *token - type - lemma.*

In the field of a corpus linguistic, every single textual form and also the smallest unit is called *token*. In other words, the token can be called *the occurrence*. The repetitiveness connected with a token can be observed. The *token* is from the point of view of a wider interpretation also named as a *position*. Every generalized token without repetition is called type. It is a unique unit located in a text (Čermák 2017, 47). If there are more tokens of the same type in a particular text, they will be counted as one type. Type stands for every unique token form in a text.

Another important term is lemma. Čermák claims that every abstract and from outside refilled representative form is in accordance with traditional linguistics is called lemma. The lemma can also be repeated in a text (Čermák 2017, 47). For the lemma *go* there can be tokens: go, went, gone, has gone, going, goes. A term lemma can also be described as a superior term for more types that are derived from the same word. To express the principles by numbers, the words *go, went, has gone and going* are one lemma. There are four types and also four tokens. In search of words that occur along with a particular word, it is talked about *concordance*, because different words are to be found together with a chosen word. This means the possibility to analyse a frequency of a particular word in connection with another. Concordance serves as a basis for determining the incidence of a particular combination of words in the corpus. It is one of the most important functions in electronic corpora tools.

Čermák's definition of concordance is slightly different from the one introduced above. He defines a concordance such a demonstration or a presentation of (all the) occurrences of a demanded term, word, form in a corpus ordered according to some key and followed with diversely long context and usually also statistical frequency (Čermák 2017, 274).

To be able to conduct corpus research, the knowledge of how the collocation works is essential to analyse the data in the right way. "*Collocation* and *colligation* are two closely related concepts associated with the distributional properties of linguistic items in actual language use. Specifically, collocation and colligation refer to the likelihood of co-occurrence of (two or more) lexical items and grammatical categories, respectively. Both terms have been attributed to J. R. Firth (1957: 194–195; 1968: 181–183; see Östman and Simon-Vandenbergen 2005 and Shore 2010 for a summary of Firth's work). Since the terms were introduced, collocation in particular has become a fundamental concept in usage-based studies in many linguistic fields, most notably lexical syntax and semantics. Typically, collocations and colligations are studied in large electronic corpora, which allows for statistical analyses of the co-occurrence patterns of linguistic items" (Lehecka 2015).

6 Types of corpora, their size, balance and representativeness

As usual by other spheres, also in case corpora there are many different types. Between these, a specialization in a particular area is to be seen. The corpora can differ according to the type of texts it includes not only from a formal point of view (spoken, written…) but also from the perspective of content, which means that a particular corpora can contain and therefore be specialized on a particular area such as *economy, law, literature* or particular period of time such as *90s of the 20th century.* Corpora also differ according to their balance and representativeness. Both these criteria together with the size of a corpus are important for the objectivity of the analysis and will be explained in detail later. Even if nowadays in most cases the use of electronic corpora is counted to be a beginning of a corpus-based research.

Kennedy opposes this statement and claims: "Corpus based-research is often assumed to have begun in the early 1960s with the availability of electronic, machine-readable corpora. However, before then there was a considerable tradition of corpus-based linguistic analysis of various kinds occurring in five main fields of scholarship: *biblical and literary studies, lexicography, dialect studies, language education studies, grammatical studies"* (Kennedy, 1998, 13). Kennedy adds that nowadays, linguists and other people who are interested in corpus linguistics may use many types of electronic corpora. These corpora vary according to the purpose for which they were compiled, their representativeness, organization and format. In the corpus linguistics literature, several different types of electronic corpora are sometimes distinguished (Kennedy 1998, 19).

As Čermák claims, there are many criteria of division of corpora that are closely connected to the needs of users (Čermák 2015, 74). An important thing is that some types of corpora can be both in written and electronic form. Written corpora can be either written on paper or stored as an electronic texts in a computer therefore the awareness of difference between the form of a corpus (*written, spoken, electronic*) and type in a sense of topic that the corpus in focused on and the period covered (diachrone, synchrone, specialized, general…) has to be taken into account.

Kennedy presents different types of corpora. Corpora that have been compiled for unspecified linguistic analysis are called *general corpora.* This type of corpus can be used for various purposes and asking many questions connected with vocabulary, discourse structure of a language or also grammar. Typically, it is used for comparative studies. General corpora are sometimes also called *core corpora, and* it is compiled to be balanced by including various genres and domains of use. They also include spoken and written, public and private language. An example of a general corpus is *SEU Corpus* (Kennedy 1998, 19). The so-called opposite of general corpora are specialized corpora. As the denomination suggests, these corpora are specialized on a particular field. They frequently serve to major commercial publishers as a source of word frequency and citation for the compilation of dictionaries (Kennedy 1998, 20).

Nowadays, many of the modern dictionaries are of this kind, such as *Macmillan Dictionary*. Kennedy adds that specialized corpora might also be focused on a particular problem or topic from many different areas. He claims specialized corpora compiled for studies of regional and sociolinguistic variations to be the major types of specialized corpora. According to him, *dialect corpora, regional corpora, non-standard corpora* and *learner's corpora belong* to this category (Kennedy 1998, 20). Another two types of corpora are *written* and *spoken corpora*. Čermák is in agreement with Kennedy's statement, when he confirms that the most common type of corpora is a written corpus (Čermák 2015, 74). He also confirms a fact, that consists in the contradiction between corpus linguistics and praxis, when in a real life the most used form of language is a spoken speech, whereas most of the corpora are written (Čermák 2015, 79). Both Kennedy and Černý mention the complicatedness of compiling spoken corpora on account of transcription, involving complex phonetic and prosodic features. These processes are described as time-consuming by both of the authors (Kennedy 1998, 20 and Čermák, 2015, 75).

