MASTERARBEIT / MASTER’S THESIS

Titel der Masterarbeit / Title of the Master‘s Thesis

"Evaluation of manual versus automatic alignment and term extraction quality in the medical domain from English into Romanian"

verfasst von / submitted by

Andreea Maria Popa

angestrebter akademischer Grad / in partial fulfilment of the requirements for the degree of

Master of Arts (MA)

Wien, 2023 / Vienna, 2023

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 070 354 331
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Masterstudium Translation Rumänisch Deutsch
Betreut von / Supervisor:	Univ.-Prof. Dragos Ciobanu, PhD

Table of Contents

Abstract

This Master's Dissertation focuses on investigating the accuracy and consistency of automatically aligned English-Romanian bilingual corpora related to Covid-19, as well as on evaluating the manual and automatic term extraction techniques. While the automatic alignment of these resources can be useful, it may not always be accurate. Moreover, there is a current shortage of Covid-19 glossaries available that contain translated terms from English into Romanian. This shortage of standardized terminology has led to inconsistencies in the translations present in the collected corpora. This research addresses these issues together with the fundamental aspects of terminology. The empirical research uses Trados Studio, MultiTerm and the Sketch Engine to evaluate the accuracy of the automatically aligned English-Romanian resources related to Covid-19 and term extraction techniques. The end result is a high-quality English-Romanian glossary and a comprehensive evaluation of both manual and automatic term extraction techniques.

Abstrakt

Diese Masterarbeit befasst sich mit der Untersuchung der Genauigkeit und Konsistenz von automatisch alignierten englisch-rumänischen zweisprachigen Korpora im Zusammenhang mit dem Thema Covid-19, sowie der Bewertung von manuellen und automatischen Terminologieextraktionstechniken. Während das automatische Alignment dieser Ressourcen nützlich sein kann, ist es möglicherweise nicht immer genau. Außerdem, gibt es derzeit einen Mangel an Covid-19-Glossaren, die übersetzte Begriffe von Englisch ins Rumänische enthalten. Dieser Mangel an standardisierte Terminologie hat zu Inkonsistenzen in den Übersetzungen geführt, die in den gesammelten Korpora vorhanden sind. Diese Forschung befasst sich mit diesen Problemen sowie mit den grundlegenden Aspekten der Terminologie. Die empirische Forschung nutzt Trados Studio, MultiTerm und die Sketch Engine, um die Genauigkeit der automatisch alignierten englisch-rumänischen Ressourcen im Zusammenhang mit dem Thema Covid-19 und der Terminologieextraktionstechniken zu bewerten. Das Ergebnis ist ein hochwertiges englisch-rumänisches Glossar und eine umfassende Bewertung sowohl der manuellen als auch der automatischen Terminologieextraktionstechniken.

Introduction

Since its initial emergence in 2019, Covid-19 has been a subject of significant interest among medical professionals. As the global pandemic persists, researchers have conducted and documented extensive studies in various formats, such as research papers, public information articles, and press releases. To ensure that this information reaches a broader audience, many of these documents have been translated into multiple languages, including English and Romanian.

Nowadays, as technology continues to develop, relevant information contained in these documents can be easily extracted by computer systems. This process involves scanning and extracting the most relevant data from various sources that belong to a certain field. The automatically extracted information is then also aligned, meaning it is organized in a standardized format that makes it easier to analyse and share.

Finally, this processed information is shared in publicly accessible collections of text data also known as corpora. This allows for greater access to the vast amount of information generated by Covid-19 research as well, allowing researchers and the general public to stay informed and up-to-date on the latest developments in the fight against the virus.

Such resources raise a number of challenges, however. Firstly, the accuracy of the automatic alignment between English and Romanian is not always guaranteed to be perfect, which can have a negative impact on further research that relies on these bilingual corpora as reference materials.

Secondly, there is currently a shortage of Covid-19 glossaries containing terms translated from English into Romanian, leading to inconsistencies in the translations present in the gathered corpora. This lack of standardized terminology makes it more difficult to accurately compare and analyse the information contained within these resources, potentially hindering the progress of Covid-19 research. Consequently, this Master's Dissertation will investigate the following research questions:

RQ1: What is the alignment accuracy of the automatically-aligned chosen bilingual corpora?

RQ2: How consistently is EN-RO Covid-19 terminology used in the chosen bilingual corpora?

RQ3: What is the precision and recall of the automatic term extraction tool the Sketch Engine against a gold standard produced through manual human terminology extraction?

The first chapter of this study will focus on providing a comprehensive overview of the Covid-19 pandemic, including its outbreak, development, and impact on daily life. The second chapter will deal with the fundamental aspects of terminology, exploring various theories from this domain, but also the processes of terminology work, including terminology extraction, as well as the tools required to perform such tasks effectively. The aim is to provide an understanding of the most important concepts and techniques used in the field of terminology. The third and fourth chapters will focus on the automatic term extraction, the translation quality assessment and the tools and software used in these processes.

The empirical research, which is thoroughly described in the fifth chapter, will rely on manually processing the chosen bilingual corpora using Trados Studio and the Qualitivity plug-in for annotation purposes to answer the first research question, the MultiTerm integration into Trados Studio to answer the second research question and also to create the gold standard for the third research question. The last research question will be answered by employing the Sketch Engine. The Discussion, found in the sixth chapter, will analyse the findings of the empirical work and the Conclusion will summarise the main findings of this research alongside limitations and areas of future work.

Upon completion of this Master's Dissertation, I will have created a high-quality English-Romanian (EN-RO) glossary by carefully selecting relevant medical terms from the chosen bilingual resources. Additionally, I will have performed a comprehensive evaluation of both manual and automatic term extraction techniques, by calculating the precision and recall for the automatically extracted terminology using the Sketch Engine against a gold-standard produced through manual extraction. The paper will also include an analysis of the quality of several automatically aligned English-Romanian resources related to Covid-19.

The end result of this research will be a comprehensive understanding of the quality and utility of automatic term extraction performed by the Sketch Engine and of the automatically aligned English-Romanian resources related to Covid-19, as well as the development of a valuable tool in the form of a high-quality bilingual glossary.

1. The COVID-19 pandemic

1.1. A brief introduction

The outbreak of the Covid-19 pandemic has brought unprecedented global challenges that affect every aspect of our lives. In this first chapter, we will take the opportunity to explore the pandemic's context, including its timeline, transmission, symptoms, and impacts on society. Moreover, we will delve into the specialised terminology used in the medical domain as well, which will help readers familiarise themselves with the subject matter as they will see highlighted specialised terminology being used in context. Therefore, specialised terminology belonging to the medical field will be written in *italics* in the following chapters. Building on this contextual understanding and specialised terminology, the next chapter will deal with the fundamentals of terminology and explore how it can aid effective communication in the pandemic domain.

In the early days of December 2019, reports began to surface in Wuhan, China of *patients* who were *suffering from* severe *symptoms* associated with *pneumonia*. Their cause remained, however, unknown. Investigations into the matter began and it was soon discovered that many of these individuals either worked or lived in close proximity to the Huanan seafood wholesale market, where live animals were also sold. As the situation progressed, some of these *patients* began to rapidly develop serious *complications* such as *acute respiratory distress syndrome* (*ARDS*) or *acute respiratory failure* (Chen et al. 2020; Hamed 2020; Ovallath 2020; Sarvepalli 2020).

As a response to this *outbreak*, the *Chinese Center for Disease Control and Prevention (CDC)* quickly took action by conducting extensive testing on the affected individuals and thoroughly analysing the results. Due to their hard efforts, a new *strain of coronavirus* was identified in a *throat swab sample* which was taken from one of the patients who was manifesting these symptoms. Based on the results, the CDC immediately initiated an emergency response to cease the spread of this new virus (Chen et al. 2020; Hamed 2020; Ovallath 2020; Sarvepalli 2020).

To make matters worse, the timing of the outbreak coincided with the Spring Festival in China, which represents a period of major population movement and a popular season among travellers. Unfortunately, this facilitated the spread of the virus and the outbreak seemed, thus, inevitable. It soon became clear that this was the case, as the disease began to spread not only across China, but also around the world (Chen et al. 2020; Hamed 2020; Sarvepalli 2020; Ovallath 2020).

This novel *coronavirus* was later named *2019-nCoV* by the *World Health Organization* (WHO), which declared the *outbreak* a "Public Health Emergency of International Concern" on January 30th. As the situation continued to worsen, the *WHO* eventually announced the *COVID-19 pandemic* as a global outbreak which has had a significant impact on public health and economies across the globe (Chen et al. 2020; Hamed 2020; Sarvepalli 2020; Ovallath 2020).

1.2. A timeline

In December 2019, the attention of public health officials worldwide was drawn to Wuhan city, which was referred to as the source of the *outbreak* of what appeared to be a severe form of *pneumonia*. Initial reports suggested that the cases were related to a seafood market, which lead to speculations around the possibility that this *disease* had been transmitted from animals to humans. It later became increasingly clear that this was a serious and quickly evolving situation (Sarvepalli 2020:2).

The *World Health Organization (WHO)* also got involved in the matter on December 31st, 2019, after being alerted by the Chinese health department about the outbreak. The WHO and the local authorities started working closely together afterwards, on investigating and appropriately responding to the situation. It was, thus, determined that this disease was caused by a new strain of coronavirus, which later became known as the COVID-19 (Sarvepalli 2020:2).

According to Sarvepalli (2020), the disease can easily spread among humans through droplets or merely direct contact, as shown by an *epidemiological investigation*. The ease of travel and the large number of people who were moving around during that time, also helped the virus to spread rapidly around the globe. By the end of January 2020, the WHO had placed this outbreak among the *public health emergencies*, as cases began to appear in countries such as Italy, Iran, the UK, Spain, France and the United States (Sarvepalli 2020:2).

The *COVID-19 pandemic* became a major problem which has affected around 1.6 million people from 205 countries and territories around the world. It has also lead to widespread *illness* and death as well as to major disruptions to economies and societies worldwide (Sarvepalli 2020:2).

1.3. SARS-CoV-2 structure

The term "corona" has a rich history and connotation, as it is derived from the Latin word "corona" meaning crown or halo. It's not just a coincidence that the name *coronavirus* was adopted by *virologists* since the *virus* has a distinct crown-like shape, with spikes protruding from its surface. These spikes are what give the virus its characteristic appearance, and they are also what make it so distinctive. S. Kang et al. (2020) refer to the shape of coronaviruses claiming that they are named after "the crown-like (or corona in Latin) spikes of the virus protruding to the periphery, with a diameter of 60–160 nm observed by electron microscopy" (S. Kang et al. 2020:1).

The virus is also relatively large, with a diameter of around 120 nm, which is visible under electron microscopy. According to Hamed (2020), coronaviruses cause severe *acute respiratory syndrome* (SARS) and is mostly *fatal*: "Coronaviruses cause severe acute respiratory syndrome (SARA) leading to death in most cases. Coronaviruses are *single-stranded RNA viruses*, about 120 nm in diameter. " (Hamed 2020:1).

According to Sarvepalli (2020:1), not only humans can be affected by this group of viruses, but also other mammals or even birds, statement confirmed by S. Pandey (2020) as well, who notes that "SARS-CoV-2 is an animal coronavirus that originated from a bat and phylogenetically belongs to the genus beta-coronavirus (subgenus: Sarbecovirus) and shares considerable similarities to human coronaviruses responsible for earlier outbreak: SARS and MERS" (S. Pandey et al. 2020:1).

COVID-19 is a highly *contagious* disease that has affected millions of people worldwide. The virus is able to remain viable in the air for up to three hours, as reported by Ovallath (2020). This means that the virus can easily be spread through the air by *droplets* produced when an infected person *sneezes* or *coughs*. The virus can also be spread through direct contact with an infected person. It's important to note that people who are infected with the virus may not show any symptoms, yet still be able to spread the virus to others. This is why it's essential for individuals to take the necessary precautions to protect themselves and others from infection. COVID-19 can affect people of all ages, but it's particularly dangerous for older adults and individuals with *underlying health conditions*. According to Ovallath (2020), these individuals are more likely to experience severe symptoms and complications caused by the disease (Ovallath 2020:4).

