BRNO UNIVERSITY OF TECHNOLOGY

Faculty of Electrical Engineering
and Communication

BACHELOR'S THESIS

Brno, 2020 Pavel Volek

BRNO UNIVERSITY OF TECHNOLOGY

VYSOKÉ UČENÍ TECHNICKÉ V BRNĚ

FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

FAKULTA ELEKTROTECHNIKY
A KOMUNIKAČNÍCH TECHNOLOGIÍ

DEPARTMENT OF FOREIGN LANGUAGES

ÚSTAV JAZYKŮ

A CORPUS PERSPECTIVE ON ENGINEERING TERMINOLOGY IN POPULARIZATION

ODBORNÁ TECHNICKÁ TERMINOLOGIE V POPULARIZACI Z HLEDISKA KORPUSOVÉ LINGVISTIKY

BACHELOR'S THESIS
BAKALÁŘSKÁ PRÁCE

AUTHOR — Pavel Volek
AUTOR PRÁCE

SUPERVISOR — Mgr. Jaromír Haupt, Ph.D.
VEDOUCÍ PRÁCE

BRNO 2020

BRNO UNIVERSITY OF TECHNOLOGY
FACULTY OF ELECTRICAL ENGINEERING AND COMMUNICATION

Bachelor's Thesis

Bachelor's study field **English in Electrical Engineering and Informatics**

Department of Foreign Languages

Student: Pavel Volek ***ID:*** 203168

Year of study: 3 ***Academic year:*** 2019/20

TITLE OF THESIS:

A corpus perspective on engineering terminology in popularization

INSTRUCTION:

The aim of the thesis is to investigate specialized terminology and its popularized versions from the point of view of the cotexts in which they occur.

RECOMMENDED LITERATURE:

McEnery, T. & Hardie A. (2012): Corpus linguistics: method, theory and practice. New York: Cambridge University Press.

Krhutová, Milena (2009). Parameters of Professional Discourse/ English for Electrical Engineering. Brno: Tribun EU.

Date of project specification: 7.2.2020 ***Deadline for submission:*** 12.6.2020

Supervisor: Mgr. Jaromír Haupt, Ph.D.

doc. PhDr. Milena Krhutová, Ph.D.
Subject Council chairman

ANOTACE

Cílem této bakalářské práce je prozkoumat vliv popularizace na technickou terminologii aplikováním korpusového přístupu při analýze technických termínů v populárních a vědeckých kontextech. Je toho docíleno porovnáváním jednotlivých výskytů termínů mezi korpusy s hlavním zaměřením na vybraných 5 termínů, které jsou prozkoumány detailněji a okomentovány. Tato práce také obsahuje krátké představení konceptu korpusové lingvistiky.

KLÍČOVÁ SLOVA

Korpus, korpusová lingvistka, terminologie, termín, SketchEngine, popularizace

ABSTRACT

The aim of this bachelor’s thesis is to examine the impact of popularization on technical terminology by employing a corpus-based approach to analysis of technical terminology in popular and scientific contexts. This is achieved by comparing the term’s occurrences between the corpora with the main focus being on 5 chosen terms, which are examined in greater detail and commented on. This work also gives a brief introduction to the concept of corpus linguistics.

KEYWORDS

Corpus, corpus linguistics, terminology, term, SketchEngine, popularization

ROZŠÍŘENÝ ABSTRAKT

Dříve pouze vědci používaná technická terminologie se s neustálým technologickým vývojem pomalu dostává do slovníků neprofesionálů. Nové technologie přináší nové nástroje pro prosperitu a pohodlí lidstva, s čímž přichází i jejich potřeba slovního popisu. Pro pojmenování těchto fenoménů se často používá jejich příslušný technický termín - jenže ne vždy se tento termín bude používat stejným způsobem v laickém prostředí jako se používá v prostředí vědeckém. Brána mezi těmito dvěma prostředími se nazývá popularizace. Ta se v médiích projevuje nespočetně mnoha způsoby - od vědecko-populárních článků až po popisy produktů v e-shopech. V dnešní době je největším zdrojem informací pro širokou veřejnost Internet, což z něj dělá i nejefektivnější nástroj v popularizaci.

Korpusová lingvistika je relativně nový přínos do světa analýzy jazyků a s neustálým technologickým rozvojem je použití rozsáhlých sbírek textů - korpusů - při výzkumu jazyků na vzestupu. Možnost uchovávání, vyhledání a sledování nespočetně mnoha vlastností jakéhokoliv aspektu jazyka v potenciálně nekonečné sbírce textů je užitečná pro každého lingvistu, ať se zabývá evolucí jazyka v čase, či se snaží naleznout nové souvislosti v části jazyka, která není plně pochopena.

Cílem této bakalářské práce je aplikovat korpusově založený přístup v analýze technických termínů se zaměřením na počítačové sítě v populárních kontextech v kontrastu s jejich použitím v kontextech vědeckých za účelem zkoumání efektu popularizace. Za tímto účelem bylo porovnáno více než 15 slov, z nichž bylo vybráno a důkladněji okomentováno 5 termínů. První část této práce má účel čtenáři přiblížit problematiku korpusově založeného přístupu a samotný předmět výzkumu - technické termíny. Tato sekce začíná úvodem do korpusové lingvistiky a popisuje její historii i současnost. Na to navazuje popis hlavního nástroje korpusové lingvistiky - korpusu - a rozčlenění různých typů korpusů. Členění korpusů je důležité, protože ne každý typ korpusu je efektivní, či alespoň použitelný v určitých kontextech výzkumů. V rámci korpusové lingvistiky je též představen problém legality sdílení použitých korpusů, jelikož sbírané texty mohou být chráněny autorskými právy a jejich rozesílání ve formě korpusu je protizákonné. Další sekce první části této práce se zabývá popisem a definicí předmětu výzkumu - terminologii v popularizaci. Součástí tohoto popisu je i představení různých typů termínů a ukázka rozdělení termínů do skupin podle jejich specifičnosti, což je dále rozvedeno v pozdější části práce.

Druhá část této práce se zabývá důkladným popisem problémů a rozhodnutí, kterým autor čelí při přípravě nástrojů a předmětu výzkumu. Tato část začíná rozdělením termínů do tří skupin podle jejich specifičnosti pro účely této práce, což je doprovázeno vysvětlením rozdílů nové klasifikace od klasifikace z předešlé části práce. Je zdůrazněno, že hlavním objektem zájmu pro tuto práci jsou termíny patřící do druhé skupiny s názvem "general networking and other technical terms". Do této skupiny patří termíny jako například "subnet", "network", či "topology". Po klasifikaci termínů následují popisy postupů tvorby korpusů, které budou použity při analýze technických termínů. První korpus, pojmenován "scientific corpus", reprezentuje vědecké a akademické psaní. Je vytvořen z 90 různých vědeckých publikací na téma počítačové sítě. Je zdůrazněna sice dostatečná, ale neideální velikost tohoto korpusu, což může způsobit, že téma jednoho vzorku textu použitého v korpusu může mít viditelný vliv na frekvence používání určitých termínů. Druhý korpus reprezentuje psaný neprofesionální diskurz na téma počítačové sítě a byl vytvořen sbíráním populárně zaměřených textů na Internetu pomocí funkce "Find text on web" použitého korpusového softwaru SketchEngine. Tato funkce vyhledává různé kombinace předem zadaných slov pomocí internetového vyhledávače Bing a kompiluje stažené texty z vyhledaných webových stránek do použitelného korpusu. V kontextu tvorby tohoto korpusu jsou představeny "Keywords", které byly použity jako slova zadaná pro jeho tvorbu. Následující sekce popisuje postup použitý pro tvorbu všech ostatních korpusů - korpusů pro analýzu jednotlivých termínů. Druhá část práce je zakončena vysvětlením pojmu "relative frequency" a výsledky výpočtů "relative frequency" každého zkoumaného termínu pro porovnání ve všech relevantních korpusech.

Třetí část práce se zabývá praktickou analýzou vybraných termínů a jejich porovnání mezi kontextem vědeckým a populárním. Prvním bodem zájmu je porovnání "keywords" z korpusu vědeckých textů a korpusu webových vyhledávání. Účelem tohoto porovnání je, mimo poukázaní na rozdíly ve frekvenci používání jednotlivých termínů, hledání vhodných termínů na hloubější analýzu pomocí dalších nástrojů použitého softwaru. Zbytek třetí části obsahuje komentář k analýze pěti termínů vybraných na základě jejich četnosti a zajímavosti. Vybrané termíny jsou "node", "topology", "packet", "bandwidth" a "wireless". Každý termín je přesně definován a jednotlivě okomentován z hlediska jeho užití ve vědeckém kontextu a v popularizaci. Komentář je doplněn obrázky ze SketchEngine, na kterých jdou vidět frekvence kolokací či blízký kontext zkoumaných termínů. Každá analýza je poté zhodnocena. Některé termíny mají v různých kontextech výrazně rozdílné frekvence tvoření kolokací s jinými slovy, jako například "topology". Většina termínů se ale chová velmi podobně, některé téměř identicky. Tento fakt by se dal vysvětlit teorií, že v popularizaci na Internetu nedochází k přílišnému zjednodušování prezentovaných informací ale že veřejnost, která tyto informace vyhledává, je v oboru více a více znalá.

Na závěr můžeme říct, že korpus je velmi užitečný nástroj, pro který může najít využití každý lingvista. Spoléhat se ale pouze na korpus jako jediný empirický zdroj informací o jazyce není ideální. Zdánlivě nevýznamné rozhodnutí při tvorbě korpusu může mít dramatický efekt na konečný výsledek výzkumu. Z tohoto důvodu, společně s poměrně malými rozsahy použitých korpusů, jsou výsledky tohoto výzkumu pouhá spekulace a jejich potvrzení či vyvrácení by potřebovalo rozsáhlejší výzkum z pohledů více lingvistických disciplín.

VOLEK, Pavel. Odborná technická terminologie v popularizaci z hlediska korpusové lingvistiky. Brno, 2020. Dostupné také z: https://www.vutbr.cz/studenti/zav-prace/detail/127132. Bakalářská práce. Vysoké učení technické v Brně, Fakulta elektrotechniky a komunikačních technologií, Ústav jazyků. Vedoucí práce Jaromír Haupt.