The next corpora mentioned is a *sample text corpus*, that is designed to be a representative sample of the total population discourse (Kennedy, 1998, 21). “That population is not necessarily “the language as a whole”. Texts can be sampled from subpopulations, according to regions, genres, or groups of users (e.g. school-children, women, journalists or immigrants)” (Kennedy 1998, 21). Very closely related to the *sample text corpus* is a *full text corpus*. The full text corpus tends to be composed of complete texts or may consist of a specified size samples adopted from complete texts (Kennedy 1998, 21). As Kennedy mentions, there are more types of corpora than above, such as *parsed corpora*, that show the sentence structure and the function within a sentence. *Concordanced versions of corpora* are also used. This type of corpus is used very frequently, because all occurrences of any word are to be seen in a context. Increasingly reliable are the *tagging* and *parsing* of a corpus (Kennedy 1998, 21). As evident from the thesis, Černý and Kennedy are not in contradiction, when talking about particular pieces of information, but their division of types of corpora differ.

6.1 Representativeness

“Questions connected with *representativeness* and *balance* are complex and often intractable Leech (1991) has suggested that a corpus is “representative” in the sense that findings based on an analysis of it can be generalized to the language as a whole or a specified part of it” (Kennedy 1998, 62). Lech also adds that the structure of the early sample corpora such as Brown or LOB was projected in a cautious way to be representative of written American and written English in this sequence (Kennedy 1998, 62).

According to Čermák a representative form of data in a corpus is usually considered in unspecified and global research of a language that strives for a complex and balanced image of usage as optimal. By making an average of extracted results, it also indicates a typical usage (Černý 2015, 17). In contrast to this mainstream corpus linguistic approach trying to achieve objectivity of information and also base data Černý mentions Noam Chomsky and his followers, who practically look down the objectives of corpus linguistics and deduce principles from single isolated and sometimes also speculative and self-constructed examples of usage, that is total in contradiction with aims of the corpus linguistics (Černý 2015, 17).

6.2 Balance

Balance is important to the corpus to be representative and corresponding with the reality of language usage. The professor Čermák and Kennedy explain both representativeness and balance together as two closely related phenomena, whose significance is really important to be aware of (Čermák 2015, 21-22; Kennedy 1998, 62). Whereas Kennedy utilizes the word *balance,* Čermák talks about *proportions.* As Čermák claims, the degree to which a corpus faithfully and proportionally records a language, can in words of corpus linguistics be expressed as r*epresentation of a language reality* by using a corpus (Čermák 2015, 22).

Both Čermák and Kennedy are in agreement, when they state that different genres and varieties of a language have to be included in a ratio which is as much as possible corresponding to the language reality to create a balanced corpus that represents the actual state of a language (Čermák 2015; 22, Kennedy 1998, 62-63). Kennedy also adds that in the past, there used to be a tendency to favor written texts or compile corpora even entirely based on written texts (Kennedy 1998, 62).

6.3 Size

Another important parameter to consider when talking about and working with corpus is *size,* that is along with other indicators *representativeness* and *balance* crucial to take into consideration, when selecting a suitable corpus for research. As Kennedy mentions, *size* is not only about the quantity of texts and that these issues are not only associated with total number of *words (tokens)* and different *words (types)* , but with how many categories the corpus should contain, how many samples the corpus should contain in each category, and how many and words there should be in each sample (Kennedy 1998, 66). Therefore, the size of a corpus is not as simple matter as could be expected, but it appears to be far more complex than its name suggests.

Čermák formulates the idea of size and its importance in a different way, but the meaning is the same as Kennedy's definition above. He emphasizes that a ratio between particular genres and their proportional representation in a language reality is necessary to be a valid image of usage of a real language as a whole (Čermák 2015, 24).

Practical part

7 Methodology and corpus

First, before starting the analysis of prepositional phrases, a corpus had to be chosen. As stated before, a great deal of caution is needed when choosing a corpus. During the selection itself, many criteria had to be taken into account such as for example *size, balance, and representativeness*. In the case of conducting a corpus research focused year of release and also the period on examining prepositions in connection with modern technologies, the of time the corpus maps had to be dealt with really carefully. The period of time the corpus collects data about ought to have been as up-to-date as possible because the newer the corpus is the higher likelihood that more texts connected with modern technologies will be included.In the following paragraphs the corpora will be get through and particular criteria will be taken into consideration.

One possible choice was naturally the Brown Corpus, which is the most known corpus among people who focus on corpus linguistics. Concerning the fact that the most extensive development of modern information technologies and also the spread of their use in households on a daily basis has been for about the last 20 years, the selection of some of the classical and also older corpora than 20 years would have been neither beneficial nor wise. That is the reason why "traditional corpora" such as LOB Corpus (compiled in 1970's) or Brown Corpus (compiled 1961 (Cvrček, Čermák, and Kopřivová 2017)) would not have served the purpose of the research focused on prepositions connected with modern technologies a little or even not at all.