However, it is important to note that the symptoms of COVID-19 can vary greatly from one person to another. Some individuals may experience only mild symptoms or no symptoms at all, while others may have severe symptoms that require *hospitalization*. Additionally, some individuals may experience symptoms that are not typically associated with COVID-19, such as *skin rashes* or *neurological symptoms*. It should also be mentioned that the virus can cause long-term effects, such as *chronic fatigue*, *difficulty breathing*, and other persistent symptoms. Ovallath (2020) mentions "*fever, dry cough, myalgia, back pain, nausea* with or without *vomiting, abdominal, discomfort, diarrhea, loss of smell, loss of taste, anorexia* and *fatigue*" among the most common symptoms (Ovallath 2020:4). Only a small number of patients require hospitalization and *ventilation* due to "*acute respiratory distress syndrome*" (Ovallath 2020:4) an even fewer suffer from *cardiac failure, renal failure, multi-organ failure* or even death.

According to Ovallath (2020), COVID-19 can be easily diagnosed using methods such as *nasopharyngeal swab, nasal swab, bronchoalveolar lavage*, or blood tests taken two weeks apart. The results typically take a few hours to two days to be received. Ovallath stresses the importance of promoting hand hygiene, proper coughing etiquette, and social distancing through public education as a means to prevent the spread of the disease. However, the most effective ways to curb the spread of COVID-19 remain isolating infected individuals and administering vaccines (Ovallath 2020:42).

1.4. Transmission

The main route of transmission is through droplets produced by infected patients through coughing, sneezing or even speaking. Ovallath claims that "an uncovered cough can lead to droplets travelling up to 4.5 meters" (Ovallath, 2022:48). Airborne could also be considered as one of the sources of *transmission*, as a large number of COVID-19 cases seem to have been caused by asymptomatic individuals through *aerosols* produced while speaking. According to Prather et al., these aerosols can remain in the air for hours and can be thus easily inhaled (Prather, Wang, and Schooley 2020:1423).

The size of respiratory droplets plays an important role in determining the distance they can travel through the air. According to Prather et al. (2020:1422), smaller droplets are more easily affected by air currents and can travel longer distances, while larger droplets are affected by gravity, which causes them to settle on various surfaces, where they can remain viable for a number of hours. Pandey et al. (2020:3) have found, for instance, that these droplets can survive up to 6.8 hours on plastic surfaces and up to 5.6 hours on stainless steel surfaces. As Ovallath (2020:47) points out, contaminated surfaces can also serve as a source of transmission, as people may unknowingly come into contact with the virus by touching these surfaces and then touching their face.

The silent spread of COVID-19 through *asymptomatic patients*, also known as "silent shedders", as described by Prather et al. (2020:1423), is a major concern for *public health officials*. These individuals may not be aware that they are infected and can unknowingly transmit the virus to others, since the *incubation period* could range from 14 to even 27 days in extreme cases, as stated by S. Pandey (2020:3). This makes it difficult to detect and contain the spread of the virus. Furthermore, Prather et al. (2020:1423) reported that 79% of COVID-19 cases in Wuhan, China were caused by asymptomatic patients, emphasizing the importance of implementing measures to detect and isolate asymptomatic cases.

1.5. Diagnostics

Kang et al. (2020) differentiate between three major testing methods: *chest imaging*, *haematological examination* and *molecular diagnosis*. Chest imaging consists of *radiography*, in the early stages of the disease, and a *chest CT* only if required, in more severe cases. During the beginning of the disease, the radiography should reveal *interstitial alterations* as well as small plaques around the edges of the *lungs*. Once the disease progresses, these lesions can also appear in the central area of the lungs "with multiple infiltrating shadows and/or ground-glass opacities" (Kang et al. 2020:3).

The second testing method, namely the *haematological examination*, has proven that COVID-19 patients have a lower *white blood cell count* and also decreased *lymphocyte* count in the early stages of the disease (Sarvepalli et al. 2020:3).

Kang et al. classify the *molecular-based diagnostic tests*, such as *real-time polymerase chain reaction* (PCR) and reverse transcription polymerase chain reaction (RT-PCR) as the most reliable and accurate testing methods which are able to detect the presence of COVID-19. The samples are collected from multiple parts of the body, including the upper respiratory tract, using nasopharyngeal and *oropharyngeal swabs*, the *lower respiratory tract*, *blood* and even *faeces*. The authors also underline the fact that they are highly sensitive and are able to provide specific results, which makes these type of tests the most accurate in terms of COVID-19 detection (Kang et al. 2020).

Even though they provide a high level of accuracy, there are several factors that could affect the results of these tests. Kang et al. (2020:3) point out that a negative test result does not exclude the possibility of a person being infected with SARS-CoV-2. The sample might get contaminated inside of the laboratory as well, which could lead to a false-positive test result. On the other hand, there is also the possibility of a false-negative test result, which might occur if there is not enough viral material provided in the sample. Since the amount of viral material present in the body varies depending during the progression of the disease, the timing of the testing is also a key element.

S. Pandey et al. mention another testing method, namely *serological tests*, which are claimed to identify *IgM and IgG antibodies* in the *samples* collected from patients. They are formed against the causative *viral antigens*, *nucleocapsid protein* and *spike protein* of the virus (S. Pandey et al. 2020:3).

According to Sarvepalli (2020), the United States of America approved on March 28th 2020 an antibody test kit that is able to provide results within 15 minutes (Sarvepalli et al. 2020:3).

1.6. Symptoms

The onset of symptoms for COVID-19 can vary greatly from one person to another, with the typical *incubation period* being around 5 days, but it can range anywhere from 0 to 21 days or longer. Factors such as the individual's *immune system* can also play a role in the length of the incubation period. The most common symptoms of COVID-19 include *fever*, *headache*, *coughing*, *muscle pain*, *difficulty breathing*, and *fatigue*. Less common symptoms include *confusion*, *sore throat*, *runny nose*, *chest pain*, *diarrhoea*, *nausea*, and *vomiting*. Some people may experience only mild symptoms at first, which can worsen over time, leading to more severe symptoms such as *shortness of breath*, *acute respiratory distress syndrome (ARDS)*, *septic shock*, *metabolic acidosis*, *coagulopathy* and *multiple organ failure* as reported by Ovallath (2022), S. Kang et al. (2020), N. Chen et al. (2020) and Guan et al. (2020).

According to Ovallath (2020), the *clinical manifestations* of COVID-19 can be divided into four levels of severity which illustrate the gravity of the disease. The first category includes all *mild diseases*, where patients experience only *mild symptoms* such as fever, headache, couch, muscle pain and fatigue, but no evidence of *pneumonia*, however. On the other hand, for patients belonging to the second category the symptoms include fever, respiratory difficulty, and evidence of pneumonia as well, as the author explains (Ovallath 2020:23f).

The third category comprises of all sever cases, while the fourth category is designated for critical cases. The patients in the third category experience *shortness of breath*, evidence of pneumonia, *respiratory failure* and even *Acute Respiratory Distress Syndrome (ARDS)*, which is a severe form and requires hospitalization. The patients in the fourth category are at a risk of death, experiencing severe respiratory failure (Ovallath 2020:23f).

According to the author, these levels of severity are important for properly *diagnosing* and efficiently treating patients, as well as for preventing further spread of the virus. Moreover, even individuals who were infected by *asymptomatic patients* were prone to sever symptoms associated with this disease, as explained by Ovallath, who, thus, stresses the importance of taking precautions and being aware of the potential spread of the virus (Ovallath 2020).

According to a study cited by Kang et al. (2020), performed on a significant number of patients diagnosed with COVID-19, it was determined that they also had decreased *white blood cell counts* and *lymphocytopenia*. This was likely due to the fact that the virus normally targets the *immune system*, leading to a suppression of the body's ability to fight off *infection*. Additionally, the study found that the majority of *chest computed tomography* (CT) *scans* of COVID-19 patients showed *lesions* in multiple *lung lobes*. These lesions were often dense, indicating that the virus was causing widespread damage to the lungs. This is consistent with the fact that COVID-19 primarily affects the *respiratory system* and can lead to severe *lung damage* in some patients (Kang et al. 2020:3).

It has been proven that *critically ill patients* experience very low to no *fever*, while patients suffering from low fever and *tiredness* are more likely to be *mild patients*. Although, most COVID-19 patients make a good *recovery*, death cannot be completely ruled out, especially in the case of elderly or individuals suffering from other *comorbidities* such as *diabetes, chronic lung disease, cardiovascular disease*, and *cancer*. Therefore, the *mortality* of SARS-CoV sums up to 10% (Chen et al. 2020; Sarvepalli 2020).

1.7. Prevention

One of the key elements in preventing the spread of COVID-19 is good hand hygiene which should comprise of washing the hands for at least 20 seconds or using hand sanitizers with at least 60% alcohol. Respiratory hygiene also plays an important role in preventing the spread of the *disease* and it includes covering the mouth and nose when coughing or sneezing. Touching the face needs to be avoided as well, especially the eyes, nose and mouth. Ovallath also mentions the importance of maintaining a decent social distance of at least 1 meter from others and avoiding crowded places as much as possible (Ovallath 2020:50).

Prather et al. stress the importance of wearing *surgical face masks* in order to keep the airborne spreading of the virus at bay. According to the authors, face masks could make a difference, especially in the case of asymptomatic individuals, who "might be a key driver in the global spread of COVID-19" or those experiencing mild symptoms who have not yet been confirmed positive (Prather et al. 2020:1424). Face masks help the wearer prevent transmitting as well as avoid getting the virus. "Thus, it is particularly important to wear masks in locations with conditions that can accumulate high concentrations of viruses, such as health care settings, airplanes, restaurants, and other crowded places with reduced ventilation" (Prather et al. 2020:1424).

Self-isolating and *quarantining* are also key elements in preventing *disease transmission*: "Quarantine of persons is the restriction of activities of or the separation of persons who are not ill, but who may have been exposed to an infectious agent or disease. This helps in early detection of cases and thus prevents its spread" (Ovallath 2020:50). The author also mentions that the duration of quarantine should be at least 14 days from the time of exposure or laboratory confirmation of COVID-19, as recommended by WHO.

In order to help avoiding further spreading of the disease, individuals should be tested for COVID-19 before and after travelling and self-isolate in case related symptoms occur.[1]

1.8. The impact of the COVID-19 pandemic

The measures put in place to curb the spread of COVID-19 have had a significant impact on various industries and sectors. The arts, entertainment, and live performance industries have been gravely affected, as events and gatherings were cancelled or postponed. The tourism industry has also been greatly affected as travel restrictions were implemented and people were encouraged to avoid nonessential travel. The transport industry, including airlines and trains, has seen a decrease in demand due to fewer travellers and gatherings in public places which has also impacted hotels and restaurants as a chain reaction (Eurofound 2021).

All in all, the economy has suffered severe consequences as a result of the COVID-19 pandemic. For instance, many people have lost their jobs due to businesses either closing or drastically reducing operating hours. This has led to an inevitable decrease in consumer spending which has further impacted the economy. Overall, the travel restrictions, social distancing, and

[1] https://www.lecturio.com/covid-19-coronavirus-disease-2019/

health protocols have had a ripple effect on various industries and sectors, and the economy as a whole (Eurofound 2021).

Despite the introduction of COVID-19 *vaccines* providing a glimmer of hope, a resurgence of the *virus*, commonly referred to as a second wave, was ultimately unavoidable. This resurgence led to a marked increase in the number of *infected individuals*, resulting in the reintroduction of measures and restrictions which, in turn, had a profound impact on both the economy and daily life. According to Eurofound, an independent agency of the European Union, unemployment rates were expected to rise to 8.6% in 2021 and stabilize at 8% in 2022. Given the European Union's effective policy response, the rates were still below those experienced during the global economic crisis of 2008-2010, even despite these increases (Eurofound 2021).

The pandemic has had a wide-reaching impact, affecting not only the economy and labour market, but also the daily lives of citizens. Many individuals have struggled financially as a result of job loss or reduced working hours, and have been unable to pay for basic necessities such as rent, mortgages, and loans. Additionally, the restrictions on social life have had a negative effect on citizens' overall well-being and have contributed to mental health issues (Eurofound 2021).

The onset of the COVID-19 pandemic has not only had *physical health* implications, but has also impacted the *mental health* of individuals. Ettman et al. (2020) state that the COVID-19 pandemic is closely related to the increase in mental health problems in the United States: "The COVID-19 pandemic has been associated with a substantial increase in *mental illness*. U.S. adults reported an estimated three-fold increase in the prevalence of elevated *depressive symptoms* at the start of COVID-19 pandemic relative to before it" (Ettman et al. 2020:1).