PROHLÁŠENÍ

Prohlašuji, že svou bakalářskou práci na téma Odborná technická terminologie v popularizaci z hlediska korpusové lingvistiky jsem vypracoval samostatně pod vedením vedoucího bakalářské práce a s použitím odborné literatury a dalších informačních zdrojů, které jsou všechny citovány v práci a uvedeny v seznamu literatury na konci práce.

Jako autor uvedené bakalářské práce dále prohlašuji, že v souvislosti s vytvořením této bakalářské práce jsem neporušil autorská práva třetích osob, zejména jsem nezasáhl nedovoleným způsobem do cizích autorských práv osobnostních a/nebo majetkových a jsem si plně vědom následků porušení ustanovení § 11 a následujících zákona č. 121/2000 Sb., o právu autorském, o právech souvisejících s právem autorským a o změně některých zákonů (autorský zákon), ve znění pozdějších předpisů, včetně možných trestněprávních důsledků vyplývajících z ustanovení části druhé, hlavy VI. díl 4 Trestního zákoníku č. 40/2009 Sb.

V Brně dne ....

(podpis autora)

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF GRAPHS

INTRODUCTION

Corpus linguistics is a relatively new approach to text and language analysis and with the increasing technological possibilities, the use of corpora in these contexts is on a rise. The ability to locate and observe the behaviour of any aspect of language in a possibly endless body of text may aid a linguist in nearly any kind of language research, be it for example a change of language in time, or even forming of new rules in parts of language which have previously not been well understood.

During everyday life and mostly when browsing the Internet, everybody is surrounded by technical terminology, even if they do not perceive it as such. This is an unavoidable consequence of the impact constant technological advancement has on society as a whole. People boot up their personal computers, launch an Internet browser, use a search engine and log in to their social media accounts. These terms are, however, recognised by many non-professionals and hobbyists and are quite often used in non-academic, but sometimes educational popular media.

Most technical terminology has its roots in academic and scientific writing. Something new may have been invented and had to be scientifically documented. However, this documentation may be difficult for a layperson to comprehend, since scientific and academic writing is addressed to other scientists or academics - professionals in the field. In order to bring the invention to the attention of the general public, a different approach to writing must be employed - providing an interpretation of the science or technology that is intended for and is comprehensible by the public audience. This process of interpretation is commonly described as ‘popularization’.

Academic and popular science writing have distinctly different qualities and it is not difficult to differentiate them based on their vocabulary or sentence structure. The academic writing style is used in scientific articles, journals, and books its target audience is scientists and academics. It is therefore expected of the readers to have a certain degree of knowledge of the field or subject and to be familiar with the terminology used. It should be concise and precise, as the goal of academic writing is to convey scientific information clearly and without redundancy. In contrast, the target audience of popular science writing are usually the interested members of the general public, and so the popular science writing tends to avoid using the academic vocabulary and describes the terminology it uses at some point. It is not uncommon for popular science writing to have a narrative, either just to catch the attention of the reader with a short story at the beginning of an article, or to guide the reader throughout its entirety.

The aim of this thesis is to search for the impact popularization has on the use of technical terminology in writing by providing a brief analysis and comparison of terms and their use in scientific and academic writing in contrast with their popular use on the Internet with the employment of a corpus linguistic approach. Using several corpora, collocations, and concordance will be explored for multiple chosen terms, highlighting points of interest for further research.

The paper begins with a chapter on the theory of corpus linguistics and terminology, highlighting their relevance to this research. In the second chapter, the methodology and process of corpus creation is explained in detail and the brief analysis and comparison of the terms is the subject of the third chapter.

1. THEORETICAL INTRODUCTION

As the title of the paper implies, the research in this publication is approached from an angle of corpus linguistics. It is therefore important to introduce both the approach applied to the study and the subject of the study itself – terminology.

1.1 Corpus linguistics

Corpus linguistics is the study of language using large collections of naturally occurring language samples – transcribed spoken utterances or written text – usually stored in corpora. This approach to linguistic study gradually emerged from the need for observational data for research purposes similar to what can be seen in other scientific fields, such as biology or chemistry, since relying solely on the intuition of native speakers appeared rather unscientific. It can be used to describe language features and to confirm or refute hypotheses which, without observation, would prove to be difficult to test.

1.1.1 Methodology or a branch of linguistics?

While it includes linguistics in the name, it is mostly accepted that corpus linguistics shares more features with a methodology rather than an area of linguistic enquiry such as semantics, pragmatics, or sociolinguistics. However, a number of researchers are still in conflict whether it could also be considered a theory in itself. A simple resolution to this conflict would be to consider corpus linguistics to be both, as Kuebler and Zinsmeister (2015:14) comment:

> *"The answer to the question whether corpus linguistics is a theory or a tool is simply that it can be both. It depends on how corpus linguistics is applied..."*

1.1.2 Past to present

Corpus-like language study has a substantial history, even though the term 'corpus linguistics' first appeared in the early 1980s (Leech, 1992). While many linguists from before the 1950s used manually written bodies of text in forms of books, they were merely considered to be collections of written or transcribed text – not corpora by our modern standards, as they were not 'representative' (McEnery, Xiao and Tono, 2006). However, the methodology they employed was similar to how we use corpora in the modern day.

During the 1950s, the corpus methodology suffered a lot of criticism, which caused the approach to be almost entirely abandoned. The biggest figure of this wave was Chomsky, who claimed that it is not possible for corpora to avoid being 'skewed'. His criticism, certainly being valid at the time it was made, lead to further steady development of the corpus methodology in the coming decades and now, with the more recent technological progress, many of Chomsky's criticisms can be dismissed, as discussed in detail by McEnery and Wilson (2001:5-13).

Technological advancement after 1980 allowed for the creation and exploitation of increasingly larger corpora, which in turn caused the interest in corpus linguistics to increase dramatically. Nowadays are corpora widely used by linguists of all branches alongside with other methods and the employment of corpus linguistics brought improvements to many of the different linguistic disciplines.

1.2 Corpus

The entirety of corpus linguistics revolves around the use and analysis of a corpus. A corpus is a collection of naturally occurring – real life – language. They are generally assembled with a particular purpose in mind and aim to be representative of some type of text or language (Leech, 1992). The importance of representativeness and purpose is stressed by many corpus linguists as these two aspects are what differentiates a corpus from just a random collection of texts or an archive – especially considering that many of the criticisms in the past were aimed at the lack of representativeness.

In modern linguistics, a corpus is considered to be a collection of sampled texts in machine-readable form, which may be annotated with linguistic information, such as part of speech tags (McEnery, Xiao and Tono, 2006).

1.2.1 Types of corpora

There are many approaches to the creation of corpora, most of which result in the corpora being viable for use only under specific conditions. A corpus can be classified into various categories by the source of the content, its relation to other corpora or even metadata, and the same corpus can fall into more than just a single category, if it fulfils the conditions.

One of the first major distinctions between corpora is whether they are **general** or **specialized**. General corpora aim for an overall representation of a language. As an example, the British National Corpus intends to represent the entirety of British English. Hence, they are usually large and contain all ranges of styles and forms of language, formal and informal. Specialized corpora tend to focus only on a specific part of language, for example a genre. This is very useful when studying aspects of a language, which may appear only under specific conditions.

Another commonly specified distinction is between **monitor corpora** and **sample corpora**. The former relies on its ever-increasing size, assuming that as the corpus grows, the data becomes more reliable and balanced. Of an important note is the concept of 'Web as Corpus' (Kilgarriff and Grefenstette, 2003), as it is an example of an always growing corpus which, however, has its own problems, as described by McEnery and Hardie (2012). Monitor corpora tend to also be general corpora. Sample corpora, on the other hand, try to represent a type of language over a specific time frame. They aim to achieve balance and representativeness using specific characteristics, which define what kind of text is collected and used, and also keep in mind how often certain types of text naturally appear in the sampled type of language. If one were to create a corpus of the journalistic style from newspapers in the 1990s, the collected text should not be 80% interviews, 15% articles and 5% advertisements if we know that interviews represent for example only around 4% of all journalistic texts from newspapers in the specified timeframe.

Corpora can be further classified as **annotated** or **unannotated**, depending on whether or not there are linguistic analyses encoded into the corpus data. This encoding can for example show parts of speech. This annotation may be either attached to the words in the text itself or, using a computer program, stored separately from the text. Annotation directly in the text is more common as, if desired, the removal of such systematically created tags using a computer is trivial. Annotation is currently an important part of corpus usage, especially when using computer software, as it streamlines the process of searching through a corpus.

Next is the difference between **monolingual, bilingual** and **multilingual** corpora. Most corpora are monolingual, as in they are limited to only one language, for example English. Bilingual and multilingual corpora are corpora representing two and three or more different languages respectively, however, when talking in a broader sense, bilingual corpora are often incorporated into the umbrella term 'multilingual'. As multilingual corpora are relatively new, there is still confusion in terminology used for the subcategories of multilingual corpora, mostly around the term 'parallel'. This paper will follow the terminology used by Baker (1993, 1995) and McEnery and Wilson (2001). A **parallel corpus** is a corpus with source texts in one language and their translation in one or more other languages. A **comparable corpus** includes a pair or a group of monolingual corpora designed using the same sampling characteristics, as mentioned in the section about sample corpora.

When working with automatized corpus tools, especially when using keyword extraction tools, one may encounter **focus** and **reference** corpora. A focus corpus in the context of keyword extraction is the corpus from which the keywords are being extracted. It usually is the studied or analyzed corpus. A reference corpus is a more general corpus to which the focus corpus is compared in order to identify keywords. It is important to note that outside of the context of keyword extraction, and automatized corpus tools in general, the term reference corpus may also hold the meaning of a corpus that is designed to provide comprehensive information about a language, so it may be used as a basis of reference for other language materials (Sinclair, 1996).

For the purpose of this study, most corpora used will fall into the category of specialized corpora, specifically written academic and scientific corpora, and popular-scientific corpora. Furthemore, they will mostly be sample corpora, monolingual, and automatically annotated.