Both British National Corpus (1991-1994) (Cvrček, Čermák, and Kopřivová 2017) and British National Corpus 64, which maps the last year covered in British National Corpus, are more modern than LOB Corpus and Brown Corpus mentioned above, but still not modern enough to cover an area of modern technologies in a sufficient way.

The COHA - Corpus of Historical American English, which covers the period from 1810s to 2000s (Davies 2010), seemed to be better from a point of view of the age of texts included but the newest texts coming from 2000's still did not appear to be new enough to describe a phenomenon of using prepositions together with modern technologies. Another situation was the COCA - Corpus of Contemporary American English. This corpus maps a period of time from 1990 to 2017(Cvrček, Čermák, and Kopřivová 2017). Even if the COCA might have been used for the analysis of prepositions in connection, because it is up-to-date enough, there was no reason why not to choose a corpus that covers a period of time from 2000 and further, that is more connected with modern technologies. From the perspective of topics included, there were five main categories of genres: spoken, fiction, popular magazines, newspaper, and academic texts. Even though results needed for the research were to be found in the COCA, the best way ever is to choose a corpus that was mainly focused on the topic demanded to obtain as much data as possible and also have the possibility to get a sufficient amount of data to analyse to be able to introduce results representing the language reality as much as possible.

The best corpus for the purpose of the analysis of prepositions co occurring with particular modern technologies provided by the Sketch Engine seemed to be the English Web 2020 (enTenTen20). There were and still are many perspectives that support the idea that the English Web 2020 (enTenTen20) was the most suitable corpus for examining the co-occurrence of modern technologies with particular prepositions.

Firstly, the English Web 2020 (enTenTen20) has a size of 43,125,207,462 tokens ("EnTenTen: Corpus of the English Web" 2023), which is far larger than The Brown Corpus which includes about 100.000.000 words (Cvrček, Čermák, and Kopřivová 2017). That is why there is no doubt that the English Web 2020 (enTenTen20) is of sufficient size for a set analysis.

Secondly, as mentioned earlier, for the purposes of an analysis connected with modern information technologies, corpora older than 20 years would not be much useful, when taking into account that modern technologies such as computers, the web, etc., and their use on daily basis has started to spread about 20 years ago, the English Web 2020 (enTenTen20) which collected the data from 2020 definitely is ideal for such an analysis described above, because as newer a particular corpora and also the data included are as more spread the use of technologies on a daily basis is and as more occurrences of phrases from the area of modern technologies it includes.

Thirdly, due to the fact that the English Web 2020 (enTenTen20) was compiled of texts collected on the internet ("EnTenTen: Corpus of the English Web" 2023), it seems to be a great precondition for the occurrence of prepositions in connection with modern technologies, nevertheless, there are also topics included that appear to be widely connected with prepositions in connection with modern technologies such as games, science, technology and also home.

It cannot be claimed that a particular topic is not connected with modern technologies or terms related to them at all, nevertheless, some topics tend to be more related to the branch of modern technologies, whereas others are less connected with them.

Finally, the English Web 2020 (enTenTen20) is located in a modern corpus linguistics platform The Sketchengine ("EnTenTen: Corpus of the English Web" 2023), which is a tool or more exactly said to be a program created by professionals of the Masaryk University Brno. The modernness and well-done processing of the Sketch Engine allows the user of English Web 2020 (enTenTen20) to work rapidly, efficiently, comfortably, and transparently, which other and usually older platforms lack.

English Web en 2020 (enTenTen20) meets all the requirements listed above, therefore it was also chosen as the basis for this analysis.

8 Selected functions of Sketch Engine examined and used for the purposes of this thesis

When a corpus-based research is conducted, the knowledge of how the functions of a corpus programme work is crucial. Decent knowledge of functions that a particular corpus analysis programme offers leads to saving time on one hand and usually also to reduction of data that do not fit either from the perspective of their form or the meaning. Good comprehension of functions that a particular corpus analysis tool offers leads therefore to the rationalization of the whole process of collecting data and their analysis.

Before starting the activity of collecting data, cautious learning of the principles of how definite functions offered by electronic corpus programmes operate is indisputable. Every time seems to be better to get to know what the principles of a particular function are than to try them blindly without knowing anything about them. Testing blind or at random might be more adventurous than studying the tutorials and theory connected with a particular function but it can also be frustrating for the sake of not reaching the demanded aim stated, nevertheless some users prefer this manner of work regardless of how time-consuming this way of using such an approach is.

Even if knowing the principles of making oneself familiar with new functions or tools, the reality does not seem to be as simple as these principles are. Oftentimes, number of tutorials either written or in a video form does not introduce a particular function as a whole but they only focus on basics and lack the introduction of more advanced usage. Sometimes the tutorials even miss at all. That is the reason why a researcher has to rely either on a more practised colleague, a classmate or a professor, to be able to gain more experience about a particular function and its application in practice.

Before the analysis itself, the individual functions had to be tried and used in practical conditions, that is the reason why the work with particular functions provided by Sketchengine was not as easy and fast as it might have been expected previously. There was no other way than to try the functions one by one.