According to different studies cited by Ettman et al. (2020:1), *depression* symptoms have been observed not only in the general public, but also in healthcare workers and even children. If left untreated, mental health problems, could lead to an increase in mortality rates, as an indirect effect of the pandemic and the measures implemented in order to keep it at bay. Isolation, uncertainty and loss of social connections have been identified as well as major factors causing mental health problems. The stress and anxiety of living with the pandemic and the fear of getting infected or losing loved ones, have also added to the mental burden of people. Additionally, the economic crisis and job insecurity have contributed to the increase of mental health issues as well.

2. Fundamentals of terminology

In the first chapter of this thesis, the readers were introduced to medical terminology used in context, highlighting its significance in the healthcare industry. However, this chapter delves deeper into the scientific perspective of the matter and aims to provide an understanding of what terminology truly entails. The current section aims to provide readers with a more comprehensive understanding of terminology and its role in facilitating accurate communication. By doing so, it is hoped to emphasize the importance of standardization and consistency in medical terminology for effective healthcare communication.

2.1. What is terminology?

Terminology refers to the specialised vocabulary of a particular field. According to the DIN 2342 standard, terminology is defined as the "total amount of terms belonging to a certain field" (DIN 2342, 1992:2). Oeser and Budin (1998) also define terminology as "a set of terms in a subject area with their associated designations or language labels" (Oeser & Budin 1998:9).

According to Sager, terminology is "the study of and the field of activity concerned with the collection, description, processing and presentation of terms, i.e. lexical items belonging to specialised areas of usage of one or more languages" (Sager 1990:2).

Terminology can be approached from a variety of angles, as explained by Sager (1990:3) in his work. One perspective is that of terminology as a set of practices, which involve the collection and description of terms specific to a particular field. This approach emphasizes the practical application of terminology, focusing on the methods and techniques used to identify and organize terms within a particular domain.

Another perspective is that of terminology as a set of theories, which aim to explain the relationship between certain terms and concepts within a field. This approach focuses on the underlying principles and theoretical framework that govern the use of terminology, and the way in which these concepts are interrelated (Sager 1990:3).

Finally, the author suggests that terminology can be viewed as the vocabulary of a specialised domain. From this perspective, terminology could be understood as the set of terms and expressions used by experts within a particular field, and the way these terms are used to communicate and convey specialised knowledge. This view emphasizes the communicative aspect of terminology, and its importance in facilitating effective communication within a specific domain (Sager, 1990:3).

According to Lynne Bowker, terminology is the discipline that focuses on the lexical items that are specific to a certain subject field, and the processes involved in collecting, processing, describing and displaying these terms: "the discipline concerned with the collection, processing, description and presentation of terms, which are lexical items belonging to a specialised subject field (e.g. medicine, law, engineering, library science or art history)" (Bowker 2015:304).

Terminology and the process of naming, organizing, describing, and translating specialised vocabulary within a specific field, has become increasingly significant in recent times due to advancements in science and technology. According to Faber Benítez, the information found in texts within specialised fields is often encoded "in terms or specialised knowledge units, which can be regarded as access points to more complex knowledge structures" (Faber Benítez, 2009:108). They act, thus, as gateways to deeper and more intricate knowledge structures. These specialised knowledge units, also known as terms, are an essential aspect of terminology and play a crucial role in facilitating communication and understanding within a specific field.

2.2. Specialised texts

In order to better illustrate the difference between terms and words it is important to define and understand the role of texts and especially the role of specialised texts. Santos C. & Costa R. define specialised texts as follows:

> The specialized text may simultaneously be understood as the production and the product of a community of controlled communication. In the text we can find all the linguistic elements that result from the interaction of language with society, which allows the text to be analysed simultaneously as a process and as a result. (Santos & Costa 2015:157f)

Santos and Costa argue that specialised texts serve as the foundation for identifying terms and specialised knowledge, as they provide a subject for examination and analysis. However, there is disagreement about what constitutes the characteristics of specialised texts: "Some say they can find terms, others concepts or even knowledge itself, while others still say they find representations of knowledge" (Santos & Costa,2015:158).

Specialised texts are known for their rich vocabulary and language specific to a particular field or domain. These texts often contain a high concentration of technical terms and jargon that may be difficult to understand for readers who are not familiar with the specific field. Because of this, specialised texts are typically intended for and understood by readers who possess a similar level of knowledge and expertise as the author. According to Santos and Costa (2015:158), these texts serve as the primary means of communication for experts within a professional community. They use specialised language and terminology to share their knowledge and insights with other members of their field. Additionally, specialised texts are also a useful tool for analysts and researchers to identify, understand and analyse the terms and specialised information that exist within a certain domain. However, the characteristics of specialised texts are a subject of debate among researchers.

2.3. Language for Specific Purposes

Language for Specific Purposes is distinguished from Language for General Purposes (LGP) based on a range of distinctive features. This distinction is often based on the subject field, also known as the subject domain, in which the language is used. ISO's definition of special language incorporates the notion of subject field, describing it as "language used in a subject field and characterized by the use of specific linguistic means of expression" (ISO 1087-1 2000). In contrast, LGP refers to a collection of words and expressions that are not associated with a specialised activity in the context they are used (Warburton 2015:370).

The author further discusses how LSP (Language for Specific Purposes) differs from LGP (Language for General Purposes) in terms of linguistic properties and communicative function. LSPs are associated with a subject field and have restricted linguistic properties such as specific rules e.g. preponderance of nominal structures, and frequency of illustrations. LSPs are strictly informative and allow for an objective, precise, concise, and unambiguous exchange of information. In contrast, LGP can be evocative, persuasive, imaginative, and even deceptive. Any text that varies from LGP text can be considered an instantiation of LSP, and all LSPs have a limited number of users, a formal or professional communicative situation, and an informative function (Warburton 2015:370f).

2.4. Terminology science - linguistics and terminology

Depecker refers to terminology science as "the study of technical and scientific vocabularies" (Depecker 2015:35). The author also mentions that language can be studied from several points of view: "from signs that are part of a language, from ideas to which language refers or from objects that language designs" (Depecker, 2015:35). In order to better understand these approaches, it is necessary to differentiate between linguistics and terminology science.

Ferdinand de Saussure, a prominent linguist, believed that while linguistics deals with the study of language and the signs used in it, terminology science focuses on the proper usage of "signs", "concepts" and "objects" in relation to each other. According to Saussure, terminology science ensures that the signs used in the language correspond to the concepts that they represent and that these concepts correspond to the objects they describe. He defines "signs" as the linguistic aspect of terminology science, and concepts as the mental representation of certain objects, which are a product of imagination. In simpler terms, terminology science ensures that the correct sign is used to denote a specific concept and that this concept is accurate to the object it represents (Depecker 2015).

Another difference worth mentioning between linguistics and terminology is that the former studies words while the latter deals with the study of terms. Depecker (2015) distinguishes between the two as follows:

"A "word" is a linguistic unit which, most of the time, is easily isolable in a language. The sense may be vague, multiple or changing. As it is when we try to define "happiness", "future" or "spirit". A "term" refers mostly to a technical or scientific reality. Most of the time, the definition of a term can be precise, and validated by the description of the concept and of the object it refers to (Depecker 2015:37).

According to Depecker (2015), the term is, thus, closely linked to the sign or linguistic unit and the concept it represents. In contrast, words have the ability to become technical or scientific terms when used in a specific field or domain. Depecker gives the word "mouse" as an example, which in the field of data processing is a technical term that refers to a specific tool. This highlights the importance of the context or the domain in which words are used, as common words can take on a technical meaning when used in a specialised field (Depecker 2015:38).

2.5. Terms and words

Terms, the "structural elements of discourse that point towards the concepts that exist outside language" (Santos & Costa 2015:158) are considered to be very similar to words by Temmerman (2000:109), who claims that both words and terms are based on prototypes.

Words are defined by Lyons as linguistic units (Lyons 1968:170f) (morpheme, word, phrase, clause) and Sager claims that terms are "a functional class of lexical units" (Sager 1998:168). Based on the affirmation that terms are a functional class of lexical units, the difference between words and terms is the domain or the field in which they are used. The domain is therefore essential in labelling linguistic units as "terms". It would be impossible to identify terms without extralinguistic information, while words can be identified only within linguistic expressions, according to Kageura (2015:47). The author emphasizes, thus, the importance of extralinguistic information in identifying terms, as it is impossible to do so solely based on linguistic expressions.

Warburton (2015:381-385) lists the most important criteria that must be met for linguistic units to be classified as "terms": frequency of occurrence, embeddedness, visibility, and translation difficulty. In other words, a term's value within an organisation is directly related to how often employees access it. If a term occurs frequently in company documents, it is more likely that employees will query it, thus increasing its value. In a commercial environment, the frequency of a term's use may be more important than its classical semantic criteria in determining its relevance. However, it is important to note that when new terms are introduced, their frequency may not be high initially, but they should not be overlooked or discarded. This means that while frequency is important in determining the value of a term, it should not be the only factor considered.

Embeddedness refers to the ability of the term in question to form other terms and is especially important in specialised domains, where terms play a crucial role in conveying information and understanding. Warburton offers the term "sustainable development" as an example. This term can either stand alone or be embedded in another term e.g. "ecologically sustainable development economics" (Warburton 2015:382). Translation difficulty is another criterion for identifying terms, as specialised terms often do not have an equivalent in another language, making translation more difficult.

Visibility is also a very important criterion. Warburton considers that the more frequently a term is used, for example within a company, both on web pages and on a software interface, in an advertising slogan or in company materials, the more important it is that the term is understood and translated correctly. The author claims that texts dealing with legal issues or even marketing texts can lead to legal conflicts if the terms are not translated correctly into the target languages. Warburton (2015) also underlines the importance of creating a term base at a company level. A term base plays an essential role especially with regard to terms that are more difficult to understand and translate. It is also of high importance for translators as it helps them to find the correct translation without having to do the same research work multiple times (Warburton 2015).

2.6. Technical terminology

Roelcke (2018) describes technical terminology as the vocabulary used in specialised human activities, which includes not only technology, but also any given science or institution. Exactness, uniqueness, and authenticity are the three properties of technical terminology that are necessary for an adequate definition of terms. Exactness refers to the clear assignment of a term to the designated objects and facts, while uniqueness refers to the absence of polysemy and synonymy of the words concerned. Authenticity is based on "a motivated assignment of meanings and their expressions themselves", meaning that metaphors, metonymies, or improper meanings are regarded as problematic (Roelcke 2018:492). However, vagueness, ambiguity, and metaphors are common in technical terminology, but these properties are not considered to imply misunderstanding, as communicative indicators can help resolve them.

2.7. Term bank

According to Bowker (2015:305), it is crucial to store all terminological information electronically, using tools like term banks, in order to ensure effective translation. Conducting thorough research on specialised terminology and finding the best equivalents can be time-consuming and repetitive, leading to inefficiency in the translation process. This problem led to the widespread use of glossaries and dictionaries, but their electronic versions have proven to be more efficient, as they allow for easy storage and accessibility of terminological data. As a result, electronic tools such as term banks have become increasingly popular among translators, since they help to save time and effort (Bowker 2015:313).

According to the author, there are two approaches to manage terminological information related to translation. One approach is to create a comprehensive, standardized multilingual resource known as a *term bank*. This type of resource is typically large in size and contains a vast array of terms in multiple languages, which makes it ideal for organizations with a wide range of translation needs. The other approach is to create a smaller, more personalized resource known as a *term base*. This type of resource is typically tailored to the specific needs of an individual or organization, and contains only the terms that are relevant to their particular field or area of expertise.

According to Bowker (2015), a term bank is a large collection of electronic records that contain information about terms and the concepts they represent. Term banks are often used to help those who need specialised information about different terms and concepts (Bowker 2015:308).

A term bank, as described by Bowker (2015:308), is a centralized, comprehensive, and standardized repository of terminological information. Its purpose is to support the information needs of professionals, experts, and academics who work in various fields and need to understand specialised terms and concepts. The information stored in a term bank is usually multilingual, meaning it contains equivalent terms in different languages. Additionally, it may include definitions, synonyms, antonyms, translations, and other information related to the respective terms and concepts.

Sager lists five different categories of term bank users: experts of a certain domain, information scientists who index specialised documents, language experts whose task is to maintain, develop and standardize language, people who are studying or teaching languages or a certain domain and those who use language for professional purposes, such as translating or communicating with others (Sager 1990:197-200).