1.2.2 Legality

One of the most crucial issues one must face when creating a corpus is whether or not he has the legal right to gather and distribute the data intended to be included in the created corpus. The expansion of the web has streamlined the gathering of large quantities of text to create a corpus, but the underlying problem of copyright laws did not disappear. The redistribution of material under copyright without permission from the author is illegal. This is a major problem for a researcher creating a corpus for a study, as the corpus data should be made publicly available in order to ensure replicability of said study.

McEnery and Hardie (2012) propose several ways of addressing this issue. The first, and arguably the most reasonable, is contacting the owners of the gathered text and asking for permission to redistribute the text within a corpus under a license. This is, however, not feasible when a large number of different texts or web pages are to be sampled. Another way of circumventing this issue is only collecting data from websites that allow the redistribution of text, such as Wikipedia. However, this kind of restriction may still have an effect on the representativeness and balance of the final corpus. Another way to approach this is to collect data without seeking permission and not distributing the entire resulting corpus. However, it is still possible to make it available to other researchers through certain online tools that do not allow copyright to be breached. Such tools may show just small sections of the entire text in concordance, which should be considered 'fair use' under Czech copyright law and cannot be reconstructed back into the full text.

Due to the nature of the study in this paper, the approach of non-distribution had to be chosen in order to avoid the unlikely, but still potential legal issues.

1.3 Terminology popularization

Terminology, in the context of this paper, is a general word describing the set of specialized words or meanings, which usually relate to a specific field. These words are commonly referred to as ‘terms’ and they are generally words or multi-word expressions that are given specific meanings in specific contexts. Terminology is constantly evolving due to the need for professionals in a field to communicate with precision, though efforts are made to protect already established terms from the natural evolution of the meaning of words in language – the meaning of a term should not change, unless it needs to be adapted due to scientific progress in the field. This paper is concerned with one kind of evolution of terminology which comes as a side-effect of popularization of science - the process of making scientific topics more accessible to non-professionals.

1.3.1 What is a term

A term is a lexical unit whose purpose is the precise definition of a concept, a phenomenon, an entity (Krhutová. 2009). Such terms play an important role in the creation of coherence in communication between professionals in a field. These terms are commonly not part of general vocabulary and a person not acquainted with the professional field is unlikely to be able to fully understand utterances or text which include such terminology.

Main features of terms are the preciseness in meaning, unambiguousness, definability and stylistic and pragmatic unmarkedness (Bozděchová, 2009:29).

Terms can be either single-word or multi-word expressions, the latter of which consists of more than one word but act as a single lexical unit, most common examples being noun phrases. Multi-word terms are often called ‘collocations’ in corpus linguistics. In technical fields of study, it is not uncommon to encounter terms in the form of abbreviations as a unique case of shortening a multi-word expression into a single-word term.

Krhutová (2009, 108-109) classified technical terms from her studies on texts on electrical engineering into three groups:

1) General scientific terms
2) General technical terms
3) Branch-specific electrotechnical terms

A similar classification can be done for terms in any professional field and as such, one will be attempted in chapter 2 for this specific study.

2 METHODOLOGY

The methods of research a corpus linguist chooses to employ may have a large impact on the resulting data and improper procedure may bring the entire research to a wrong conclusion. It is therefore important to have a well thought out methodology behind every corpus-based research to achieve plausible results (McEnery and Wilson, 2001). This chapter's purpose is to document the process and explain the reasoning behind the decisions made during the gathering and compilation of research data.

The first subchapter is concerned with the initial classification of terms in the context of computer networking. This classification is important as it divides terms into multiple groups, from which only one is of particular interest for this study.

The second subchapter is concerned with describing the method, difficulties and choices regarding the creation of multiple corpora, all of which are later used in the analysis in the third chapter of this paper. The first corpus described is the corpus of scientific and academic text. In addition to its use in the last chapter, it acts as a point of reference for the creation of all other corpora. The second section consists of the description of the creation of the corpus of web searches. It intends to represent popular texts on the topic of computer networking on the Internet. The third section is concerned with the creation of 5 corpora each specialized around a single term.

The last subchapter is concerned with the normalization of word counts using relative frequency of terms and its calculation.

2.1 Term classification

Looking back at chapter 1.3.1, it was mentioned that a classification similar to the one devised by Krhutová (2009) could be made specifically for this study, and that by modifying the names of the categories to fit the context of this study subject:

1) Academic vocabulary (general scientific terms)
2) General networking and other technical terms
3) Specialist networking terms

The first group remains relatively unchanged and describes the most commonly used terms in scientific and academic writing, regardless of the topic of the writing. Examples of terms belonging to this group may be 'theorem' and 'hypothesis'. These terms appear in popular writing only sparingly.

The second group now includes terms more specific to the field at question in addition to general technical terms. This was done because of the change in the third group, which will be described in the next paragraph. The change pushed out terms which are often used by non-professionals who are relatively acquainted with the field into this group, making it very important for this study, as it is mostly terminology from this group which is being used in popularization.

The third group was changed as it appeared too broad for researching not the entirety of electrotechnics, but just a subsection - like computer networks. Without this change, nearly all terms coming from the keyword extraction would belong to the third group. The changed group should now better reflect who uses the terminology belonging to it, that being professionals highly skilled and specialized in the field of computer networking.

By observing the lists of keywords, it is simple to differentiate many keywords which should belong to the second and third groups by looking at the frequency at which they occur in the reference corpus. Terms with a low frequency of occurrence in the reference corpus are very likely to be used mostly in highly specific professional contexts, thus they can be considered Specialist networking terms. Some examples of specialist networking terms may be 'VNF', 'ExpressRoute', and 'EtherChannel'. Consequently, most terms with a high frequency of occurrence should belong to the second group, with examples such as 'node' and 'bandwidth'. One must keep in mind that just the occurrence of these terms in a general reference corpus is not enough to classify all terms. Some amount of linguistic introspection and also knowledge of the terminology is essential. The first group is fairly underrepresented in the lists of keywords for both analyzed corpora. That was expected for the corpus of web searches, as it is not comprised of scientific writing - it is unlikely for such expressions to be used in popular writing. However, the first keyword that can be classified as general scientific in the scientific corpus is 'denote' on the 35. position with 499 occurrences in the scientific corpus and 92,091 occurrences in the reference corpus. The second is 'Theorem' and the third is 'theorem' - note the capitalisation - on the 42. and 230. position respectively.

2.2 Creation of corpora used in this study

2.2.1 Choosing a scientific corpus

The first step is finding or creating a corpus of scientific writing suitable for our research purposes. The corpus must be monolingual, contain only scientific and academic written text and it must be specialized in a technical field of study. However, these specifications make searching for a freely available corpus on the internet a difficult task. Most specialized corpora are not accessible to the general public for a multitude of reasons, such as legality and confidentiality. Some English corpora available on the internet with free public access (such as British National Corpus) include subcorpora of academic writing, but there is no possibility of filtering out specific fields of study, therefore they cannot be used to conduct this research.

A different option available to us is the manual creation of a corpus specifically for this research, which would ensure that all the conditions specified above are met. For the purpose of creating such a corpus, 90 different scientific publications on the topic of computer networks have been gathered. It is necessary to rid the text of certain elements (such as tables) which could have an unwanted impact on the frequency of occurrence of some terms in relation to others, as having terms repeat multiple times in a non-coherent and unnatural text will disproportionately increase its number of occurrences. The collected body of text was then uploaded onto an online application called SketchEngine for automatic text processing, parsing, other tagging and final compilation using SketchEngine's 'Create corpus' function. For the automatic text processing, the default and recommended settings, as presented by SketchEngine, were used.

This has yielded a corpus fulfilling all of the aforementioned specifications with a total of 670,578 words in 31,102 sentences, which should be fully sufficient for the purposes of this research. However, it is important to note that a corpus of this size can still be considered small and may be prone to having resulting analyses affected by the subject matter of individual text samples. A larger corpus composed of additional differently themed samples would proportionately decrease the impact of a single sample - which would be desirable, but not possible while using SketchEngine due to a storage space limitation of 1 million words maximum.

2.2.2 Creation of a web search corpus

The first subject for comparison with the scientific text corpus will be a corpus created by gathering text from non-academic websites found using the ‘Find text on the web’ function during corpus creation in SketchEngine instead of manually inserting text as done previously. After writing at least 3 words or phrases, SketchEngine proceeds to use the bing.com search engine to search for different combinations of 3 of these words.

To achieve the most relevant search results, the decision to use a number of keywords appearing in the scientific text corpus in addition to generic networking terms was made. The keywords can be extracted in SketchEngine’s Keyword section with the scientific text corpus selected, where it can be seen how the application’s algorithm rates words as keywords by comparing the frequencies at which they occur in the studied corpus with the frequencies of the same words appearing in a different corpus - the reference corpus. SketchEngine’s default reference corpus is the ‘English Web 2013 (enTenTen13)’ corpus, however the ‘English Web 2015 (enTenTen15)’ corpus was used in the case of this keyword extraction due to its data being more recent. Other settings were kept as default with the exception of the ‘Focus on’ slider, which dictates the overall rarity of the extracted keywords in relation to general language or the reference corpus. The slider ranges from numerical values 0.001 to 1000000, with the default setting being 1. Smaller values prioritize more rare keywords, while larger values highlight more commonly used keywords.

Under default ‘focus on’ settings for keyword extraction in SketchEngine, most high-rated keywords appear to be abbreviations related to specific research topics from the scientific articles comprising the corpus and have very few occurrences outside of such topics. Such terms most likely belong to the third group previously specified in Chapter 2.1 and are unlikely to be used in popular writing, and therefore are of little interest for this study. By changing the value of the ‘focus on’ slider, it is possible to extract keywords which are more fitting to belong to the second group of terms specified in Chapter 2.1 - the general networking terms and other technical terms - which are more common in popular writing, making them an interesting subject for this research. The results of the keyword extraction employed will be further explored in chapter 3.1.