For the purpose of this analysis Sketchengine, which is a corpus linguistics programme created by linguists of the Masaryk University Brno, was chosen. The simplest function provided by this programme is called *word sketch*. It is a basic function everyone gets acquainted with at the very beginning of work with Sketch Engine. This particular function allows the users to input a word and then obtain collocations and word combinations with this particular word. Closely related to *word sketch* is function *word sketch difference,* which is basically the same function as the *word sketch.* The difference is that the *word sketch difference* can provide an input of two different words whose collocation and word combinations are displayed along with a graphical representation afterwards. The graphic has a form of a comparative overview. Both *word sketch* and *word sketch difference* are useful functions for the purposes of corpus analysis, but they only allow one to input one single word to a box, therefore complex phrases such as prepositional phrases *on your screen* or in *your screen* cannot be analysed via these tools.

The function providing what is needed for the analysis of prepositional phrases such as *in your screen* or *on your screen* is the one called *concordance*, which does not only allow the user to input one word but also makes it possible to input queries composed of more than one word. That is why *concordance* appears to be the only function that provides a search of phrases such as *in your screen* or on *your screen*. When using a *simple concordance,* the search term is simply filled in the box without doing anything else. The problem that *simple concordance* does not deal with is which part of speech the demanded query should contain or what meaning should the query have as such, therefore in case of a search query *in your screen,* when a sense of an adverbial of place is demanded, the word *screen* might also occur as a verb, more precisely past participle of a verb functioning as a adjective such as in these sentences: Grill or relax <u>*in your screened patio*</u>! or *If you choose "< .25" you should still have plenty of choice* <u>*in your screened*</u> list. For this reason, the analysis of queries provided by Sketch Engine is time demanding, because the more unfitting results there are the more time is spent by analyzing.

To avoid and reduce the number of results that do not have a demanded meaning seems to be more wise to use *advanced concordance* with tags. Tags could be described as specifiers because they specify which characteristics a particular part of a phrase has. Simply said, the *tags* can define criteria that a particular part of a phrase should have such as in this CQL code *[word="in"][tag="PPZ"] [word="screen"].* Instead of using a simple concordance without using the *CQL code*, which is a chain of *tags,* a better idea might be using *advanced concordance* which means CQL code including tags. The tag *word* makes a particular part of a phrase to have a firm form, which means that this element will not be able to change its form in any way. The reason why in the example above is that instead of obtaining results including only *in your screen,* by using the tag *PPZ*, which stands for possessive pronouns, the range of results is more varied because it does not only include the pronoun *your*.

9 Results

9.1 Classification of data provided by Sketch Engine

Before starting the analysis there is one necessary operation to do – to state which form or meaning the chain of words should have and according to which criteria decide whether a particular chain of words (phrase) does fit the meaning and form demanded or not. The demanded meaning of a phrase *on/in your screen* is an adverbial of place such as in the case of the examples below:

On your screen

1. *Although a smaller image may look fine* ***on your screen****, it will not print well; a printer needs a much larger sized image than a screen does to compose the image.*
2. *If you saw a pop up* ***on your screen****, this is most likely what you've experienced.*
3. *Tibetan Mastiff Screensaver is a nice screensaver that will show* ***on your screen*** *many images of Tibetan Mastiff of different breeds and sizes.*

In your screen

4. *if your interested in getting very clear & detailed, ingame, explosive flying debris-artifacts* ***in your screens*** *(they actually look like pieces of identifiable units & structs scatering in all diff directions... works for firing*
5. *Up to 200 words of text, 3 photos to fit on 800x600 pixel area (The amount you can see at ONE time* ***in your screen*** *) Includes listing in Stallion Service Directory or Service, Farm Directory etc.. <s>*
6. *(the particular email is not open at this point, you see your spambox or inbox* ***in your screen*** *). Hold down simultaneously Alt-Shift-F. You now have a new email with the checked email as the attachment.*

After inputting the search query in the a form: *in your screen*, desired result, after the first sight, were not obtained to a full extent. The degree of fitting results, those which have the meaning of an adverbial of place, was not clearly speaking enough to only present the number of fitting results and cease doing the further analysis, because many invalid results occurred and there was a necessity of finding, where there was any possibility to reduce such results. On the other hand, in the case of inputting the search query in the form: *on your screen*, the results seemed to speak quite clearly. No invalid results appeared to be there, nevertheless a further examination was done to find out whether some change in a ratio of valid and invalid results by both the noun phrases *in / on your screen* could arise.

As the previous paragraph outlines, mainly in case of *in your screen* there were many invalid results present. Because of many invalid results when a simple concordance was utilized, the description of these results from both syntactic and semantic perspective was done. The following step, when all the non fitting results were classified, consisted in establishing and describing groups to which the non fitting results were assigned.

Out of non-fitting results two main groups can be established, first of all, the group that has *a very different meaning* than demanded, second *word screen functioning as a descriptor*.

Group called *I. Very different semantics* includes those phrases that have a different form and also meaning to *on/in screen*. The central criterion is semantic meaning of a phrase or its parts.

6. Enjoy your big lot while sitting out back ***in your screened in porch.***

7. *Town gazebo? You might even be able to catch the sunset while out there eating your wonderful meal. Play Outdoor Chess* ***in Your Screened Gazebo*** *Staying home*

8. *When those hot summer days seem unbearable, go ahead and cool off* ***in your screened in pool*** *overlooking your basketball court. The possibilities are endless in this incredible home. It's not going to blow any holes in your screens and they will come cleaner.*

Group *II. Screen as a descriptor* refers to cases when the word *screen* acts as a descriptor within a phrase. This group can be divided into two subcategories:

II. A) screen as a descriptor, which structure is: preposition, determiner, descriptor, head.

II. B) screen as a descriptor plus another descriptor, which structure is: preposition, determiner, descriptor, head, determiner.