According to Bowker, professional translators are the primary users of multilingual term banks, such as TERMIUM Plus, Inter-Active Terminology for Europe (IATE), UNTERM, and the Grand dictionnaire terminologique (GDT). These term banks are known to be reliable sources for translators, and are often created by significant institutions for their internal translators. These are, however, also accessible to the general public through the internet (Bowker 2015: 308).

On the other hand, terminologists must carefully review documents, texts, and corpora to identify potential terms that could be added to a term bank. The selection process is based on various criteria, including the frequency of word usage and the relationships between different types of words (Bowker 2015).

2.8. Term base

According to Bowker, term bases are typically created by using pre-existing translated materials for terminology purposes, whereas term banks are the result of a thorough selection process which is carried out by terminologists. Term bases are smaller and more personalized collections of terminology that can be managed using computer-aided translation tools like translation memory (TM) systems. (Bowker 2015:310f).

Translators can reuse the terminology used in the original translation if they are confident that it was of good quality. However, as translators do not typically engage in exhaustive research, term extractors have not been a popular tool for them until recently. With the integration of term extractors into TM systems, translators can now use them to generate a list of all the lexical items that appear a specified number of times in the TM database, making the terminology management process easier and more efficient (Bowker 2015:310-314). However, it is important to note that many of the term extractors built into CAT tools are just statistical in nature and may not always be reliable. These extractors generate a list of lexical items that appear a specified number of times in the TM database, but they do not necessarily take into account the context or meaning of those terms. Therefore, while term extractors can make the terminology management process more efficient, they should not be relied upon blindly and should always be used with caution.

2.9. Translation Memory

Translation Memories (TMs) are extensively used in the translation profession as a database of pre-translated texts, which are usually aligned with their corresponding source texts at the sentence level, as explained by Bowker (2015:312). However, it is important to note that TMs may also contain individual segments that are not part of proper texts, which can make the terminology management process more difficult.

These databases can help identify equivalent term pairs, which can be directly inserted into a new target text or added to a term base. The author claims that translators may occasionally refer to the TM for terminology and if the TM contains translations previously done by the same translator, they may reuse them with confidence. Bowker goes on to stress the fact that while translators may feel comfortable using their own terminology research, clients may provide TMs of unknown origin that consist of contributions from various translators with different levels of proficiency. In such cases, translators may have to rely on others' translated material for terminological research, which may perpetuate the errors contained in the TM in future translations (Bowker 2015:312).

According to Reynolds (2015), the usual TM tools present the translator with a window that consists of the source text, the target text, and the translation results. The translation results show matches found in the TM and, depending on the term base, can propose potential term matches from the terminology database for the terms in the source text (Reynolds 2015:279).

2.10. Theories of terminology

Faber Benítez (2009) addresses two different kinds of terminology theories: the prescriptive terminology theory and the descriptive terminology theory. The author goes on to explain that the two categories encompass a wider range of sub-theories that have developed over time. Generally, prescriptive terminology theory is concerned with providing guidelines for the proper use of terms, whereas descriptive terminology theory is focused on describing the actual use of terms in a particular field or domain. The General Terminology Theory (GTT), which is further explained in the next sub-chapter, was the first theoretical proposal in the field of terminology, and it is primarily prescriptive in nature. However, subsequent theories in this area have taken a more descriptive approach, increasingly incorporating principles from cognitive linguistics. This shift in focus towards the social, communicative, and cognitive aspects of terminology has resulted in a more realistic view of how terms are used and behave in texts. As a result, these newer theories offer a more accurate and nuanced understanding of terminology (Faber Benítez 2009:110f). They are further explained in the subchapters addressing the Communicative Theory of Terminology, Cognitive-based Terminology and Frame-based Terminology.

2.10.1. General Terminology Theory (GTT)

The General Terminology Theory, which is the first theoretical approach to the study of terminology, falls under the category of prescriptive theories. This theory was developed by Eugen Wüster, who is also known for creating a series of French-English dictionaries containing standardized terms. Wüster believed in the general standardization of scientific language. However, as Faber Benítez notes, this theory has a rigid character, as it separates terms from the concepts they denote. This means that terms are seen as merely a linguistic label for the concepts they represent, whereas concepts are seen as separate entities that represent real-world objects. In other words, GTT focuses on specialised knowledge concepts, viewing them as separate from their linguistic designation (terms), which are merely their linguistic labels. This emphasis on differences conveyed the idea that terms were not even language at all, but rather abstract symbols referring to concepts in the real world. However, GTT's rigid view of the semantics of specialised language limited the representation of knowledge concepts, without allowing for their multidimensional nature. GTT did not account for syntax or pragmatics as well (Faber Benítez 2009:111f).

The General Terminology Theory posits that specialised language terms are distinct from words in general language. This idea is based on the belief that there is a clear, one-to-one relationship between specialised terms and the concepts they represent. GTT asserted that terms in specialised language are inherently different from general language words due to their monosemic reference between terms and concepts. However, specialised language texts show that terminological variation is frequent, and that such variation seems to stem from parameters of specialised communication, such as knowledge, prestige, text function, and content. Additionally, terms have distinctive syntactic projections and can behave differently in texts, depending on their conceptual focus. Syntax was not considered part of Terminology within GTT, which also exclusively focused on synchronic language and ignored the diachronic dimension of terms (Pavel et al. 2001:19).

In the 1990s, a shift occurred in the field of terminology as a response to the limitations of the General Terminology Theory. Instead of solely focusing on the technical aspects of terms, new theories began to emerge that emphasized the social, communicative, and linguistic aspects of terminology. This shift resulted in a broader perspective that integrated terminology into the larger context of communication. These new theories examine how terms are used in different communicative situations, providing a more realistic view of terminology and its role in communication. Additionally, they investigate the sociological and discourse factors that contribute to the creation of various types of texts (Faber Benítez 2009:111f).

2.10.2. Communicative Theory of Terminology (CTT)

The communicative Theory of Terminology brings linguistics and terminology closer together as it explores the specialised language from a more complex point of view (Faber Benítez 2009:114).

Cabré's (2003) theory of terminology, known as the Cognitive Theory of Terminology (CTT), emphasizes that specialised knowledge units have cognitive, linguistic, and sociocommunicative components, which make them multidimensional. In other words, Cabré (2003) points out that knowledge units can be studied and analysed from a cognitive, linguistic and communicative perspective, as they behave like words from the general language. Cabré argues that focusing on one perspective does not negate the presence of the other two dimensions, which remain implicit. The CTT proposes the Theory of the Doors, a metaphor that represents the three dimensions of a terminological unit: a cognitive dimension, a linguistic dimension, and a communicative dimension, which are separate doors through which terminological units can be accessed (Cabré, 2003).

The CTT regards terminological units as "sets of conditions". Therefore, according to the CTT, the term is a result of a number of influencing factors: "their particular knowledge area, conceptual structure, meaning, lexical and syntactic structure and valence, as well as the communicative context of specialised discourse" (Faber Benítez, 2009:114).

The CTT has led to a valuable body of research on different aspects of terminology, such as conceptual relations, terminological variation, term extraction, and the application of different linguistic models to terminology. However, the CTT has some shortcomings. It avoids opting for any specific linguistic model and does not seem to give much importance to conceptual representation, knowledge structure or ontology, and category organization. Another area in need of clarification in the CTT is semantic meaning, as it seems to avoid the question of what specialised meaning is and what its components are. Cabré's theory seems to implicitly support some type of semantic decomposition, but nothing is explicitly said about the semantic analysis of specialised language units (Faber Benítez, 2009:115).

2.10.3. Cognitive-based Terminology

The Cognitive-based Terminology Theory and the Communicative Theory of Terminology both analyse terms within the context of text and discourse and do not perceive them as isolated entities. However, cognitive-based theories take this approach one step further by incorporating elements from cognitive linguistics and psychology. This provides a more universal understanding of how terms are used and processed. The Sociocognitive Terminology proposed by Temmerman (2000), which looks as the social and cognitive aspects of terms and their usage, and the Frame-based Terminology by Benítez, which examines how terms are framed within different communicative contexts and how this influences their meaning and usage, are among the most notable subcategories of the Cognitive-based Terminology (Faber Benítez 2009).

2.10.4. Frame-based Terminology

Frame-based terminology theory views specialised knowledge units as integral parts of a larger context and examines them in relation to their behaviour within texts, rather than as separate entities. This theory posits that there is little difference between specialised terms and general language words and suggests that specialised terms should be studied in terms of how they function within the context of texts. This perspective is similar to the Communicative Theory of Terminology (CCT) and is also considered a cognitive approach to terminology as it emphasizes how terms are perceived, used and processed. Frame-based terminology, therefore, emphasizes the importance of considering the communicative context in which terms are used in order to understand their meaning and usage.

Faber Benítez (2009) underlines the three main focuses the frame-based terminology: "(1) conceptual organization; (2) the multidimensional nature of terminological units; and (3) the extraction of semantic and syntactic information through the use of multilingual corpora" (Faber Benítez 2009:123).

2.11. Terminology work

Terminology work is defined by ISO 1087-1 (2000:10) as follows: "work concerned with the systematic collection, description, processing and presentation of concepts and their designations".

Karsch (2015) outlines a set of six activities that terminologists perform to ensure accurate terminology work. These activities include investigating concepts and determining the meanings of different terms, identifying the designations assigned to these concepts, and documenting the findings in the form of terminological entries. Another important task that terminologists are responsible for is finding equivalents in the target language. Additionally, they may also assist in naming new concepts or terms. Overall, terminologists play a crucial role in ensuring the consistency and accuracy of terms within a specific field or organization. The activities mentioned by Karsch are important steps in the process of terminology management, which aims to organize and maintain terms and their associated information in a systematic and consistent manner (Karsch 2015:293).

Faber (2015) explains, in simpler terms, that terminology work involves collecting, analysing and distributing specialised terms. He claims that terminology work plays an important part in further various tasks and activities such as technical writing and communication, knowledge acquisition, specialised translation, development of knowledge resources and information retrieval. Terminology work helps, thus, ensure consistency and accuracy of terms, which, in turn, improves communication and understanding within a specific field or organization. The author describes terminology work as an important aspect of knowledge management, which deals with organizing, maintaining and leveraging knowledge within an organization. It helps in making the information more accessible and understandable for the intended audience (Faber 2015:14).

Dobrina (2015) also argues that the underlying motivation behind terminological work is the need for effective communication, whether within or between fields. According to the author, this need arises in situations such as when a researcher is attempting to explain a concept to colleagues, or when translators are searching for an appropriate equivalent. The main purpose of terminological work is to facilitate clear and accurate communication by creating reliable terminology resources (Dobrina 2015:180).

Castellví (1998) differentiates between two types of terminology: ad-hoc and systematic. Ad-hoc terminology is typically limited to 60 terms or less and is performed in response to a specific task, whereas systematic terminology involves more than 60 terms and focuses on establishing a system of concepts within a specific subject field (Castellví 1998:152).

Terminology can therefore be either ad-hoc, or take the form of systematic work. It can be regarded as a single-project on a certain subject, or as part of a bigger context and can be, therefore, used either in an academic setting or as a task attributed by a public institution or civil society with the scope of improving communication and learning.

2.11.1. Data categories

Cerrella Bauer (2015:335) emphasizes the importance of the terminological entry. In her opinion, the terminological entry is the core component of a terminology management system, and she claims that it should encompass all information related to a specific term. As Warburton (2015:385) explains, the information is divided among fields in the term base, which reflect different data categories (such as Part of speech, Definition, and Subject field).

According to ISO 12616 (2021), the mandatory data categories for every terminological entry are the term, its source, and the entry date. These categories may either follow the ISO 12616 guidelines or be customized to meet the user's specific needs. For example, a terminology project that covers multiple subject fields may consider the subject field a mandatory data category. At the same time, Bauer (2009) highlights the significance of the source of the term, as it provides information about the quality of that certain term. Additionally, definitions are also important because they provide essential information for a deeper understanding of the term or for its proper translation (Cerrella Bauer 2009).

Warburton claims, on the other hand, that "definitions are time-consuming to prepare, the effort is not commercially justifiable" (Warburton, 2015:386) and introduces the alternative of context sentences.