Keywords chosen from the extracted keywords list were ‘node’, ‘algorithm’, ’topology’, ‘packet’, ‘routing’ and ‘latency’. Other generic networking terms added to the search, which either appear on lower tiers of the keyword list or do not appear at all, were ‘subnet’, ‘bandwidth’, ‘gateway’ and ‘router’, making a total of 10 words used for 120 different combinations of 3 words employed in web searches, from which the application finds the top 20 results. SketchEngine then shows a list of all the combinations and their results where one can pick and choose which pages should be added as text to the final corpus. All scientific and academic texts were discarded, including Wikipedia, and a large number of websites had to be left out due to a storage limitation in SketchEngine.

The resulting corpus consists of 238,323 words, which, when compared with the previously created scientific corpus, should be a large enough sample to represent non-academic texts found on the Internet when searching for multiple relatively generic terms in the area of computer networks and networking.

2.2.3 Corpora for term analysis

These corpora were designed specifically for the analysis of a single term in non-academic text per corpus. There is not a great need for them to be large, as the term itself should have a very high occurrence even in a small sample. However, the samples were not deliberately shortened, as new contexts for the term's use may appear with more data.

The creation of these corpora was done using the 'Find text on the web' feature in SketchEngine, similarly to the Web search corpus in 2.2. However, this time only 4 words were used for the search, those being the term to be analyzed taken from the Keywords list of the corpus of scientific text and three other common networking terms, which remained the same for every other corpus created for the analysis of different terms. 'Network', 'routing' and 'system' were chosen as the three common words. Thus, when creating the analysis corpus for the term 'node', words 'node', 'network', 'routing' and 'system' were used. In the list of results, all academic texts and Wikipedia were discarded, as well as the results from the one possible 3-word combination which does not include the term subject to analysis.

The corpus for analysis of the term 'node' consists of 133,002 words, the corpus for 'topology' of 96,462 words, the corpus for 'packet' of 24,049 words, the corpus for 'bandwidth' of 37,918 words, and the corpus for 'wireless' of 54,023 words.

Table 1. Total sizes of all created corpora

Corpus name	Total size (in words)
Scientific c.	670 578
Web search c.	216 575
‘Node’ c.	133 002
‘Topology’ c.	96 462
‘Packet’ c.	24 049
‘Bandwidth’ c.	37 918
‘Wireless’ c.	54 023

2.3 Normalizing word counts

The comparison of two different corpora is a very common practice in corpus linguistics in general and it is an important part of this study as well. However, just a direct comparison of the numbers of occurrences for a term in differently sized corpora without taking this discrepancy into account could be misleading. In order to accurately compare corpora of different sizes, it is necessary to ‘normalize’ the frequencies of occurrence of the studied word for both corpora. In other words, to calculate the word’s relative frequencies of occurrence.

There are two types of frequencies considered in corpus linguistics. The first is absolute frequency, which is the raw number of a word’s occurrences in a corpus. It usually requires further specification, such as the total size of the corpus or another point of reference, to be useful in frequency comparison. Relative frequency is the absolute frequency in proportion to the total size of the corpus. By converting absolute frequencies in different corpora to relative frequencies, it is possible to reliably compare corpora of vastly different sizes. It is generally agreed to calculate relative frequency per 1,000,000 words for large corpora and per 10,000 words for small corpora. The general formula for the calculation of relative frequency is:

$$RF = \frac{AF}{S} \cdot N \qquad (2.3-1)$$

where RF is the relative frequency, AF is the absolute frequency (number of occurrences), S is the size of the corpus, and N is the number of words to which the frequency will be relative (1,000,000 or 10,000).

An example of a calculation of the relative frequency for the term ‘node’ in the scientific corpus would be:

$$RF = \frac{4478}{670578} \cdot 10000 \doteq 66.778 \qquad (2.3-2)$$

The results of all relative frequency calculations for each term in every corpus are shown in Tables 2., 3., and 4..

Table 2. Relative frequencies for the terms in the scientific corpus

Term	**Occurrences**	**Relative frequency (per 10 000 words)**
Node	4478	66.778
Topology	433	6.457
Packet	1902	28.364
Bandwidth	561	8.366
Wireless	306	4.563

Table 3. Relative frequencies for the terms in the corpus of web searches

Term	**Occurrences**	**Relative frequency (per 10 000 words)**
Node	661	30.521
Topology	130	6.002
Packet	434	20.039
Bandwidth	976	45.065
Wireless	64	2.955

Table 4. Relative frequencies for the terms in their respective analysis corpus

Term	**Corpus size**	**Occurrences**	**Relative frequency (per 10 000 words)**
Node	133 002	2 021	151.953
Topology	96 462	2 363	244.967
Packet	24 049	352	146.368
Bandwidth	37 918	420	110.765
Wireless	54 023	601	111.249

ons, but that may be my only recourse! </s><s> Re: WAN link	**bandwidth**	testing </s><s> Hi Justin, </s><s> For your scenario, I would sugge
ween two routers / between two LAN's. </s><s> Re: WAN link	**bandwidth**	testing </s><s> Hi Naidu, </s><s> Thanks very much for the respor
it using TTCP. </s><s> I'm not sure what each TTCP stream's	**bandwidth**	is, but it seems insufficient for my needs. </s><s> Do you know if th
iin, thanks for the assistance everyone! </s><s> Re: WAN link	**bandwidth**	testing </s><s> Hi Justin, </s><s> Since you mentioned, the links a
ak.sys.gtei.net 0.0% 1.1 1.1 1.0 1.1 0.1 </s><s> Re: WAN link	**bandwidth**	testing </s><s> Manish - Thank you for taking the time to reply, that
or the response, it's much appreciated. </s><s> Re: WAN link	**bandwidth**	testing </s><s> hi, </s><s> it seems strange to me that the ISP war
Vhile this being one of the easiest and most practical for WAN	**bandwidth**	testing I would like to emphasize that it can be CPU intensive. </s><
k. </s><s> you can use "http://speedtest.net/" to test the link's	**bandwidth**	. it is pretty accurately for where i am located. </s><s> you may alsc

Figure 1. Concordance lines for the term 'bandwidth' from the web search corpus

There is a large increase in relative frequency of the term ‘bandwidth’ in the corpus of web searches compared to the relative frequency of the term in the scientific corpus. To investigate this phenomenon, the term was searched through concordance in the web search corpus. It appears that a large number of occurrences of this term in said corpus stems from the amount of same repeating occurrences. An example can be seen in Figure 1., which shows a sample of the concordance, where the term is repeated in the subject of a forum post.

3 ANALYSIS

3.1 Corpus keywords

3.1.1 Comparison between corpora

The first subject of interest is the comparison of keywords extracted from the scientific corpus and the corpus of web searches.

	Word	Frequency? Focus	Reference	
1	VNF	346	626	•••
2	D2D	334	1,671	•••
3	eNB	224	738	•••
4	NFV	269	7,552	•••
5	node	4,478	474,024	•••
6	Sdn	462	32,918	•••
7	UE	310	17,706	•••
8	SPQ	149	272	•••
9	RLC	190	5,666	•••
10	PDCP	142	190	•••

Figure 2. Keyword extraction for the corpus of scientific text with default 'focus on' value

	Word	Frequency? Focus	Reference	
1	subnet	500	16,029	•••
2	ExpressRoute	248	273	•••
3	Vnet	133	444	•••
4	Azure	505	66,199	•••
5	bandwidth	976	154,858	•••
6	pred	106	984	•••
7	latency	408	56,259	•••
8	Bandwidth	132	7,084	•••
9	VPC	112	3,421	•••
10	ZyWALL	92	89	•••

Figure 3. Keyword extraction for the corpus of web searches with default 'focus on' value

Figures 2. and 3. show a list of 10 highest rated keywords from their respective corpus under the same conditions with the ‘focus on’ slider set to the default value of 1. The first column, titled ‘Word’, presents the keywords themselves, the ‘Focus’ column shows the number of occurrences the term exhibits in the analyzed corpus, and the ‘Reference’ column shows the number of occurrences the term exhibits in the reference corpus - the 15 billion word enTenTen15 corpus mentioned first in chapter 2.2. From the provided images in figures 2. and 3., while they only show a small sample of the entire list, one can deduce that the keyword lists are very different despite both of the corpora revolving around the subject of computer networks. It should be noted that SketchEngine’s keyword extraction is case sensitive, which can cause duplicate words to appear, as can be seen in Figure 3. with the keywords ‘bandwidth’ and ‘Bandwidth’. While ignoring it is not optimal for the preciseness of results, there is no option in SketchEngine which could help with fixing this issue.

It is important to note how the scientific corpus has mostly abbreviated terms in the highest positions of the list. That may be because many of these abbreviated terms were the subjects of research in the sampled academic and scientific texts, which would explain their high number of occurrences in contrast with their low number of occurrences in the reference corpus when compared to more general terms - ‘VNF’ appears only 626 times, while ‘node’ appears 474,024 times. One possible method to circumvent the issue of specific terms overshadowing the common terms could be by increasing the size of the focus corpus, resulting in more occurrences for the common terms, but that is not possible at the time being due to the 1 million word limit in SketchEngine.

Thankfully, by changing the ‘focus on’ value, it is possible for SketchEngine to filter out keywords which are undesirably specific, leaving only the more common keywords without the need for a larger focus corpus. Examples of keyword extraction for both focus corpora with the ‘focus on’ slider set to the value 10 can be seen in Figures 4. and 5. These keywords are viable candidates for deeper analysis due to them being used generously in both of the focus corpora and the reference corpus - which is a non-specific sample of written english text on the internet - implying a frequent use in popularization.

	Frequency?			
	Word	Focus	Reference	
1	node	4,478	474,024	•••
2	packet	1,935	338,940	•••
3	algorithm	2,614	521,651	•••
4	throughput	527	59,604	•••
5	Sdn	462	32,918	•••
6	VNF	346	626	•••
7	routing	469	74,248	•••
8	D2D	334	1,671	•••
9	denote	499	92,091	•••
10	topology	433	58,509	•••

Figure 4. Keyword extraction for the corpus of scientific text with the 'focus on' value 10

	Frequency?			
	Word	Focus	Reference	
1	bandwidth	976	154,858	•••
2	subnet	500	16,029	•••
3	Azure	505	66,199	•••
4	VPN	476	91,110	•••
5	latency	408	56,259	•••
6	Wan	323	47,228	•••
7	ExpressRoute	248	273	•••
8	gateway	434	180,790	•••
9	routing	290	74,248	•••
10	router	381	177,220	•••

Figure 5. Keyword extraction for the corpus of web searches with the 'focus on' value 10

3.2 Analysis of specific terms

This subchapter focuses on the dissection and commentary on a number of terms which appeared in the keyword extraction list of the scientific corpus. The terms were chosen specifically due to them being a part of the general networking term group, as the specialist network terms are very uncommon in popularization, as well as due to their high position in the keyword lists.