Appropriate examples of *screen* functioning as a descriptor and also belonging to subtype

II. A) are:

9. *Rainier Railroad and Logging Museum in Washington, the Great Smoky Mo... Please login to TwinTurbo.NET by entering* ***in your screen name*** *and password below.*

10. *Optically, it's also near-perfectly clear - meaning you won't notice any reduction i****n your screen brightness*** *or any dulling in image sharpness.*

11. *Your Expert Shield uses a 'dry fit' system. With our should be in the Internet format (your AOL screen name followed by "@aol.com"), such as: EMAIL. Note that any spaces* ***in your screen name*** *must be removed, so "My Screen Name" would become EMAIL.*

Examples representing subtype II. B) are:

12. A few pages later, ORIGIN FX (remember that product?) gets a nice write-up, too. 'If you want variety ***in your screen saver presentation***,' Charles Idol writes, 'ORIGIN FX does an excellent job.' Wing Academy gets a full-fledged

13. A few pages later, ORIGIN FX (remember that product?) gets a nice write-up, too. 'If you want variety ***in your screen saver presentation***,' Charles Idol writes, 'ORIGIN FX does an excellent job.' Wing Academy gets a full-fledged.

To get exact statisticadata in the form of whole numbers, the whole group of 100 samples (lines in the Sketch Engine) had to be classified from a point of view of validity and subsequently sorted into groups according to the type of invalidity. Only by this means the first exact data could be received. After the non-fitting results when the *simple concordance* was used was dealt with, the data below were obtained.

9.2 Raw results of the analysis – on / in your screen

On your screen vs in your screen and simple concordance

If a simple concordance in case *on your screen* was used, the results speak quite clearly. As evident from the table below *(Table 1),* there are 100 of 100 valid results *(Appendix 1-5 - on your screen, simple concordance),* those that fit the criterion to have a meaning of an adverbial place. No invalid results were to be found, therefore all the occurrences of the word *screen* within all the one hundred prepositional phrases had a meaning of an adverbial of place. The number of hits (occurrences) is 17,680. S

In the event of a phrase *in your screen* and the usage of simple concordance there are 734 total hits (occurrences). After the analysis of a sample of 100 occurrences (*Appendix 6-10 - in your screen, simple concordance)* there are 73 valid results, and 27 results are invalid, as presented in the table below *(Table 1)*.

Table 1 - IN your screen x ON your screen – raw results (simple concordance)

IN your screen x ON your screen – raw results (simple concordance)		
function	simple concordance	
search query	in your screen	on your screen
Search link	https://ske.li/tku	https://ske.li/tkv
I. VALID RESULTS	73	100
II. INVALID RESULTS	27	0
1. TYPE I. (very different semantics)	9	0
2. TYPE II. (*screen* as a descriptor)	18	0
II. A) (screen a descriptor)	14	0
II. B) (screen as a descriptor + another descriptor)	4	0
III. TOTAL NUMBER OF SAMPLES (I+II)	100	100

On your screen vs in your (or another possessive pronoun) screen and advanced concordance

To avoid much irrelevant data, the advanced concordance using CQL code was used. The CQL code [word="on"][tag="PPZ"][word=”screen"] provided a possibility to obtain data including various possessive pronouns, not only your. The code also provided via its tag *word* in a form of *[word=”screen"]* firm form of the word *screen* that is the reason why none different form of the word screen appeared.The *[tag="PPZ"]* caused the number of total hits was larger and the sample of 100 lines was multifarious in comparison to searches when a simple concordance was applied. After analyzing the data obtained when the advanced CQL code was utilized, the following outcomes were reached, as shown in the table below *(Table 2).*

In case of searching *on your screen* using the CQL code and setting PPZ as tag, which stands for possessive pronouns, as expected, the total range of hits (occurrences) of a phrase given is larger due to the expansion of other possessive pronouns than *your*, nevertheless only 1 invalid result out of 100 samples appears *(Appendix 11-15 - in possesive pronoun screen, advanced concordance*). The total number of hits (occurrences) is 27,428. By using a CQL with the tag PPZ the number of total hits (1,771) is more significant than if using only a simple concordance. After the extraction of irrelevant invalid results, there are 46 valid results, and 56 are invalid, as summarized in the table below (*Table 2).*

Table 2 - IN your screen x ON your screen – raw results (advanced concordance)

IN your screen x ON your screen – raw results (advanced concordance)		
function	advanced concordance (CQL)	
search query	in + possessive pronoun + screen[1]	on + possessive pronoun + screen
CQL code	[word="on"][tag="PPZ"] [word="screen"]	[word="in"][tag="PPZ"] [word="screen"]

Search link	https://ske.li/tkw	https://ske.li/tkx
I. VALID RESULTS	50	99
II. INVALID RESULTS	50	1
1. TYPE I. (very different semantics)	36	0
TYPE II. (*screen* as a descriptor)	14	1
II. A) (screen a descriptor)	12	1
II. B) (screen as a descriptor + another descriptor)	2	0
III. TOTAL NUMBER OF SAMPLES (I+II)	100	100

To be able to present the findings of the analysis well and make the outcomes in a form of statistical data comprehensible to the reader it is needed to state in advance in which form will the data be presented, how large the number of samples will be and also how the process of analyzing itself will be like.