While ISO 12616 does not mandate definitions or context sentences as compulsory data categories, they are widely considered essential for effective terminology management. Definitions provide a clear understanding of the meaning of a term, while context sentences provide examples of how the term is used in a particular context. Without these elements, it is difficult for users to fully understand and use the term correctly, leading to potential errors and misunderstandings. Therefore, while not compulsory, including definitions and context sentences in a termbase can greatly enhance its usefulness and effectiveness. It is up to the terminology manager to decide which data categories to include in the termbase, based on the specific needs of the project and users.

Since data categories can sometimes play an important role in the process of comprehending the meaning of a certain term, the Part of speech should almost always be included among the data categories, according to Warburton (2015). This is due to the fact that the part of speech value can be used to determine the meaning of a word that has several possible meanings. The author gives the example of the word "port", which can either be a noun indicating the location on a computer where devices like printers can be connected, or a verb meaning the action of converting a software so that it can work on a different operating system. Consequently, the Part of speech plays a decisive role, especially when dealing with homographs, words with the same form, but different meaning (Warburton 2015:385).

All in all, the terminological entry is a valuable component of a terminology management system, and it should provide all information needed in order to comprehend the meaning of a certain term. Apart from the data categories proposed by ISO 12616, other categories can be added as well, based on the user's specific needs, such as synonyms, antonyms, usage notes, which represent information on how the term is used in the context, including any regional or dialectical variations that may exist, contextual information, representing information about the domain or subject area the term is part of, including any relevant cultural or historical background and even illustrations or images that can help to clarify the meaning of a term.

2.11.2. Concepts and definitions

The concept is the mental representation of objects in both the physical and abstract world, and serves as a way of labelling those objects. However, when a concept is unknown, clarification is often necessary. This is why definitions play an essential role in the process of understanding concepts and facilitate effective communication by providing a clear and concise explanation of the object being designated. In essence, definitions help cover the gap between an abstract idea and its representation, allowing for smoother and more accurate communication.

As Pius ten Hacken (2015) explains, the terminological definition should provide enough information for understanding the concept. Delivering a precise definition may be difficult, though, as most concepts rely on prototypes. The author also claims that an accurate definition could be achieved by determining exact boundaries. This is usually the case in legal and scientific fields: "The enforcement of laws and the evaluation of scientific claims depends on precise definitions of the underlying concepts" (Pius ten Hacken 2015:3).

Imposing a terminological definition can be challenging due to a number of reasons. One possibility is that the concept or idea exists already in people's minds and is attributed to a certain prototype structure. Another factor can be the introduction of different concepts under the same name in various theories and domains. Moreover, definitions may also need to be modified when new information becomes available. However, according to Pius ten Hacken, a terminological definition establishes a brand-new, abstract object that exists regardless of the speakers' linguistic competence (Pius ten Hacken 2015:3).

The need for additional clarification arises if a word or phrase describing a particular concept fails to be understood due to a lack of necessary background information. The additional information may be introduced as a definition, which is the most common method of clarification, defining context, or a detailed explanation. The definition can be further divided into a number of subcategories (Löckinger et al. 2015:60).

The intensional definition, which is the most common type of definition, is explained by ISO 1087-1 (2000:6) as follows: "definition which describes the intension of a concept by stating the superordinate concept and the delimiting characteristics". Therefore, intensional definitions tend to focus more on the superordinate concept. This means, that in order to be able to fully understand a certain concept which, in turn, relies on more superordinate concepts, one may need to consult the other definitions first. This may require reading several other intensional definitions for understanding one concept. Consequently, intensional definitions tend to be more abstract (Löckinger et al. 2015:63).

The intensional definition has gained widespread acceptance and has become the norm for defining concepts in various disciplines, including logic, chemistry, and linguistics. It is frequently utilized in dictionaries, encyclopaedias, and specialised databases of terms, and has proven to be an effective way to describe the meaning and essence of a concept. This type of definition has become so prevalent that it is now considered the standard method of concept illustration across a range of fields (Löckinger et al. 2015:63).

For a better understanding of what the intensional definition consists of, the authors offer the following example, quoting Freeman (1984:56):

> (1) convention
>
> treaty, usually between more than two States, concerning matters of mutual interest
>
> 1. Superordinate concept (generic concept): "treaty"
>
> 2. Delimiting characteristics:
>
> a. signed between more than two States
>
> b. concerns matters of mutual interest (Löckinger et al. 2015:63)

Intensional definitions, while useful, have limitations that must be considered. Unlike extensional definitions, intensional definitions lack the specific terms used to describe subordinate concepts. This means that intensional definitions can be more abstract than extensional definitions, making it difficult to understand related concepts from the definition itself, except for the explicitly mentioned superordinate concept. Furthermore, a single intensional definition may refer to other concepts defined elsewhere, requiring one to read one or more additional intensional definitions to fully comprehend the concept at hand. While consistent drafting of intensional definitions is essential for systematic terminology work, the potential for multiple cross-referenced definitions complicates their practical application (Löckinger, Kockaert, & Budin 2015:67).

On the other hand, ISO 1087-1 (2000) defines an extensional definition as a "description of a concept by enumerating all of its subordinate concepts under one criterion of subdivision." This definition highlights that extensional definitions focus on the extension of a concept and enumerate all of its subordinate concepts. However, this type of definition cannot be considered a proper terminological definition as it does not describe the concept in its entirety. This is in contrast to intensional definitions, which typically provide a genus and species distinction, presenting a superordinate concept followed by necessary delimiting characteristics to describe the intension of the concept (Nilsson 2015:85f).

This type of definition is based on the extension of the concept being defined. An example of an extensional definition can be found in the discussion around the definition of a planet in our solar system. Prior to 2006, Pluto was considered a planet, and it belonged to the extension of the concept "planet". In fact, Professor Hans Rickman from the department of Physics and Astronomy at Uppsala University and former secretary-general of the International Astronomical Union (IAU) remarked that the only way to define a planet in our solar system was to enumerate all the planets from Mercury to Pluto. This type of definition is characteristic of extensional definitions and differs from intensional definitions in which a concept is defined by its necessary and sufficient conditions (Nilsson 2015:81).

2.11.3. Termhood

Warburton (2015:378) explains that termhood "refers to the properties that qualify a linguistic expression as a term", while Cabre Castellvi claims that the term is a "meaningful text unit in a specialised discourse considered useful for an application" (Cabre Castellvi 2007:1f). Warburton is of the opinion that terminologists should avoid including the following in the term base:

1. A higher proportion of non-nouns;
2. General lexicon words and expressions;
3. Widespread variants;
4. Proper nouns (names or titles of organizations, programmes, services, products and documents);
5. Sentence fragments;
6. Boilerplate text (such as copyright statements and legal notices). (Warburton 2015:378)

According to Martin, there are several key factors that should be taken into account when evaluating the "termhood" of a lexical unit in the field of terminology, which places a greater emphasis on the quality of terms rather than their quantity. These criteria include: "conceptual analysis; the nature of pre- and post-modifiers; term frequency; the context and domain and term embeddedness" (Martin 2011:10).

3. Automatic term extraction

The use of correct terminology is vital for translators. However, their perspective on terminology is different from that of terminologists. According to Cabré (2010), translators see terminology as a means to solve particular problems in translation. From the translator's standpoint, terminology encompasses all the unknown words in the source text that need to be further investigated. Therefore, translators focus not only on linguistic units that would be considered as proper terms by terminologists, but also on other units that require specialised translation within a specific subject area (Oliver 2017:150).

The process of identifying and compiling a list of important terms or term candidates (TCs) specific to a particular field or domain is known as *terminology extraction*. This process can be conducted either manually, by a trained terminologist with the assistance of a domain expert, or automatically using computer algorithms. Although manual terminology extraction is generally considered more precise and accurate, it can be time-consuming and subjective. On the other hand, automatic terminology extraction utilizes computer-based methods to quickly identify term candidates, but it relies solely on the evidence present in the text corpus, as well as the reference corpus, should such a resource be available, without considering the nuances of language or the specific context of the terms. As a consequence, the final list of terms still requires the confirmation of a domain expert, as computers struggle with the complexities of language and meaning (Heylen and De Hertog 2015:216).

Automatic terminology extraction is also known as terminology mining, term recognition, glossary extraction, term identification, and term acquisition. The process of automatic term extraction is a multi-step approach and has been thoroughly described by Heylen and De Hertog. The first step involves collecting a relevant corpus of texts from a specific domain, known as corpus collection. The second and third steps involve the detection of unithood and termhood, respectively. The detection of unithood refers to identifying linguistic elements that represent a single conceptual unit, while the detection of termhood involves determining which extracted linguistic units are valid terms (Heylen and De Hertog 2015).

It is important to note that high unithood does not guarantee high termhood. For example, even though the collocation "most of the time" may have a strong unithood, it has weak termhood within a specialised domain. Unithood refers to the degree of cohesion between words or collocations in a group, as explained by Vo et al. (2022). On the other hand, termhood is achieved when a word or group of words refers to a specific concept within a domain (Vo et al. 2022:122).

The fourth and fifth steps in automatic term extraction, as outlined by Heylen and De Hertog (2015:204), are the detection of term variants and evaluation and validation of the extracted terms. This involves comparing the results of the automatic extraction process to a manual extraction performed by a domain expert or a pre-existing gold-standard glossary of domain terminology. The precision and recall of the term candidates are also calculated in this step.

Precision refers to the number of correctly identified terms from the list of term candidates. Due to the limitations of the gold-standard glossaries, correctly identified terms can sometimes be, however, considered as incorrectly identified. Recall, on the other hand, refers to the proportion of terms identified out of all terms present in the specialised corpus, and can be calculated if an exhaustive gold-standard glossary is available. However, recall cannot be determined in the case of manual term extraction (Heylen and De Hertog, 2015:216).

The importance of terminology extraction is emphasized by Thurmair (2003), who sees it as a fundamental step towards various practical applications. One such application is terminography, where the list of term candidates extracted automatically is used to create specialised dictionaries or term databases for a specific field. This list also plays a crucial role in the translation process, serving as an ad-hoc glossary that can assist in identifying unknown words that require clarification, and helping to maintain consistency throughout the translation.

Terminology extraction is closely related to Term Alignment, also known as Bilingual Term Extraction, in a translation context. The goal of term alignment is to identify equivalent domain-specific terms across two or more languages, as explained by Heylen and De Hertog. The list of term candidates also acts as an index for a collection of documents, making it easier for users to search for specific topics (Heylen and De Hertog 2015:204).

The automatic term extraction relies on various methods used to determine whether a single word or a group of words qualify as a term or not. Heylen and De Hertog (2015) explain that words that are commonly present in multiple texts within a corpus are more likely to belong to the general vocabulary, which means that they do not necessarily meet the criteria for being specialised terms. On the other hand, should a word or a certain combination of words appear in only a limited number of texts, then they are considered domain-specific and could potentially qualify as specialised-terms. In other words, the less frequently a term candidate is utilized in the documents, the higher the likelihood that it is, in fact, a specialised term.

Another method worth mentioning with regard to determining the termhood of a term candidate is the analysis of its morphological structure. The medical field is given in this case as an example by the authors since Latin and Greek words and expressions are predominant here and are also an indication of termhood (Heylen and De Hertog 2015:213).

Contrastive Term Extraction is another approach used to determine the termhood of a term candidate. This method involves comparing the frequency of a particular word in a corpus from a specialised domain with the frequency of the same word in a corpus that belongs to either the general vocabulary or a different domain. This comparison helps to identify which words are specific to the specialised domain, and which words are part of the more general vocabulary. By contrasting the frequency of the word in two different contexts, the method aims to determine if a word is a specialised term or if it has a broader usage in the general language (Heylen and De Hertog 2015:214).

3.1. Statistic and linguistic automatic term extraction

The Automatic Terminology Extraction (ATE) methods can be broadly classified into two categories: statistic and linguistic. The first category includes statistical methods, which often consist of generating n-grams and filtering them using a list of stop-words (common words that are not useful in analysis). Among the various statistical methods, frequency is the most common and simplest characteristic used to obtain and rank term candidates, according to Astrakhantsev et al. (2015), who emphasize, thus, the importance of frequency in ATE methods. By focusing on the frequency of words, ATE methods aim to identify terms that are commonly used in the specialised domain and distinguish them from words that are part of the general vocabulary (Astrakhantsev et al. 2015).