Using the ‘Word Sketch’ feature of SketchEngine, it is possible to examine the word’s collocates and other words often surrounding it. SketchEngine automatically summarises all of the collocates found and rates them based on a score. The score represents how strong the collocation is. A high score means that the examined collocate is often found with the analyzed word and at the same time the number of other words the collocate combines with is low. A low score means that the collocate often combines with many other words. This is a valuable tool to quickly gain insight into the behaviour of a word without having to manually examine the concordance generated by SketchEngine.

3.2.1 Node

According to the Oxford English Dictionary "a node is a point at which lines or pathways intersect or branch, a central or connecting point" ("Node, n."). In the context of computer networking this usually describes all devices connected to a network.

Figures 6. and 7. show a sample of the Word Sketch results displayed by SketchEngine for the scientific corpus and the corpus for analysis of 'node' respectively. Each column shows a different type of collocation and presents the results ordered from top to bottom by score. The left side of every row in each column shows the collocate together with an example of a collocation with the analyzed word, the centre number shows the frequency of occurrence for the collocation in the corpus and the rightmost number is the calculated score.

It should be noted that the term 'node' as a noun was found 4,478 times in the scientific corpus, while it appeared only 2,021 times in the corpus for analysis of 'node'. This means that the frequencies of occurrence for collocations shown in figures 6. and 7. should be considered with the difference in sample size in mind.

By observing the results of Word Sketch, it can be deduced that the term 'node' acts in a similar manner in both examined corpora and that the term is used in both scientific and academic contexts relatively equally. This is supported mainly by the amount of similar collocations found in the corpora as well as the similar frequencies at which most collocations occur relative to the number of times the term 'node' was found in each corpus. Most of the difference in the collocations can be attributed to the subject matter of texts sampled in the analyzed corpora.

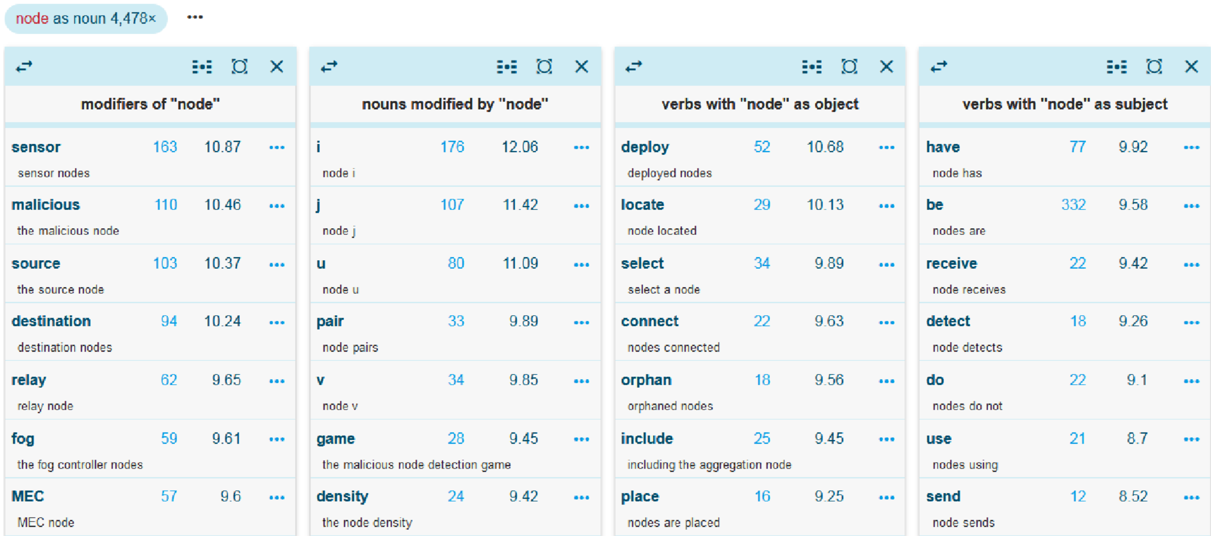

Figure 2. Word sketch results for analysis of the term 'node' in the scientific corpus

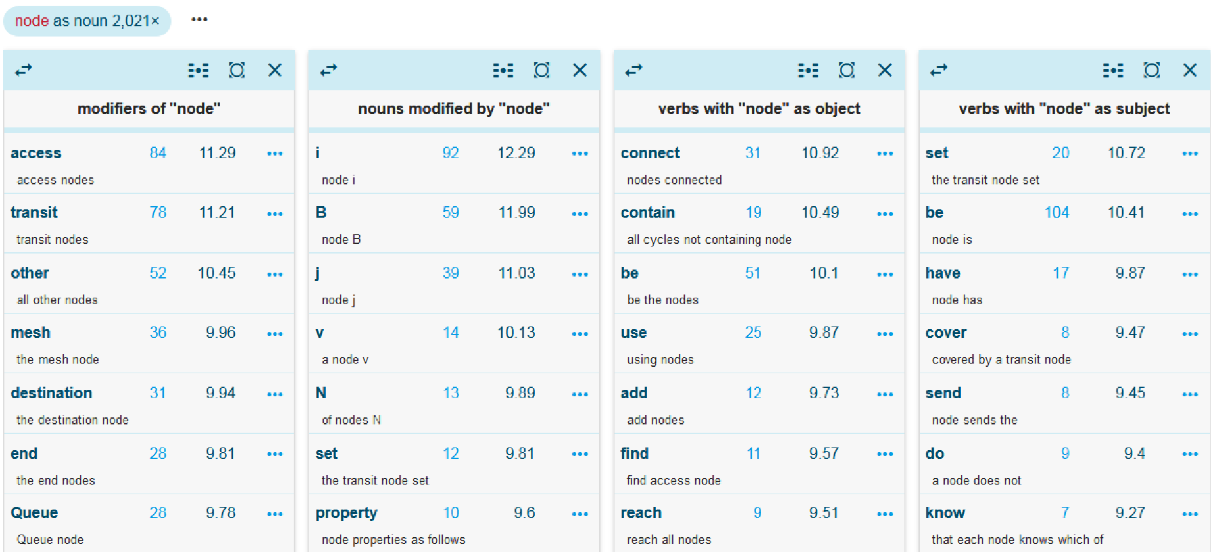

Figure 3. Word sketch results for analysis of the term 'node' in the corpus for analysis of 'node'

3.2.2 Topology

According to the Oxford English Dictionary topology is "the way in which constituent parts are interrelated or arranged" ("Topology, n."). In the context of computer networks this usually refers to the arrangement of a network. This is a very common topic of interest in non-professional circles, as it is one of the most basic cornerstones of network building.

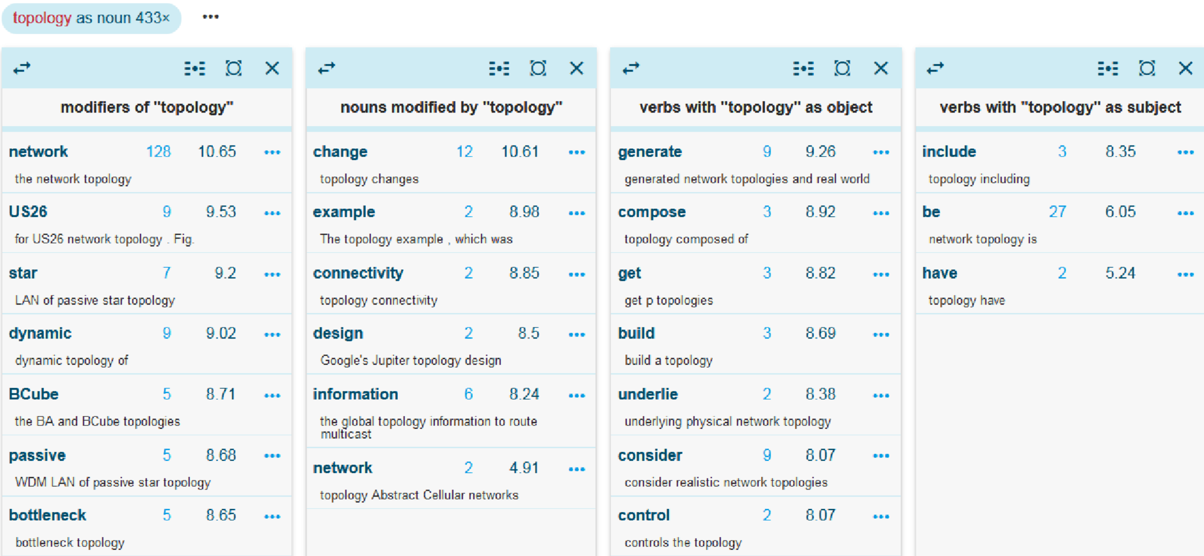

Figure 4. Word sketch results for analysis of the term 'topology' in the scientific corpus