For this thesis, a number of 100 samples of a given query was selected to make the potential recalculation to percent easier. To make different search queries comparable, there was a necessity to select the same number of samples, because different queries have a different total number of occurrences in a particular corpus.

After collecting and also analyzing a given number of samples of a particular prepositional phrase such as *in your screen, on your screen* etc., the data were filled to a table containing the results. First of all, the number of valid samples matching the criterion which have a sense of an adverbial of place were marked as valid results and filled to a corresponding cell named as *valid results*. The number of those that have not got a meaning of an adverbial of place were being filled into a cell marked as *invalid results*.

In contrast to the category of *valid results*, the category of *invalid results* is further divided into subcategories such a*s screen as a descriptor, screen and another descriptor, very different semantics.* The group of invalid results is therefore analytical in a sense of subcategories. The table also contains the name of a function used to make clear what the data refer to. Taking into consideration the usage of CQL codes, the form of a search query are to be seen to make the function used even better imaginable for the readers.

At the end of a table the total number of hits or simply said the total number of occurrences of a particular query is to be seen. Also, the common number of samples, which is 100, used for the analysis is being recorded there.

Even if most of the linguists including Madsen perceive determiners to be a group of many word classes or their subcategories of pronouns, numerals or adverbs such as possessives, ordinals, distributives, quantifiers, numerals, articles (Madsen 2020, 56) in the Sketch Engine only articles and demonstrative pronouns are included. According to this fact the amount of results using the tag “DT” is expected to be quite limited and specific in comparison to a situation when all the determiners would be included. There is a special separate tag “PPZ?” when possessive pronouns are looked for.

9.3 Adjusted results of the analysis – on / in your screen

After the analysis of a sample containing 100 lines of: *on / in your screen,* by means of simple and advanced (CQL code) concordance, the estimated number of invalid and valid hits was calculated on a basis of the total number of hits to allow the ratio of valid results to be calculated. The tables below not only serve as a summary of previous results, they also make the outcomes clearly arranged to provide the reader a more transparent evidence of which of the varieties is more used than the others and they even add new outcomes such as *the ratio of valid results.* All the data are displayed in the tables below*: (Table 3, Table 4), where* the column named as *Number of hits* represents total number of hits including invalid results. The *Invalidity rate* is based on the occurrence of invalid results detected in a sample of 100 lines. The *Estimated number of invalid hits* was calculated by multiplication the *Invalidity rate* by the *Total number of hits. The* Estimated number *of valid hits* was calculated by extraction of the *Estimated number of invalid hits* of *Number of hits* (Total number of hits). The *Ratio of valid results* is a proportion that was calculated by the following key: *Estimated number of valid hits* of "ON" *divided by Estimated number of valid hits* "IN". These principles might be also used for other prepositional phrases variants.

Table 3 - IN your screen x ON your screen – adjusted results (simple concordance)

IN your screen x ON your screen – adjusted results (simple concordance)					
	Function used	simple concordance			
	Number of hits	**Invalidity rate**	**Estimated number of invalid hits**	**Estimated number of valid hits**	**Ratio of valid results (in:on)**
IN	734	0.27	198	536	1 : 33
ON	17 680	0.00	0	17 680	

Table 4 - IN your screen x ON your screen – adjusted results (advanced concordance)

IN your screen x ON your screen – adjusted results (advanced concordance)					
	Function used	advanced concordance (CQL code)			
	Number of hits	**Invalidity rate**	**Estimated number of invalid hits**	**Estimated number of valid hits**	**Ratio of valid results (in:on)**
IN	1 771	50	885,5	885,5	1 : 31
ON	27 428	1	274	27 154	

The comparison of results of simple and advanced concordance shows a rising tendency of total number of hits (occurrences) is to be seen. In the case of *on your screen,* the growth of total hits including invalid results is from 17 680 when using a simple concordance to 27 428 when using a CQL code (advanced concordance), whereas in the case of *in your screen,* the growth is from 734 when using a simple concordance to 1771 when using a CQL code, which is more than a 100 percent difference between both values.

The overall statistic shows that the phrase *on your screen* is not only represented by a larger number of valid results, but the ratio of valid and invalid results seems to be far better (100/100 valid results using a simple concordance and 99 valid results using advanced concordance) than the phrase *in your screen*, where not only the number of valid results but also the ratio between valid and invalid results seems to be far worse than by *on your screen*. Therefore the variety which tends to be more common, as is *on your screen.*

9.4 Raw results of the analysis – in / on / at Wikipedia

In / on / at Wikipedia

Based on the experience with examination of the prepositional phrase on / in your screen, the same groups of invalid results were stated and also described. The demanded meaning of valid results also remained - the prepositional phrase had to have the meaning of an adverbial of place. The only change was, of course, the replacement of the prepositional phrase head *screen* by the word *Wikipedia*. Then the procedure was the same as by the analysis of *on / in your screen.*

The two main groups of invalid results are, as you can also see in the table below *(Table 5)*:

I. Very different semantics which includes words that contain those phrases that are neither related to the word *Wikipedia* functioning as an adverbial of place, nor have anything to do with the word Wikipedia at all.