The second category, the linguistic methods, is based on linguistic properties. Many ATE systems make use of a set of predefined morphosyntactic patterns and frequency is only utilized to rank the obtained term candidates (Daille et al. 1994). To further refine the terminology, additional techniques can be implemented. For instance, a stop-list of unwanted words can be utilized to eliminate term candidates that contain any of those words. Oliver (2017:367) suggests that stop-lists typically include function words and generic words that are frequently used in the language, such as "this", "that" and "thing". Most ATE approaches extract stop-list words automatically from a general corpus using frequency, which are then validated with the help of human experts. The use of a stop-list can remove false terms that are commonly used in generic collocations in the language.

The outcome of the above-mentioned methods of automatic terminology extraction is a list of term candidates, or units that present characteristics that may qualify them to be actual terms. However, it is important to note that this list is generated through an automated process, and therefore, requires a manual review to determine the true terms from the list of potential terms. This manual review is necessary to ensure that the correct terms are selected and to eliminate any false positive results proposed by the automated process (Oliver 2017).

3.2. Parallel corpora and bilingual terminology extraction

In order to properly understand the functioning of bilingual term extraction, it is necessary to first comprehend the concept of parallel corpus. Ivanović et al. define the parallel corpus as follows: "a pair of electronic texts that represent a direct translation of each other" (Ivanović et al. 2022: 234).

Bilingual terminology extraction takes place at the sentence level and is performed on sentence-aligned bilingual texts. Bilingual terminology extraction relies on the "monolingual extraction of terms from the same corpus for both languages involved" (Ivanović et al. 2022: 234).

According to Oliver (2017), consistency in the selection of translation equivalents for source language terms is an essential aspect of specialised translation. He highlights the fact that many terms encountered in translation projects are not included in available term bases (Oliver 2017:148). Although it may be true that many terms encountered in translation projects are not included in available term bases, it is important also take into consideration the influence of various factors such as the languages involved, the domain of the translation project, and the resources provided by clients. Depending on the specific project requirements and available resources, the use of term bases, glossaries, and other reference materials may aid translators in identifying appropriate translation equivalents. However, it is also important to acknowledge that the effectiveness of such resources may vary and that the identification of appropriate translation equivalents may ultimately depend on the translator's expertise, knowledge of the domain, and an understanding of the specific context and requirements of the project.

An automatic method for identifying source terms and choosing their translation equivalents would significantly aid translators and other language professionals. The implementation of an automated approach for term extraction would streamline the translation process, reducing the time and effort required for manual term identification and selection. Oliver (2017) claims that in an ideal scenario, when a translation project is initiated, the project manager or translator assigns appropriate resources to the project, including a specific term base if available. In this ideal scenario it is also possible to perform automatic terminology extraction on the texts to be translated, select relevant term candidates, and automatically select their translation equivalents in available translation memories or parallel corpora. This process can aid the translator in creating a project-specific term base and reduce the time and effort required for terminological queries (Oliver 2017:148). However, in practice, the availability of resources and the extent of automation may vary depending on the project's scope and requirements, leading to a more complex and nuanced approach.

In the same ideal scenario, a computer-assisted translation (CAT) tool can aid the translator in the automatic bilingual terminology extraction process by presenting term candidates and their corresponding translation equivalents in the source and target languages. This process can facilitate the translation process and improve the consistency of the terminology used (Oliver 2017:148). However, in practice, the effectiveness of this approach may depend on the quality and accuracy of the automatic extraction and selection of term candidates, which may vary depending on the complexity of the text, the domain, and the languages involved. The approach also assumes that the term candidates are not known in advance and that they do not come from an existing term base, which may not always be the case.

Morin and Hazem address the concept of comparable corpora as well, which are a type of non-parallel corpus consisting of texts that share similar features such as genre, domain, or time period. The authors claim that term extraction can also be performed on comparable corpora, since the parallel corpora are scarce resources especially in specialised domains (Morin and Hazem 2016:482). For the purpose of this paper, however, the bilingual terminology extraction will take place at the sentence level and will be performed on sentence-aligned bilingual texts belonging to the medical field.

Since medicine is "a rapidly growing and evolving science" (Vo et al. 2022:300), it is a very difficult task for dictionaries to remain up-to-date. According to Vo et. al, terminology lists have been developed in terminologies such as UnifiedMedical Language System (UMLS), Medical Subject Headings (MeSH), and the Systematized Nomenclature of Medicine (SNOMED) for English. Moreover, every language has its own terminology which makes it impossible for dictionaries to cover everything. Therefore, term extraction plays a very important role in the medical field. As manual term extraction is very time-consuming, there is a preference for automatic term extraction in order to keep terminology up to date. Vo et al. also states that: "ATE plays an important role in many systems such as information retrieval, query systems, ontologies, semantic networks, or in automatically building and extending dictionaries" (Vo et al. 2022:300).

3.3. Term alignment

The field of Automatic Term Extraction (ATE) has close connections to several other related disciplines. In a translation context, the ATE is particularly related to Term Alignment, also known as Bilingual Term Extraction. The broader concept of alignment involves the task of identifying matching sentences, words, and phrases in parallel corpora. Term alignment is, on the other hand, more specific and concentrates on pairing domain-specific terms in two or more languages (Heylen and De Hertog 2015).

In essence, term alignment helps to identify terms that have the same meaning across languages and to create a bilingual lexicon or glossary, which can be used to support the translation process and ensure terminological consistency.

3.4. Terminology tools and management

According to Steurs et al. (2015), the knowledge included in specialised documentation is always changing, as most specialised fields are characterized by progress and development. According to Faber (2011:9-29), specialised language is dynamic, and its representation should reflect this. Therefore, the terminology used to describe concepts also changes and evolves. As a consequence, conceptualization or concept formation is also dynamic. A specialised knowledge concept cannot be understood in isolation but is instead part of a dynamic process that requires contextual and situational information. When discussing and analysing concepts in a specialised domain, corpus-based research is typically recommended. Accurately monitoring and describing term variation and evolution within a specialised domain requires the use of up-to-date corpora.

Despite the preference for a direct relation between the term and the concept, there are also cases where polysemy can alter this direct connection. Polysemy can occur in specialised domains as well, since words and concepts can go from general into specialised contexts and from one domain into another. The authors argue that a strategy that combines descriptive, linguistic, and semasiological elements, as well as an analysis of specialised texts, is necessary to gain a deeper understanding of the communication rules in specialised domain (Steurs et al. 2015:224).

Both monolingual and multilingual knowledge management, document management, as well as translation depend on terminology work. It is also of great importance for the understanding of the concept designated by linguistic symbols in the source language and for consequently finding the right equivalent in the target language. Terminology work plays a vital role in ensuring consistency and accuracy in the transfer of information from one language to another.

Steurs et al. emphasize the significance of terminology management, which encompasses the collection, processing, and updating of terminological data: "the capturing, processing, updating and preparation of terminological data" (Steurs et al. 2015:222).

Proper terminology management (TM) is crucial for professionals who deal with specialised knowledge. Effective management of terminology enhances the consistency of texts, leading to improved usability, readability, and transferability to other languages. The TM process involves several stages, including identifying terminological units, extracting them, sorting and validating the terminology, and storing and manipulating it (Steurs et al. 2015).

The term base, which is considered one of the most important components for successful terminology management, is defined by Melby (2012) as a computer database that holds information about specific domain concepts and the terms that denote them: "a computer data base containing information of domain specific concepts and the terms that designate them" (Melby 2012:8). Moreover, to effectively establish and maintain a term base, it is necessary to utilize a software solution.

According to Steurs et al. (2015), there are numerous software solutions available for terminology management that address different needs and goals. These solutions vary in their complexity, ranging from simple solutions that primarily focus on data entry into a basic database structure to more advanced solutions that offer various tools to aid in the input, handling, collection, storage, and retrieval of terminological data throughout the terminology management process.

3.4.1. MultiTerm

MultiTerm is widely considered as one of the top terminology management solutions in the market, as stated by Steurs et al. (2015:225). MultiTerm, which has been under the ownership of RWS since 2021, is well recognized among the translation and terminologist community, as well as project managers, as a top-notch terminology tool. The software was first introduced by Trados in 1990 and since then, it has established a reputation for its robust terminology capabilities.

MultiTerm can be accessed either online through the web browser when using a server-based term base, or locally on the computer via the preinstalled application, MultiTerm Desktop. According to the authors, the software offers extensive search capabilities, allowing users to set source and target languages, utilize wildcards, recognize fuzzy matches, and apply various filters for improved search results. In addition, users have the option to add or remove fields, such as "context" or "context source" as well as rearrange their order (Steurs et al. 2015:229). The platform enables not only the manual entry of terminology, but also swift addition of terms while using Trados Studio for translation.

According to the authors, terms can be extracted from existing source or target documents using the separate software MultiTerm Extract, as well as from pre-existing word lists (such as Microsoft Excel glossaries). The platform also offers the ability for users to add synonyms or additional information related to the term, which facilitates the understanding of the concept. For instance, users can specify that a term is an abbreviation or that it is prohibited from use. Additionally, if a user attempts to add a term that already exists in the terminology database, they will receive a notification and be prompted to either merge or duplicate the records.

As previously mentioned, MultiTerm can be integrated into the translation software Trados Studio, allowing for term recognition in the source text. This integration provides translators with a more comprehensive view of the terms present in the text and their translations, as they are presented in a separate window as a list (Steurs et al. 2015:231).

3.4.2. The Sketch Engine

The Sketch Engine is a corpus analysis software that allows users to build, explore, and analyse large text corpora. It is primarily used for natural language processing and computational linguistics research, as well as for language learning and teaching.

The Sketch Engine allows users to upload and annotate text data, which is then indexed and made searchable. The software provides a variety of tools for analysing the data, including concordances, word sketches, collocates, and word lists. It enables data visualization, text classification, and other advanced features such as machine learning, that involves using algorithms and statistical models to enable computer systems to automatically learn from data.[2]

Machine learning algorithms are designed to identify patterns, relationships, and insights in data, and to use these insights to make predictions or take actions. The process of machine learning typically involves feeding large amounts of data into a model, which is then trained to identify patterns and relationships within the data. Once the model has been trained, it can be used to make predictions or decisions based on new data. The Sketch Engine also provides pre-built corpora which is ready to be used. The software's powerful search functionality allows users to easily find specific words, phrases, or grammatical patterns in the corpus, and to compare the frequencies and collocational patterns of different words.

The Sketch Engine is a powerful tool for understanding language, using algorithms to analyse large amounts of authentic text to identify common and unique usage patterns. It is widely used by linguists, students, teachers and professionals in the publishing and language industries, and contains a vast collection of language corpora in over 90 languages, each containing up to 60 billion words.[3]

The Sketch Engine is a highly beneficial tool for translators as it provides access to parallel corpora that have been precisely aligned. This valuable resource is crucial for translators in obtaining accurate and authentic translation suggestions. The feature of Word Sketches in Sketch Engine can be particularly useful for translators as it helps in identifying the most appropriate word combinations and phrases. This allows for a more natural and fluent translation that is true to the language and culture of the target audience. The Sketch Engine, with its advanced features, can significantly improve the quality of translation work and make the process more efficient.[4]

Terminologists can greatly benefit from the use of the Sketch Engine tool. The advanced monolingual and bilingual term extraction capabilities offered by the Sketch Engine incorporate both statistical and linguistic analysis to produce reliable results. This enables terminologists to easily identify the most relevant and accurate terms in the target language. The extracted terms are

[2] https://www.sketchengine.eu/
[3] https://www.sketchengine.eu/
[4] https://www.sketchengine.eu/

usually of high quality and require minimal additional cleaning or processing, saving terminologists valuable time and effort in their work. Additionally, the combination of statistical and linguistic analysis provides a more comprehensive and thorough analysis, ensuring that the extracted terms are relevant, accurate, and up-to-date. The Sketch Engine is therefore a valuable resource for terminologists looking to improve their term extraction processes and enhance the accuracy of their terminology work.[5]

3.4.3. Terminology extraction using the Sketch Engine

Automatic Bilingual Terminology Extraction (ABTE) has been a topic of research for several decades, according to Baisa et al. (2015). In spite of the continuous research, the field is still not well established and there are some challenges that need to be addressed. According to the authors, the quality of ABTE methods remains low and, moreover, there is a lack of information available on the quality of commercial tools such as MultiTerm4 and Araya Bilingual Extraction Tool5.