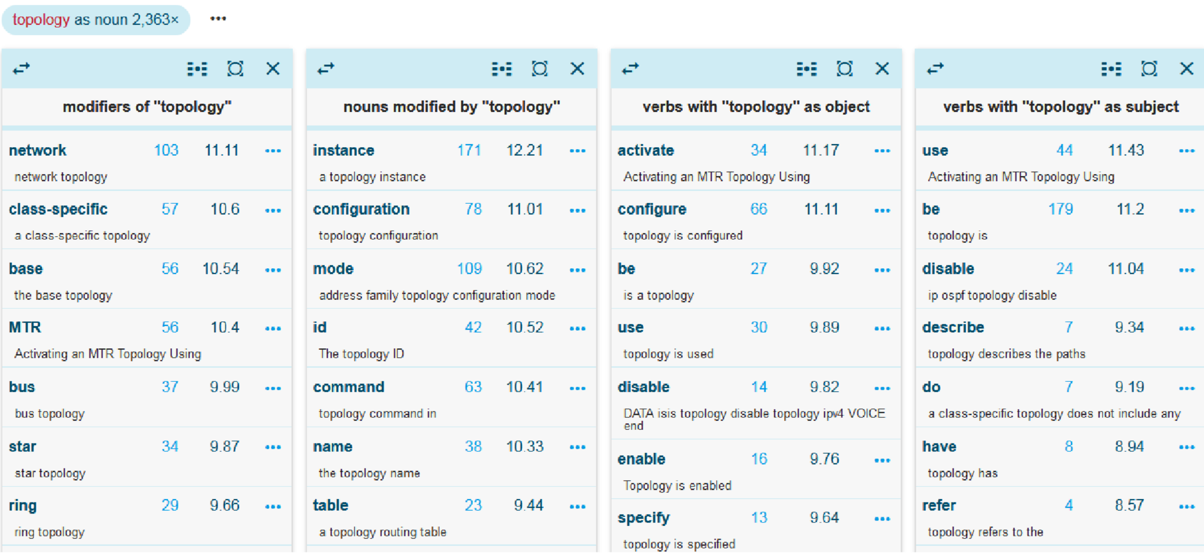

Figure 5. Word sketch results for analysis of the term 'topology' in the corpus for analysis of 'topology'

Figures 8. and 9. show the Word Sketch results from SketchEngine with the same layout and labels as described in chapter 3.2.1. Again, the difference in the relative frequency of the term occurring in the corpora should be noted. In the scientific corpus, the term 'topology' appears only 433 times, which is a considerably low amount considering the size of the corpus, while in the corpus for analysis of 'topology' the term appears 2,363 times.

It can be seen that in the scientific corpus the term tends to mainly form collocations with the word ‘network’, with only a few other occurrences of other collocations. This contrasts with the much larger pool of collocates surrounding the studied term and their higher frequency of occurrence in the corpus for analysis of ‘topology’.

From this data we may assume that the term ‘topology’ is a very common subject in popular writing, while in scientific writing it rarely appears as a subject of interest and mostly just complements a different subject of research.

3.2.3 Packet

In the context of computer networking, a packet is a unit of data transported across a network. A packet consists of control information, such as its source and destination, and user data, which upon reaching its destination may merge with other packets’ data into data blocks or files. In the current state of the Internet - and most computer networks in general - everything involves packets.

It is worth nothing that even though packets are a largely prominent element in computer networks, the corpus for analysis of the term ‘packet’ is much smaller compared to the other corpora for term analysis, which were created using the same procedure. This may be because the concept of packets is not complex and does not require a lengthy explanation to satisfy a curious layman. Another explanation could be that packets as a topic are not interesting enough to non-professionals to be widely talked about, resulting in scientific and academic writing overshadowing popular writing in search results.

The term ‘packet’ appears in the corpus of scientific text 1,902 times and in the corpus for analysis of the term ‘packet’ 352 times. The small number of occurrences in the latter corpus may be attributed to the small sample size of text in the corpus, as mentioned in the previous paragraph.

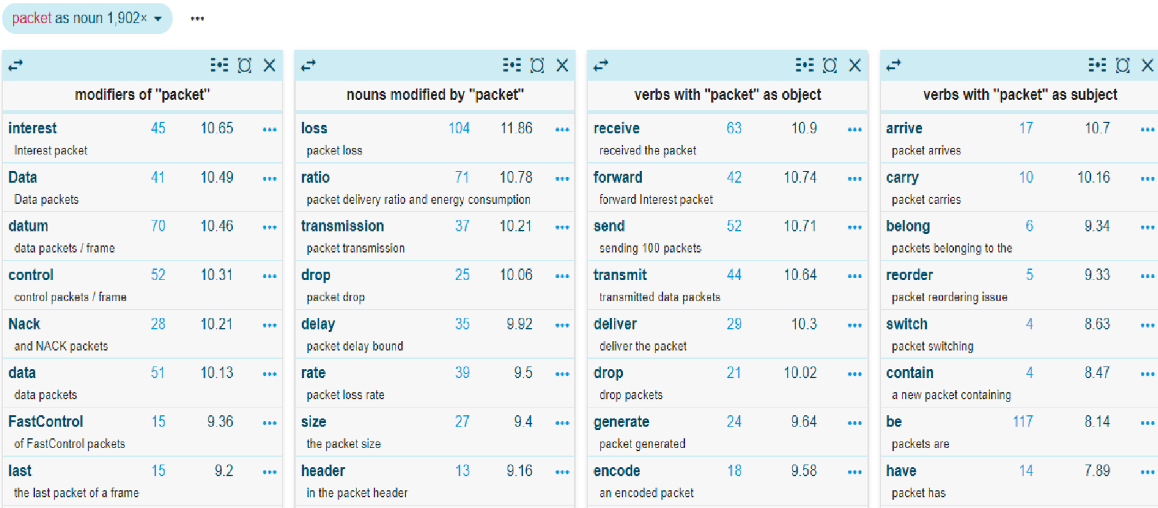

Figure 6. Word sketch results for analysis of the term 'packet' in the scientific corpus

Figure 7. Word sketch results for analysis of the term 'packet' in the corpus for analysis of 'packet'

Figures 10. and 11. show that the term 'packet' frequently acts as an object in a sentence in both scientific and popular writing. This implies the term serving a similar role in both contexts, with the most commonly appearing being 'send', 'forward', and 'receive'. However, there is a noticeable difference when it comes to the ratio of the term's most common modifier between the two studied corpora. It can be deduced that, in both contexts, the term 'packet' has a tendency to be modified by other nouns or adjectives, further specifying the type of packet the writer is referring to. In popular writing these modifiers appear to be more general, with examples such as 'data' and 'network', with an apparent outlier in the form of the collocate ‘IP packets’. While it may at first seem that the modifier ‘IP’ creates a narrow specification of a certain type of packet, it is not the case. When talking about packets in non-scientific contexts it is very uncommon for them to belong under other network protocols than TCP/IP - hence ‘IP packets’. This would make the modifier ‘IP’ a rather general or in some cases even redundant modifier, which can be corroborated by examining the concordance of ‘IP packets’ in the corpus for analysis of ‘packet’. An example of randomly chosen concordance lines can be seen in Figure 12.. In the scientific text corpus these modifiers tend to be both general and specific in relatively equal numbers, with examples of general modifiers again being ‘data’ and ‘network’, and examples of specific modifiers being ‘interest’, ‘NACK’, and ‘FastControl’.

to a network requires some type of routing instructions for network TCP/ IP	packets	when they leave the local host. </s><s> This is usually very straightforward bec
ttached printers, need to make decisions about where to route TCP/ IP data	packets	. </s><s> The routing table provides the configuration information required to m
thing to do with the steps that routers have to take when they forward an IP	packet	from one interface to another. </s><s> In this lesson, I will walk you through an
:1 and R2. </s><s> IP Routing Process </s><s> The actual forwarding of IP	packets	by routers is called IP routing. </s><s> This has nothing to do with the "learning
thing to do with the steps that routers have to take when they forward an IP	packet	from one interface to anothe </s><s> Let's look at this step-by-step, device-by-c
ın ARP request. </s><s> We now have an Ethernet frame that carries an IP	packet	with the following addresses: </s><s> The frame will be on its way to R1. </s><
interface so we will process it. </s><s> We de-encapsulate (extract) the IP	packet	out of the Ethernet frame which is then discarded: </s><s> The router will now l
et frame which is then discarded: </s><s> The router will now look at the IP	packet	, and the first thing it does is check if the header checksum is OK: </s><s> If the
the GigabitEthernet 0/2 interface and R2 as the destination. </s><s> The IP	packet	is then encapsulated in this new Ethernet frame. </s><s> And the frame will be
it to R2. </s><s> Like R1 it will first do this: </s><s> De-encapsulates the IP	packet	, discard the frame. </s><s> Check the IP header checksum. </s><s> Check th

Figure 8. Random concordance lines for the collocation 'IP packet' from the corpus for analysis of the term 'packet'

Similarly to ‘IP’ in the column of term modifiers, the noun ‘loss’ shows an overwhelming presence between nouns modified by the term ‘packet’. This may be due to packet loss being a common problem in networks where constant connection is necessary. Packet loss is the name for the event where packets are sent out, but some, or sometimes even all, do not manage to reach their destination. Such a phenomenon is likely to become a topic of interest for both laymen and professionals, resulting in a higher number of occurrences in both scientific and popular writing.

3.2.4 Bandwidth

In networking, bandwidth is the maximum data transfer rate across a given path within a network. It is a compound word consisting of the words ‘band’ and ‘width’. While bandwidth is one of the more important aspects of a network, essentially determining the amount of data that can pass through the network at a singular moment, it is often wrongly named ‘speed’ by both professionals and non-professionals alike. This may greatly affect the use of the term ‘bandwidth’ in popular contexts.