The second main group - *II. Wikipedia as a descriptor,* illustrated by examples such as:

15. *Students write throughout the quarter, building up to a substantive intervention* ***<u>in Wikipedia's coverage</u>*** *of the internet industries and their cultural implications.*

16. *If you plan to make breaking changes to this template, move it, or nominate it for deletion, please notify Twinkle's users and maintainers* ***at Wikipedia talk****:Twinkle as a courtesy, as this template is used in the standard installation of Twinkle.*

,where on the other are included phrases , within whose the word *Wikipedia* functions as a descriptor. There are also two subcategories of the second group *II. Wikipedia as a descriptor:* A) Wikipedia as a descriptor, whose structure is: preposition, determiner, descriptor, head such as in these examples:

17. Indian-language Wikipedia projects are directly impacted by this global drive, be it the Women's History Month edit-a-thon where Wikipedia content largely related to women are improved every year or the Lilavati's Daughters project where biographies of Indian women scientists were created and enriched ***in Wikipedia projects****.*
18. The simplest and best way to look ***at Wikipedia trends.***

and B) Wikipedia as a descriptor plus another descriptor, whose structure is: preposition, determiner, descriptor, head, determiner such as in examples bellow:

19. *By its popular acronym it's known as POODLE, and you can find information about it on the web: Article about POODLE* ***on Wikipedia Google announcement*** *and many others (do a search) Why is this a problem on some servers?*
20. Land Rover ***at Wikipedia The Land Rover page*** *at Wikipedia offers an extensive look at every aspect of the Land Rover brand.*

Table 5 – IN x ON x AT Wikipedia

IN x ON x AT Wikipedia			
Function	simple concordance		
Search query	in Wikipedia	on Wikipedia	at Wikipedia
Search link	https://ske.li/u3e	https://ske.li/u3k	https://ske.li/u3l
I. VALID RESULTS	91	94	89
II.INVALID	9	6	11

RESULTS			
1. TYPE I. (very different semantics)	0	0	0
TYPE II. (*Wikipedia* as a descriptor)	9	6	11
II. A) (Wikipedia as a descriptor)	6	2	7
B. II. B) (Wikipedia as a descriptor + another descriptor)	3	4	4
III. TOTAL NUMBER OF SAMPLES (I+II)	100	100	100

As in case of in / on Wikipedia also in case of in / on / at Wikipedia not only simple concordance but also advanced concordance with CQL code was used to get more accurate data for the sake of the analysis. There are results displayed in the table below *(Table 6).* See also the appendices which contain lines provided by Sketch Engine *(Appendix 21-25 - in Wikipedia, simple concordance, Appendix 26-30 - on Wikipedia, simple concordance and Appendix 31-35 - at Wikipedia, simple concordance).*

Table 6 - IN x ON x AT Wikipedia

IN x ON x AT Wikipedia			
Function	advanced concordance (CQL)		
Search query	in Wikipedia	on Wikipedia	at Wikipedia
CQL code	[word="in"] [word="Wikipedia"]	[word="on"] [word="Wikipedia"]	[word="at"] [word="Wikipedia"]
Search link	https://ske.li/u3m	https://ske.li/u34	https://ske.li/vj4

	I. VALID RESULTS	95	96	93
	II. INVALID RESULTS	5	3	7
1. 2.	TYPE I. (very different semantics)	0	0	0
	TYPE II. (*Wikipedia* as a descriptor)	5	3	7
	II.A) (Wikipedia as a descriptor)	3	2	5
B. C.	II.B) (Wikipedia as a descriptor + another descriptor)	2	2	2
	III. TOTAL NUMBER OF SAMPLES (I+II)	100	100	100

9.5 Adjusted results of the analysis – in / on / at Wikipedia

As in case of *in / on your screen* also in case *in / on / at Wikipedia* the sample of 100 lines was examined *(Appendix 36-40 - in Wikipedia, advanced concordance, Appendix 41-45 - on Wikipedia, advanced concordance and 46-50 – at Wikipedia, advanced concordance),* by means of both simple and both advanced (CQL code) concordance. The estimated number of invalid and valid hits was calculated on a basis of the total number of hits to allow the ratio of valid results to be counted. The tables below not only serve as a summary of previous results, they also make the outcomes clearly arranged to provide the reader a more transparent evidence of which of the varieties is more frequently used than the others and they even add new outcomes such as *the ratio of valid results,* which is a value computed out of the ratio of valid results and the sample of 100 lines.

As mentioned above, the steps done to analyse the prepositional phrase *in / on / at Wikipedia* were nearly identical or even the same as by examining the prepositional phrase in / on your screen. For the first sight the results displayed in the table the numbers seem to be quite balanced. In case of examining *in / on / at Wikipedia* via the simple concordance the proportion of valid results expressed in per 91 valid results out of a sample of 100 for *in Wikipedia*, 94 for *on Wikipedia* and 89 for *at Wikipedia*.