Baisa et al. (2015) categorize research in the field of ABTE into two groups: those that rely on a combination of monolingual extraction and co-occurrence statistics, and those that use alternative methods. Alternative methods include using bootstrapping techniques to identify reliable terminology pairs and establish correspondence rules between terms in different languages, and using a phrase-based translation model approach.

The authors claim that according to a study, the ratio of multi-word to single-word terms is 7:36, making it unfeasible to rely solely on single-word alignment methods. This highlights the importance of considering alternative methods in ABTE, as well as the need for continued research in the field. According to Baisa et al. (2015), further progress in ABTE will require a better understanding of the challenges facing the field, as well as more comprehensive and innovative approaches to addressing these challenges.

Traditional methods of term extraction, which only consider frequency, tend to result in a list of term candidates that still require manual cleaning and processing. This can be a time-

[5] https://www.sketchengine.eu/

consuming and labour-intensive task. On the other hand, an effective term extraction method is one that generates a clean and accurate list of terms that require minimal manual processing.[6]

One way to achieve this is by incorporating linguistic criteria together with statistics. By combining both elements, the method can reduce manual processing, as it takes into account the structure and language-specific features of terms. This leads to a more accurate extraction of terms, which better reflects the focus text and its language.[7]

To qualify as a term, a phrase must have the structure and characteristics of terms in the language, and must also appear more frequently in the focus text compared to a general text, when the Contrastive Term extraction is used, for example. This ensures that the terms extracted are specific to the focus text, and that they have sufficient frequency and relevance to be considered meaningful and important. The use of linguistic criteria and statistics in term extraction helps to minimize manual processing, improve the accuracy of extracted terms, and support more effective term management and analysis.[8]

It is important to consider the structure of terms when examining term extraction across different languages. Different languages may have varying requirements for the format of terms. For instance, in English, terms can be composed of nouns, adjectives, and prepositions. However, a phrase that includes a preposition followed by an article and adjective is typically not considered a term, as terms are typically expected to be noun phrases.[9]

Therefore, it is crucial to understand the specific language requirements for terms when extracting them, as this can greatly affect the accuracy and relevance of the extracted terms. This highlights the importance of considering linguistic criteria, in addition to frequency and other statistical measures, when performing term extraction, to ensure that the extracted terms match the language-specific format requirements.[10]

---

[6] https://terms.sketchengine.eu/how-does-it-work
[7] https://terms.sketchengine.eu/how-does-it-work
[8] https://terms.sketchengine.eu/how-does-it-work
[9] https://terms.sketchengine.eu/how-does-it-work
[10] https://terms.sketchengine.eu/how-does-it-work

The frequency of phrases in a text is an important factor in determining whether they are terms or just frequent phrases. When comparing texts from different sources or genres, such as tabloid newspapers and texts on accounting, certain phrases may appear in both, but their frequencies are likely to be different. For example, the phrases "income tax" and "best way" may appear in both types of texts, but "income tax" is likely to appear more frequently in texts on accounting.[11]

This difference in frequency is what allows the system to differentiate between a frequent phrase and a term. By analysing the frequency of a phrase in a particular text or set of texts, the system can identify "income tax" as a term, as it appears more frequently in texts on accounting than it does in tabloid newspapers or other texts.[12]

OneClick Terms is a user-friendly online platform that enables individuals to effortlessly extract terms from text. It leverages the technology from Sketch Engine, a software for corpus building and analysis. The platform performs part-of-speech tagging and lemmatization automatically and then determines the frequency of each phrase that fits the term structure. As explained by Nuțu (2021), lemmatization refers to the process of identifying the dictionary form of a given word, known as its lemma. Within linguistic domains, lemmatization is employed to consolidate all inflected forms of a word into a singular unit for analysis. The application of lemmatization varies based on the specific language and its corresponding rules. In Romanian, for instance, the masculine singular nominative often serves as the lemma for a noun, while a verb's lemma is represented by its infinitive form.

> "Lemmatization is the process of determining the lemma (the word's dictionary form) of a certain word. In the linguistic fields, through lemmatization, all flexional forms of a word are grouped together to be analysed as a single entity. The lemmatization is language dependent and adheres to certain rules. In Romanian, the lemma of a noun is often the masculine singular nominative, while a verb's lemma is the infinitive form." (Nuţu 2021:51)

Lemmatization means that different inflected forms of a term will be considered as the same term. If not for lemmatization, terms like "palette knife" and "palette knives" would be viewed as two

[11] https://terms.sketchengine.eu/how-does-it-work
[12] https://terms.sketchengine.eu/how-does-it-work

distinct terms, and their frequencies would be counted separately, leading to incorrect information about their actual frequency.

The frequency in the specific text is compared to that in a comprehensive corpus of general language text, and a score reflecting the term-like nature of each phrase is computed. Finally, users can download the extracted terms in either txt, csv, or tbx format. [13]

Bilingual terms can be extracted from parallel corpora stored in a Sketch Engine account, but not from preloaded corpora. To extract bilingual terms, the user must first build a parallel corpus in Sketch Engine. The extraction process is initiated from the Sketch Engine interface, but completed in the OneClick Terms interface. The user must use the corpus selector to pick the first language in the corpus, then go to the dashboard and click bilingual terms. The user will then confirm the selected languages and be taken to the OneClick Terms interface where the extraction will be carried out automatically.[14]

[13] https://terms.sketchengine.eu/how-does-it-work
[14] https://www.sketchengine.eu/guide/bilingual-term-extraction/

4. Translation Quality Assessment

4.1. Trados Studio

Trados Studio is a computer-assisted translation (CAT) tool and is widely used in the translation process and industry. It provides a variety of features to make the translation process easier and less time-consuming, including translation memory and terminology management among others.

This software is designed to increase productivity and consistency for translators by providing previously translated content to speed up the process of translation. The translation memory feature allows translators to store their previously translated content and reuse it in future translations. This not only saves time, but also helps to ensure consistency in terminology and phrasing.

The terminology management feature allows users to maintain a consistent and accurate terminology database, ensuring that translations use the correct terminology for the subject matter. This is particularly useful for technical or specialised content, where consistency is especially important.[15]

4.2. The Qualitivity plug-in

Qualitivity is a plug-in developed for use with Trados Studio, and is designed to fully integrate with the API (Application Programming Interface). Its purpose is to facilitate the management of productivity and quality-related data by allowing users to monitor their work on documents directly within the Studio Editor. The plug-in provides users with the ability to track the time spent on translating, revising, and post-editing segments of documents. However, this plug-in goes beyond mere time tracking by also offering granular tracking of each change made to individual segments. This allows for the generation of comprehensive reports based on the data collected, presented in a structured and easily readable format. Overall, Qualitivity is a powerful tool that provides a robust

[15] https://www.trados.com/products/trados-studio/

framework for users to monitor and improve their productivity and quality when working with Trados Studio.

One of the main advantages of using such a tool is its ability to keep a detailed and comparable record of every action taken by the linguist that involves adding, deleting, or modifying the properties of the segments within a document. This means that the record contains not only the most recent version of the segment following the changes, but also a copy of the segment prior to these changes. A very detailed report that captures every modification made to the segments can be generated whenever needed.

The document overview feature in the Qualitivity plug-in summarizes all the changes that a user made while editing a document in one or multiple sessions. It puts all these changes together and presents them in a clear and easy-to-understand format.

The Qualitivity plug-in also includes a function that can be activated from within the Studio Editor, which enables users to import widely used error annotation standards, such as TAUS DQF, SAE J2450, and MQM Core. These standards are based on a series of different quality metrics. The ability to use these standards has significant advantages in terms of measuring the quality of translations during the revision process. The Quality Assessment report that is generated from this data provides an overview of the frequency and context of linguistic errors, as well as who or what translation memory, machine translation, etc. was applied. It is important to note that the type of reports generated from these assessments are relative to the specific evaluation being conducted, meaning that they may also be applicable in evaluating translations from an MT engine within a specific domain. The data can also be exported in Excel and/or XML format. This allows for interoperability by providing a way to integrate the plug-in's data into other platforms or systems, according to the needs of the user. Therefore, after importing the data, further statistical reports, charts, and so on can easily be generated.[16]

[16] https://community.rws.com/product-groups/trados-portfolio/rwsappstore/w/wiki/2251/qualitivity

4.3. The TAUS Dynamic Quality Framework

The assessment of translation quality is a significant topic of concern within the translation industry nowadays, according to Görög (2014). He states that, in order to address this subject, the Translation Automation User Society (TAUS) developed the Dynamic Quality Framework (DQF) in 2011, with the goal of standardizing the evaluation of translation quality. The author explains that in the DQF framework, the notion of quality is considered dynamic because it varies according to the type of content, purpose, and intended audience. DQF encompasses an extensive knowledge base, evaluation resources, and a variety of tools for profiling and assessing translated content (Görög 2014:446). TAUS DQF is also one of the quality metrics integrated in the Qualitivity plug-in in Trados Studio.

As mentioned before, in order to improve the human evaluation of translated content, the Translation Automation User Society (TAUS) has developed the Dynamic Quality Framework (DQF), which, according to Görög (2014), is comprised of an extensive knowledge base on Quality Evaluation. This includes best practices, reports, templates, and a range of tools for evaluating translations produced by both human translators and machine translation engines. These tools enable evaluators to compare translations, assess their accuracy and fluency, measure post-editing productivity, and score translated segments based on an error typology. Additionally, users have the option to select the most appropriate evaluation methods for a given content type (Görög 2014:453).

In DQF's view, the evaluation of translation quality should therefore be flexible and adaptable and should depend entirely on specific quality requirements. This should allow for the selection of the most suitable quality evaluation model. Influencing factors in selecting the most appropriate quality evaluation model are the underlying process, technology, and resources used in the translation process. For instance, a translation provided by machine translation (MT) requires a different quality evaluation model from the one used for human translation (Görög 2014:446).

However, worth mentioning is also the fact that TAUS made the announcement recently that they will no longer be supporting and updating their quality evaluation, the DQF. They made this decision due to the evolving nature of the language industry, which now offers various alternative solutions for quality evaluation and management. Although DQF will still be available for download on their website, TAUS has clarified that they will no longer offer any technical support or updates for the tool.

The Multidimensional Quality Metrics (MQM) is another error typology metric worth mentioning which was developed through careful examination and expansion of existing quality models as part of the QTLaunchPad project that was supported by funding from the European Union. TAUS, on the other hand, developed the DQF in collaboration with its members.[17]

Error typology is the main method for evaluating translation quality, according to Görög (2014:452). While there is some degree of uniformity in its implementation across the industry, there are also variations in the categories, level of detail, penalties, and other factors used. To address this, the Dynamic Quality Framework (DQF) provides an error typology tool that establishes a consistent way of categorizing and quantifying translation errors based on widely recognized industry criteria such as accuracy, language, terminology, style, and country-specific standards.

MQM serves as a comprehensive system for defining and describing quality metrics utilized to evaluate the quality of translated texts, and to pinpoint particular issues in those texts. The framework offers a structured approach for outlining quality metrics by identifying textual features, providing a systematic way to evaluate translations. Its objective is to provide a standardized set of criteria that can be employed to assess the quality of translations (Görög 2014:452).

[17] https://www.taus.net/resources/blog/the-8-most-used-standards-and-metrics-for-translation-quality-evaluation

5. Methodology

The empirical research conducted for this master's thesis relies on manually processing the chosen bilingual corpora using the Qualitivity plug-in in Trados Studio for annotation purposes and the Multiterm integration into Trados Studio to create the gold standard for the automatic term extraction performed by the Sketch Engine. The Qualitivity plug-in in Trados Studio was used to facilitate the annotation process whereby the quality metrics, in accordance with the TAUS DQF standard, were manually added to highlight any translation errors. The main goals of this study is to compare the effectiveness of manual and automatic term extraction methods in identifying relevant medical terms from bilingual corpora, as well as to evaluate the quality of the chosen bilingual corpora.

In this study, the chosen automatically aligned corpora, namely the "Covid-19 antibiotic dataset", which consists of a total of 1034 segments, was analysed for a variety of potential errors. The errors of interest included alignment errors, mistranslations, terminology errors, punctuation errors, spelling mistakes, and other related issues. These errors were manually identified using the aforementioned quality metrics integrated in the Qualitivity plug-in. Following the analysis, an activities report was generated to provide a comprehensive overview of the annotations made, including a summary of the segments where errors were identified, the respective quality metrics that were used, their severity, and any relevant content associated with them. The resulting report serves as a valuable tool for tracking and monitoring errors encountered during the analysis of the automatically aligned corpora, facilitating a better understanding of the dataset and providing a starting point for future research endeavours.