The term ‘bandwidth’ appears 561 times in the corpus of scientific text and 420 times in the corpus for analysis of the term ‘bandwidth’.

bandwidth as noun 561× ···

modifiers of "bandwidth"			
available	38	11.31	the available bandwidth
adaptive	8	9.89	adaptive bandwidth binning (ABB
fiber	5	9.4	fiber bandwidth
additional	7	9.13	some additional bandwidth
affected	3	8.62	the ratio of affected bandwidth routed through Gi
sufficient	3	8.62	sufficient bandwidth to
high	12	8.61	high bandwidth and
maximum	7	8.6	the maximum bandwidth that

nouns modified by "bandwidth"			
constraint	21	10.28	bandwidth constraints
utilization	19	10.26	bandwidth utilization
limitation	10	9.89	with bandwidth limitation
requirement	21	9.87	bandwidth requirement
allocation	17	9.8	bandwidth allocation
loss	13	9.75	upstream bandwidth loss
guarantee	8	9.6	bandwidth guarantees for
resource	19	9.44	compute and bandwidth resources

verbs with "bandwidth" as object			
allocate	16	10.65	bandwidth allocated to
consume	12	10.54	bandwidth consumed
waste	5	9.95	waste network bandwidth
buy	4	9.67	buying additional bandwidth
save	6	9.64	to save network bandwidth
reserve	3	9.23	reserve sufficient bandwidth
guarantee	4	9.01	with guaranteed bandwidth
divide	4	8.94	bandwidth is divided into

verbs with "bandwidth" as subject			
bin	8	12.11	adaptive bandwidth binning (ABB
do	2	7.17	bandwidth does
be	28	6.1	bandwidth is
have	3	5.82	bandwidth has

Figure 9. Word sketch results for analysis of the term ‘bandwidth’ in the scientific corpus

bandwidth as noun 420× ···

modifiers of "bandwidth"			
much	7	10.98	how much bandwidth can vary
more	8	10.86	more bandwidth
available	6	10.84	available bandwidth
network	16	10.56	network bandwidth
total	4	10.23	out of whatever total bandwidth we might be
low	4	10.04	It uses low bandwidth
high	4	10.04	a high bandwidth but
term	2	9.37	The term bandwidth

nouns modified by "bandwidth"			
usage	9	10.71	bandwidth usage
command	8	10.52	the bandwidth command
maximization	7	10.49	for bandwidth maximization
availability	6	10.28	bandwidth availability . The
value	7	10.18	bandwidth value
statement	5	10.02	the bandwidth statement
utilization	5	10	bandwidth utilization
consumption	5	9.99	limit their total bandwidth consumption to like

verbs with "bandwidth" as object			
maximize	6	10.68	to maximize bandwidth
use	12	10.57	using bandwidth
divide	4	10.15	then EIGRP will divide the bandwidth of the physical
reduce	4	9.97	reduce the bandwidth
limit	4	9.95	would like to limit the bandwidth used by a
configure	4	9.95	the bandwidth configured
expand	3	9.72	expand the bandwidth
specify	3	9.65	network connection that specified a bandwidth of 1Mbps means

verbs with "bandwidth" as subject			
throttle	17	12.68	Re : Bandwidth throttling via QoS
be	32	10.62	bandwidth is
mean	3	10.33	bandwidth means the bandwidth
make	2	9.75	most of the bandwidth making it useless for
need	2	9.61	bandwidth needs
have	3	9.27	bandwidth has

Figure 10. Word sketch results for analysis of the term ‘bandwidth’ in the corpus for analysis of ‘bandwidth’

According to Figures 13. and 14., there is little difference in the usage of the word between the different contexts. A minor difference is in the slight decrease in complexity of some of the term's collocates in popular contexts, such as 'additional' being replaced with 'more'. A closer look at randomly selected concordance lines for 'bandwidth' in both the corpus of scientific text and corpus for analysis of the term does not disprove this sentiment either, as can be seen in Figures 15. and 16..

v reclassification system will detect and attempt to mitigate instances of	**bandwidth**	allocation that are not weighted max-min fair, but there is no way in gene
on is designed to assess the ability of ABB to isolate unresponsive high	**bandwidth**	UDP flows. </s><s> Our prior work [32] showed that single queue schem
d conditions would provide improved performance. </s><s> The present	**bandwidth**	consumption and bin assignment algorithms also merit further study. </s
ributed computation and storage facilities in addition to connectivity and	**bandwidth**	[1]. </s><s> These characteristics of the 5G systems open the door for m
imal solution of the restoration problem, called RT-O. </s><s> Since the	**bandwidth**	of each demand is given, yd is a fixed parameter in RT-O instead of a va
</s><s> We define . In each small increment ε of vl, the ratio of affected	**bandwidth**	routed through Gi is computed as follows:(34)where(35) The detail of all
of admitted user requests can be maximized, subject to computing and	**bandwidth**	resource constraints. </s><s> In this paper, we study the fault-tolerant st
services :: For start-up service providers with both limited compute and	**bandwidth**	resources,we formulate the fault-tolerant VNF placement problem to max
ion is minimized, subject to compute resource capacity C(DCi), network	**bandwidth**	capacity B(e) for e ∈ E, and end-to-end delay constraints. </s><s> Let xj
period of time. </s><s> However, the evaluation value of link's available	**bandwidth**	to user's bandwidth requirement approaches 0 rather than being equal to

Figure 11. Randomly selected concordance lines for the term 'bandwidth' from the scientific corpus

vidth statement of 4096, so the network should be getting 4 Mbps of	**bandwidth**	. </s><s> Unfortunately, it just doesn't work that way. </s><s> Let's fi
><s> OSPF uses cost as its routing metric, which it calculates using	**bandwidth**	. </s><s> For example, OSPF takes 108 and divides it by the bandw
Here's an example: </s><s> This command has only one option–the	**bandwidth**	, in kilobits, of the interface. </s><s> There are always default bandw
lefault may be incorrect. </s><s> As you can see, setting the correct	**bandwidth**	on each interface is very important when it comes to routing protocol
dth is lower, make sure you configure the correct bandwidth with the	**bandwidth**	command. </s><s> Telling EIGRP to use less bandwidth is simple, th
t their total bandwidth consumption to like 2MB out of whatever total	**bandwidth**	we might be assigned. </s><s> I have read some resources on Clas
jateways can also free up a lot of bandwidth. </s><s> More Network	**Bandwidth**	Troubleshooting </s><s> Encountering bandwidth issues </s><s> Sr
on, you can see the low signal areas and deadzones that reduce the	**bandwidth**	available to your users. </s><s> Switching between the Network Per
op-of-the-line combination of network tools to help you manage your	**bandwidth**	. </s><s> However, SolarWinds provides an alternative pack of utiliti
nments. </s><s> The Server Manager module won't directly sort out	**bandwidth**	problems. </s><s> It identifies any server issues that may be slowing
er second (bps). </s><s> An ISP network connection that specified a	**bandwidth**	of 1Mbps means that in one second, maximum 100000 bits can be tr
e: </s><s> Therefore theoretically an internet connection of 512kbps	**bandwidth**	can download at a maximum speed of 62.5KBps. </s><s> If you do r

Figure 12. Randomly selected concordance lines for the term 'bandwidth' from the corpus for analysis of 'bandwidth'

An assumption can be made that the reason for 'bandwidth' staying essentially the same is the incorrect name 'speed' taking its place in everyday conversation, leaving 'bandwidth' to be used in the more educational and knowledgeable parts of popular writing which more closely imitate its usage in scientific and academic writing.

3.2.5 Wireless

The term 'wireless' is different from the previously examined terms, since it most commonly behaves as an adjective - modifying other nouns and creating multi-word terms. It consists of the free morpheme 'wire' and the derivational morpheme '-less'. As described in the Oxford English Dictionary, the word 'wireless' as an adjective holds the meaning "using radio, microwaves, etc. (as opposed to wires or cables) to transmit signals" ("Wireless, adj."). However, searching for the term 'wireless' as an adjective in our corpora using SketchEngine's automatic part-of-speech parser results in no occurrences found. The software recognises the term as a noun while also showing nearly exclusive use as a modifier in both the scientific corpus as well as the corpus for analysis of the term 'wireless', which can be seen in Figures 17. and 18.. This could imply that the term 'wireless' may be considered a noun adjunct. Such an implication would however contradict many dictionaries, which classify the term in these contexts as an adjective. There is no major difference in the modifier's usage between the two analyzed corpora. Although it is not important for this research to further explore whether it is or is not a noun adjunct and for this reason will keep to the definition in dictionaries, the difficulty the not ideal parsing presents reaches into the examination of another dictionary definition for 'wireless', with this instance being a noun.

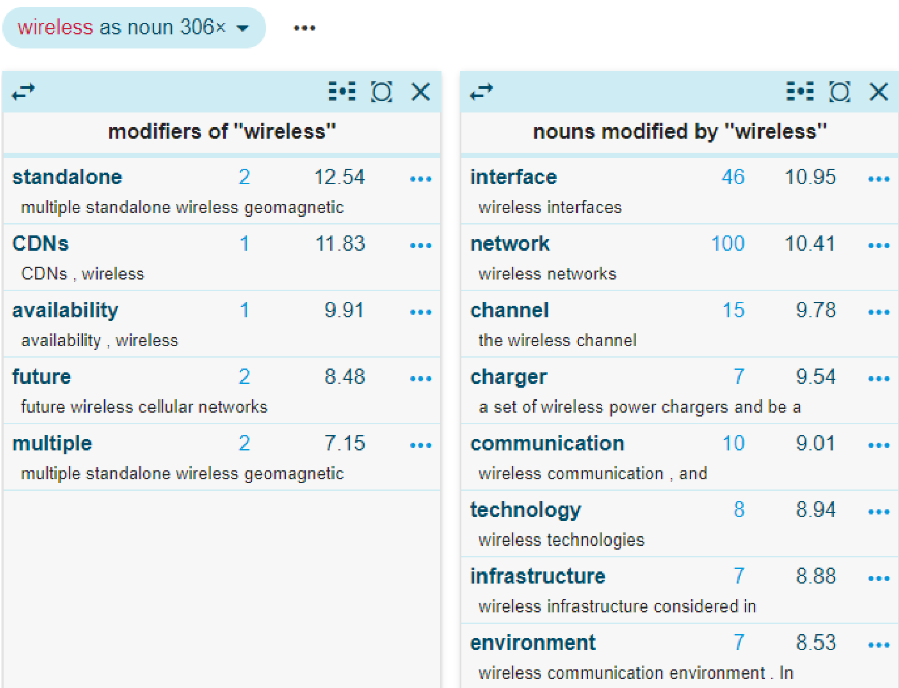

Figure 13. Word sketch results for analysis of the term 'wireless' in the scientific corpus