When an advanced concordance was used, there was quite a shift of valid results to be seen , 95 valid results out of a sample of 100 for *in Wikipedia*, 96 for o*n Wikipedia,* 93 for *at Wikipedia*. There was no change in the order of valid results by every single variety. The numbers have risen, but the sequence of the varieties according to the biggest number of valid results is not at all. From the point of view of the total number of samples, either valid or invalid, the numbers speak for themselves. The number of hits (including both valid and invalid results) when the simple concordance was used, by *on Wikipedia* was 30,771, whereas in case of *in Wikipedia* it was only 12,826 and *at Wikipedia* even 7,288. According to both the percentage of valid results and both the total number of hits the most used variant seems to be *on Wikipedia*. In comparison to phrases on / in your screen there was a signicant change, because no samples of the group *I. Very different semantics* were to be found at all. The numbers from the text are presented in the tables below (*Table 7 and Table 8).*

IN x ON x AT Wikipedia – adjusted results (simple concordance)					
	Function used:	simple concordance			
	Number of hits	**Invalidity rate**	**Estimated number of invalid hits**	**Estimated number of valid hits**	**Ratio of valid results (in:on:at)**
IN	12 826	9	115	12 711	1.8 : 4.3 : 1
ON	30 771	6	185	30 586	
AT	7 228	1	80	7 148	

Table 7 - IN x ON x AT Wikipedia – adjusted results (simple concordance)

Table 8 - IN x ON x AT Wikipedia – adjusted results

IN x ON x AT Wikipedia – adjusted results					
	Function used	advanced concordance (CQL code)			
	Number of hits	**Invalidity rate**	**Estimated number of invalid hits**	**Estimated number of valid hits**	**Ratio of valid results (in:on:at)**
IN	10 442	5	52	10 390	1.7 : 4.2 :1
ON	26 026	4	104	25 922	
AT	6 206	7	43	6 163	

10 Summary

In the course of the analysis of prepositional phrases from the area of modern technologies, particular observations were done. **Firstly**, the corpus based analysis is not only about the knowledge of some of the basic principles of this scientific field but also about a decent knowledge of other linguistic disciplines, above all syntax, morphology, semantics, therefore a complex linguistic knowledge is needed to be able to use the corpus based approach. **Secondly**, the corpus linguistics combines both theoretical acquaintance with the principles and phenomena of this science and both the capacity of learning how to work with corpus linguistics programmes and use their tools in a right and efficient way and no of these two skills is separable from the other one. **Thirdly,** there are two discovered types of invalid results that were detected and also defined during the analysis of prepositional phrases connected with modern technologies. The first group *I. Very different semantics* includes those phrases that have a different form and also meaning to *on/in your screen*. The second one *II. Screen as a descriptor* refers to cases when the word *screen* acts as a descriptor within a phrase. This group can be further divided into two subcategories:

II. A) screen as a descriptor, which structure is: preposition, determiner, descriptor, head.

II. B) screen as a descriptor plus another descriptor, which structure is: preposition, determiner, descriptor, head, determiner. **Fourthly,** the research question was answered in the following wording: the variant *on your screen* is used more often than *in your screen. The* ratio of valid results when a simple concordance was used is 1 *in* : 33 *on*. When advanced concordance with CQL code was used, the ratio was: 1 *in* : 31 *on*.

The variant on Wikipedia is more frequently used than the variants *in Wikipedia* and *at Wikipedia.* The ratio of valid results when a *simple concordance* was used is: 1.8 *in* : 4.3 *on* : 1 *at*, in case of *advanced concordance with CQL code* the ratio was nearly the same as in the case when a simple concordance was used: 1.7 *in* : 4.2 *on* :1 *at.*

Finally, the principles used for this analysis, including the two types of invalid results, might be applyed on further examination of prepositional phrases from the area of modern technologies.

List of bibliography

1. Cvrček, Václav, František Čermák, and Marie Kopřivová. 2017. “*Hlavní světové korpusy”*. CzechEncy - Nový encyklopedický slovník češtiny. 2017. https://www.czechency.org/slovnik/HLAVN%C3%8D%20SV%C4%9ATOV%C3%89%20KORPUSY.
2. Čermák, František. 2017. *Korpus a korpusová lingvistika*. Praha: Univerzita Karlova, nakladatelství Karolinum.
3. Davies, Mark. (2010) *The Corpus of Historical American English (COHA)*. Available online at https://www.english-corpora.org/coha/.
4. Dušková, Libuše, and Libuše Dušková. 1994. *Mluvnice současné angličtiny na pozadí češtiny*. 2. vydání. Praha: Academia.
5. “EnTenTen: *Corpus of the English Web*”. 2023. https://www.sketchengine.eu/ententen-english-corpus/.
6. Kennedy, Graeme D. 1998. *An introduction to corpus linguistics*. Studies in language and linguistics. London: Longman.
7. Lehecka, Tomas. 2015. *“Handbook of Pragmatics online: Collocation and colligation”.* John Benjamins Publishing Company: Collocation and colligation. 27 November 2015. https://benjamins.com/online/hop/articles/col2.
8. Madsen, Richard Skultety. 2022. *Morphology: introduction to English morphology for university students*. Liberec: Technická univerzita v Liberci, Fakulta přírodovědně-humanitní a pedagogická, Katedra anglického jazyka.
9. Meyer, Charles F. 2002. *English corpus linguistics: an introduction*. Studies in English language. Cambridge: Cambridge University Press.
10. Quirk, Randolph, and Douglas Biber. 1999. *Longman grammar of spoken and written English*. Harlow: Longman.

11. Swan, Michael. 2016. *Practical English Usage: New International Student's Edition*. 4. ed. Oxford: Oxford University Press.

12. Tunaz, Mehmet, Emrah, Muyan, Muratoglu, 2016. *"A Corpus Based Study on the Preposition Error Types in Turkish Efl Learners Essays"*. Uhbab: Uluslararasi Hakemli Beşeri Ve Akademik Bilimler Dergisi: 16. https://files.eric.ed.gov/fulltext/ed573230.pdf.

Appendices