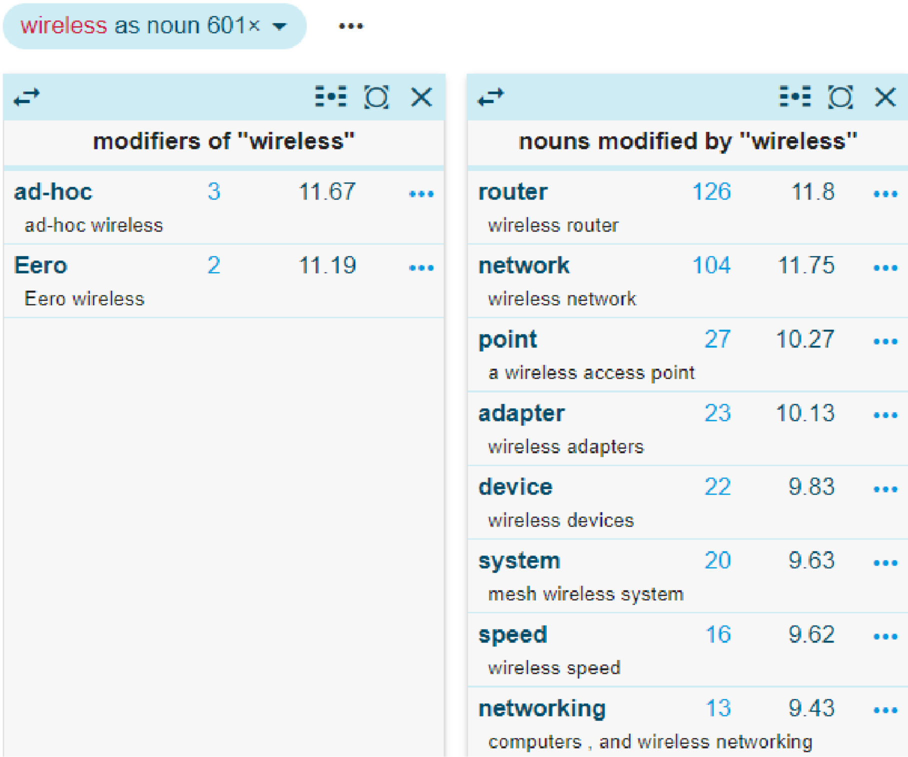

Figure 14. Word sketch results for analysis of the term 'wireless' in the corpus for analysis of 'wireless'

According to the Oxford English Dictionary 'wireless' as a noun may mean "broadcasting, computer networking, or other communication using radio signals, microwaves, etc." ("Wireless, n."). One could assume that a word which is commonly used as an adjective taking the form of a noun would be a phenomenon mostly appearing in popular contexts, as using this simple term may seem less professional than using more specific multi-word terms. In order to find the occurrences of the noun 'wireless' fitting this definition, every single instance of the word had to be examined through concordance by hand, as Word Sketch was unable to automatically differentiate between the noun and adjective form. Out of the total 306 occurrences in the scientific corpus, only one would be considered a proper noun, which can be seen in Figure 19. in context

.

<s> While fiber is capital intensive (cost function of distance) and offers limited availability,	**wireless**	is highly cost effective and flexible [1]. </s><s> Besides, fiber is more exp

Figure 15. Concordance line for the occurrence of 'wireless' as a proper noun in the scientific corpus

A similar manual concordance examination in the corpus for analysis of the term ‘wireless’ yielded 6 occurrences out of the total 601, which are shown in Figure 20.. Since the number of occurrences in both contexts is extremely low due to small sample size, this proves to be no conclusive evidence to say whether or not the noun form of the word ‘wireless’ is mostly a popular phenomenon, even though the frequencies of the term appearing as a noun may imply a larger use in popular text. It does however show that it is not a purely popular phenomenon, as an instance of the term being used in scientific writing was found.

Identify the WLAN design that's best for your situation. </s><s> Benefits of	**Wireless**	</s> <s> Wireless offers tangible benefits over traditional wired networking. </s
l design that's best for your situation. </s><s> Benefits of Wireless </s> <s>	**Wireless**	offers tangible benefits over traditional wired networking. </s><s> Ever tried to
relaxing on your outdoor patio? </s><s> These are just some of the things	**wireless**	can do for you </s><s> Terminology of Wireless Networking </s><s> The field
earch source </s><s> You should find these settings under a menu called "	**Wireless**	", "Wi-Fi setup", or something similar </s><s> If you're unsure of Router 1's SS
as wired routers are connected. </s><s> Connecting two home routers over	**wireless**	is also possible, but in most configurations, the second router can only functio
it closely for a moment. </s><s> On it, you can see several tabs like Setup,	**Wireless**	, Security, Access Restrictions, Application & Gaming, Administration and WRT

Figure 16. Concordance lines for the occurrences of 'wireless' as a proper noun in the corpus for analysis of 'wireless'

3.2.6 Other terms

Although some of the above analyzed terms show a certain degree of change in their use between the two contexts, it is not the case for most of the terms appearing in the keyword extraction lists which can be seen at the beginning of this chapter. A quick Word Sketch result and concordance comparison between the scientific corpus and the corpus of web searches may reveal that the apparent behaviour of many terms in the popular context is extremely similar to that of the scientific context, even to the point of frequently forming the same collocations. Some examples of such terms are ‘server’, ‘routing’, and ‘protocol’. Some other terms, such as ‘latency’ or ‘link’ differ between the two contexts only in the specificity of the words they are being modified by. In scientific writing, these terms are frequently modified by other terms which could belong to the third group that has been specified in chapter 2.1 - specialist networking terms -, while in popular writing, their noun modifiers are much more common general networking terms. This trend could imply that in written popularization, the value of scientific information of described phenomena is usually not diminished.

CONCLUSION

This thesis focused on the corpus-based approach to the study of technical terminology and contrasting its use in both scientific and popular writing. Technical terminology has a great deal of importance in the life of all people, academics and laymen alike. The access to the infinite pool of information, the Internet, may nowadays help the common non-academic understand many concepts, which were previously only available to the most elite of professionals. The Internet today is filled with non-professional informative text, blurring the lines between professionals and laymen, all because of the human need for knowledge which facilitates a market for presentable information on topics that tend to appear in people's lives.

The main objective of the thesis was to introduce the reader into the topic of corpus linguistics and the use of corpora in language analysis, and to provide an analysis of technical terms in popularization by contrasting their use between scientific or academic and popular contexts from the perspective of corpus linguistics. The greatest obstacle when employing a corpus-based approach in any linguistic study is the large impact every choice made may have. Due to this fact it is extremely important to provide an extensive and detailed description of the methodology used while conducting any corpus-based research. In the methodology section of this thesis, a classification has been made to highlight the type of words which were later analyzed. Furthermore, the process of creation of all corpora used in this study has been described with an attempt to justify the choices made during the creation of the corpora.

The analysis of the chosen common networking terms yielded mixed results, with differences described on a case-by-case basis. While many of the specifically chosen and analyzed networking terms have been found to show differences in use between the scientific and popular contexts, an overwhelming majority of technical and networking terms which have not been described in this thesis behave with little difference in both contexts. It appears as if rather than the information being simplified to appease the wide audience of the general populace, it is the audience which is getting more knowledgeable and able to familiarize itself with the previously mentioned professional terminology without too much difficulty. However, this research is merely a shallow lexicological and syntactic look into the growing subject of technical terminology in popular contexts and is not intended to provide definite conclusions on the whole topic due to its small scope. As such, further research into this subject could be done from the perspectives of lexicology and syntax, as well as other linguistic disciplines. Any deeper research would greatly benefit from an increased size of corpora used, as well as a more evenly distributed sampling of the different types of texts used to create the corpora, considering that even in a corpus of nearly 700,000 words a single article had the ability to notably influence the number of occurrences of a term.

LIST OF REFERENCES

BAKER, M. Corpus Linguistics and Translation Studies — Implications and Applications. *Text and Technology*, 1993, p. 233-352. DOI: 10.1075/z.64.15bak.

BAKER, M. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target International Journal of Translation Studies Target*, vol. 7, no. 2, Jan. 1995, pp. 223–243., DOI: 10.1075/target.7.2.03bak.

BOZDĚCHOVÁ, I. *Současná terminologie: se zaměřením na kolokační termíny z lékařství*. Prague: Karolinum, 2009. Acta Universitatis Carolinae. ISBN 978-80-246-1539-4.

KILGARRIFF, A., GREFENSTETTE, G. Introduction to the Special Issue on the Web as Corpus. *Computational Linguistics*, 2003, vol. 29, no. 3, pp. 333–347. DOI: 10.1162/089120103322711569.

KILGARRIFF, A., RYCHLÝ, P. *SketchEngine* [online], 2003. http://www.sketchengine.eu. Accessed 12. June 2020.

KUEBLER, S., ZINSMEISTER, H. *Corpus Linguistics and Linguistically Annotated Corpora.* New York: Bloomsbury, 2015. ISBN 9781441116758.

KRHUTOVÁ, M.. *Parameters of Professional Discourse/ English for Electrical Engineering*. Brno: Tribun EU, 2009.

LEECH, G. Corpora and Theories of Linguistic Performance. *Directions in Corpus Linguistics,* 1992. DOI: 10.1515/9783110867275.105.

MCENERY, T., HARDIE, A. *Corpus linguistics: method, theory and practice.* New York: Cambridge University Press, 2012. ISBN 978-0-521-54736-9.

MCENERY, T., WILSON, A. *Corpus Linguistics: an Introduction*. Edinburgh: Edinburgh University Press, 2001. ISBN 0-7486-1165-7

MCENERY, T., XIAO, R., TONO, X. *Corpus-based language studies: an advanced resource book*. London: Routledge, 2006. ISBN 0-415-28622-0

"Node, n.." *Oxford English Dictionary*, Oxford University Press, 2020, https://www.lexico.com/en/definition/node. Accessed 12. June 2020.

SINCLAIR, J. *Preliminary recommendations on corpus typology*. EAGLES Document TCWG-CTYP/P (available from http://www.ilc.pi.cnr.it/EAGLES/corpustyp/corpustyp.html), 1996.

"Topology, n.." *Oxford English Dictionary*, Oxford University Press, 2020, https://www.lexico.com/en/definition/topology. Accessed 12. June 2020.

"Wireless, adj.." *Oxford English Dictionary*, Oxford University Press, 2020, https://www.lexico.com/en/definition/wireless. Accessed 12. June 2020.

"Wireless, n.." *Oxford English Dictionary*, Oxford University Press, 2020, https://www.lexico.com/en/definition/wireless. Accessed 12. June 2020.