During the analysis process, each segment was examined for the aforementioned potential errors, as well as for specialised medical terms specific to the use of antibiotics in the context of Covid-19. Any identified English terms and their corresponding Romanian translations were subsequently integrated into a previously created term base, using the "quick add new term" function in Trados Studio. This was accomplished by highlighting the relevant term in both the source and target segments, right-clicking and selecting the function. The resulting term database was then opened in MultiTerm, a terminology management tool, which allows for the creation and organization of a database of terms, including their translations in multiple languages, definitions, and other relevant information. In order to ensure the accuracy and consistency of the terms, these were reviewed for spelling errors and adjusted to conform to their appropriate dictionary form where necessary.

To answer the first research question about how accurate the alignment is for the chosen bilingual corpora, it needed to be determined how many segments out of a total of 1034 were misaligned. Therefore, aside from the errors found using the Qulitivity plug-in's quality metrics, I also searched for instances of misalignment, which occurred when the translation of a particular unit of text, such as a sentence or phrase, failed to capture the original meaning. These mistakes were labelled as "mistranslations." Each misaligned segment was then marked as "translation rejected" in Trados Studio to indicate these errors. To measure alignment accuracy for the selected bilingual corpora, a performance metric was used that determines the proportion of correctly classified instances within a dataset, known as accuracy. The alignment accuracy was calculated by dividing the number of correctly predicted data points by the total number of data points, and then multiplying by 100.

After the manual term extraction, the same "Covid-19 dataset" was uploaded to the Sketch Engine interface and underwent the process of automatic term extraction, which was performed by the same Sketch Engine software. The bilingual term extraction functionality was, however, not accessible for the English-Romanian language pair. Therefore, the Sketch Engine was required to perform a separate term extraction procedure on the English, respectively Romanian corpora, yielding distinct sets of results. These results were then exported into four distinct Excel documents: one for the single English term candidates, one for the multi-word English term candidates, one for the single Romanian term candidates, and one for the multi-word Romanian term candidates. However, only a threshold of 1200 English term candidates and 500 Romanian term candidates were then further analysed in this paper.

After generating these four documents, the two Excel documents containing English term candidates were merged into a single list, and the same step was, subsequently, repeated for the Romanian documents as well. The two lists thus obtained, one containing English term candidates and the other Romanian term candidates, were manually filtered and a subset of 1200 English term candidates and 500 Romanian term candidates were selected for further analysis. The selected term candidates were then classified into the following five categories: "not a term", "not a medical term", "similar to term already in the term base", "wrong", and "overlooked during manual term extraction".

Each category is represented by a different colour on the pie chart that is thoroughly explained in the next chapter addressing the findings of this research. This pie chart was created to better illustrate these results: "not a term", denoted by the colour yellow; "not a medical term", represented by the colour orange; "similar to term already in the term base", indicated by the colour blue; "overlooked during the manual term extraction" assigned the colour green, and "wrong", assigned the colour red. The category "not a term" refers to word combinations that were automatically extracted but do not constitute an actual term. The category "not a medical term" encompasses words that could be perceived as actual terms but do not belong to the medical field, however. The category "similar to term already in the term base" was implemented to identify those terms that were extracted by the Sketch Engine, but were already present in the term base, albeit in the form of derivative words or word combinations. The category "overlooked during manual term extraction" designates those terms that were correctly identified as terms by the Sketch Engine, but were not included in the term base as they were overlooked in the process of manual, human extraction. Finally, the category marked red designates words that were either unidentifiable in the source text or erroneous in terms of spelling, for example.

To enable the analysis and categorization of the samples of automatically extracted term candidates, the following criteria, listed by Warburton (2015:382), was taken into consideration: frequency of occurrence, embeddedness, visibility, and translation difficulty. The occurrence frequency of a linguistic unit in a text, particularly in specialised domains, was considered indicative of its term status. Embeddedness, or the ability of a term to form other terms, was also considered essential in conveying information and enabling understanding in specialised domains. Warburton (2015) gives the term "sustainable development" as an example, which can either stand alone or be embedded in another term. Translation difficulty was also among the criteria for identifying terms, since specialised terms may lack equivalents in other languages, making translation more challenging.

The Sketch Engine was the tool utilized for the automatic term extraction which allows users to create and access personal corpora, allowing them to view and download separate lists of up to 1000 terms. The software, therefore, generated four Excel documents: one comprising of single English term candidates, one of multi-word English term candidates, another of single Romanian term candidates, and, finally, one comprising of multi-word Romanian term candidates. The automatic term extraction performed by the Sketch Engine generated 2000 English and 2000 Romanian term candidates. By merging the two lists of English term candidates and, subsequently, their corresponding Romanian term candidates, a total of 4000 term candidates were, thus, automatically extracted. These term candidates were then further filtered and arranged in alphabetical order.

However, for the analysis were selected only the following term candidates: the first 1200 English term candidates, arranged in alphabetical order and the first 500 Romanian term candidates, also arranged in alphabetical order, which were then attributed to the aforementioned categories. Different percentages were obtained for each category, based on the total of 1200, respectively 500 term candidates analysed. The experiment aimed to determine whether the percentages of the "not a term" and "not a medical term" categories were higher than the percentage of the "overlooked during the manual term extraction" category. Should this be the case, it would suggest that automatic term extraction is less efficient than manual, human term extraction.

As already mentioned, the main goal of this study is to compare the effectiveness of manual and automatic term extraction methods in identifying relevant medical terms from bilingual corpora. To achieve this, the two-step approach explained above was adopted. First, the terms were manually extracted from the chosen bilingual resources using Trados Studio. Subsequently, the same task was performed using the automatic term extraction tool in the Sketch Engine. The resulting lists of terms, which were narrowed down to 1200 English term candidates and 500 Romanian term candidates, were then compared to determine the precision and recall of the automated tool in comparison to the gold standard, which was established through manual, human terminology extraction. In the context of terminology extraction, precision refers to the proportion of identified terms that are actually relevant or correctly labelled as terms. It measures how accurate the term extraction process is in identifying valid terms. For example, if a tool identifies 100 words as term candidates, but only 36 of those words are valid terms, then the precision of the tool is 36%. In other words, precision measures the quality of the extracted terms, indicating how many of the identified terms are accurate. In the context of terminology extraction, recall refers to the proportion of actual terms that are identified or retrieved by the term extraction process. It measures how well the term extraction process is able to identify all the relevant terms in a given text corpus. For example, if there are 100 actual terms in a corpus, but a tool only identifies 77 of them, then the recall of the tool is 77%. In other words, recall measures the completeness of the extracted terms, indicating how many of the actual valid terms were correctly identified by the tool. The aim is to investigate whether the manual, human way is superior to the automatic method in terms of identifying the relevant medical terms in the bilingual corpora.

The research conducted and the different methods applied in this study yielded several outcomes that are thoroughly explained in the following chapter.

6. Discussion

6.1. Quality evaluation of translation errors

As already mentioned in the previous chapter, the analysis of the "Covid-19 antibiotic dataset", which comprises of 1034 segments, required the utilization of the Qualitivity plug-in in Trados Studio to annotate different types of errors encountered. This plug-in generated a detailed activities report that provides a thorough overview of the annotations made during the analysis process of the given dataset. The Qualitivity plug-in facilitated the process of manually adding quality metrics, based on the TAUS DQF standard, to identify and highlight translation errors. The analysis sought to identify various errors such as misalignments, mistranslations, terminology errors, punctuation errors, spelling mistakes and other related issues. Whenever encountering any type of error, the quality metrics provided by the Qualitivity plug-in were employed to label and identify the errors. The Quality Assessment report produced from this data offers valuable insights into the frequency, location, context, and responsible party of linguistic errors. It should be noted that the nature of the reports resulting from such assessments depends on the user's evaluation criteria.

The activities report can be generated directly from Trados Studio, namely from the Qualitivity plug-in, in the form of an Excel table that looks like this:

Activity ID	Activity Name	Segment ID	QM Name	QM Status	QM Severity	QM Content	QM Comment	QM Created
1	en-ro_covid19_ant	11	Terminology > Inconsistent	Open	Minor	medicii din asistenţa medicală primară		2023-02-16 20:02:45.123
1	en-ro_covid19_ant	11	Terminology > Inconsistent	Open	Minor	medicii de familie		2023-02-16 20:02:54.914
2	en-ro_covid19_ant	117	Accuracy > Mistranslation	Open	Minor			2023-02-18 12:41:53.165
2	en-ro_covid19_ant	119	Accuracy > Mistranslation	Open	Minor			2023-02-18 12:42:46.361
2	en-ro_covid19_ant	120	Accuracy > Mistranslation	Open	Major			2023-02-18 12:43:18.342
2	en-ro_covid19_ant	121	Accuracy > Mistranslation	Open	Major			2023-02-18 12:44:02.767
2	en-ro_covid19_ant	122	Accuracy > Mistranslation	Open	Minor			2023-02-18 12:44:32.740
2	en-ro_covid19_ant	122	Accuracy > Mistranslation	Open	Major			2023-02-18 12:44:53.074
2	en-ro_covid19_ant	123	Accuracy > Mistranslation	Open	Major			2023-02-18 12:45:25.688
2	en-ro_covid19_ant	131	Accuracy > Addition	Open	Minor			2023-02-18 14:11:49.625
1	en-ro_covid19_ant	14	Accuracy > Mistranslation	Open	Minor	Material		2023-02-16 20:03:28.193
1	en-ro_covid19_ant	14	Terminology > Inconsistent	Open	Minor	medici		2023-02-16 20:04:25.095
2	en-ro_covid19_ant	141	Accuracy > Addition	Open	Minor			2023-02-18 14:31:25.211
2	en-ro_covid19_ant	142	Accuracy > Addition	Open	Minor			2023-02-18 14:32:15.138
2	en-ro_covid19_ant	143	Accuracy > Mistranslation	Open	Minor			2023-02-18 14:32:33.076
2	en-ro_covid19_ant	144	Accuracy > Untranslated	Open	Minor	Remove this filter Audience		2023-02-18 14:33:07.098
2	en-ro_covid19_ant	145	Accuracy > Mistranslation	Open	Minor			2023-02-18 14:33:25.355
2	en-ro_covid19_ant	148	Accuracy > Addition	Open	Minor	pentru îngrijiri medicale primare		2023-02-18 14:37:57.462
3	en-ro_covid19_ant	152	Accuracy > Untranslated	Open	Minor	Remove this filter Audience		2023-02-18 14:50:50.610
3	en-ro_covid19_ant	153	Accuracy > Addition	Open	Minor	antibiotice		2023-02-18 14:51:04.123
3	en-ro_covid19_ant	153	Accuracy > Mistranslation	Open	Minor	Pagina 1 din 1 // 5 Rezultate găsite		2023-02-18 14:51:15.685
3	en-ro_covid19_ant	155	Accuracy > Omission	Open	Minor			2023-02-18 14:54:07.711
3	en-ro_covid19_ant	157	Accuracy > Ambiguous	Open	Minor			2023-02-18 14:54:52.673
1	en-ro_covid19_ant	16	Accuracy > Addition	Open	Minor	care prescriu antibiotice		2023-02-16 20:10:01.675
1	en-ro_covid19_ant	16	Accuracy > Addition	Open	Minor	utile atunci când prescriu antibiotice.		2023-02-16 20:13:06.347
3	en-ro_covid19_ant	163	Accuracy > Addition	Open	Minor	dumneavoastră		2023-02-18 15:09:40.402
3	en-ro_covid19_ant	171	Accuracy > Addition	Open	Minor			2023-02-18 15:58:08.777
3	en-ro_covid19_ant	171	Accuracy > Ambiguous	Open	Minor			2023-02-18 15:57:24.077
3	en-ro_covid19_ant	174	Language > Grammar	Open	Minor			2023-02-18 16:07:26.005
3	en-ro_covid19_ant	176	Accuracy > Addition	Open	Minor			2023-02-18 16:14:21.186
3	en-ro_covid19_ant	176	Language > Spelling	Open	Minor			2023-02-18 16:09:26.408
3	en-ro_covid19_ant	177	Accuracy > Mistranslation	Open	Minor			2023-02-18 16:14:45.000
3	en-ro_covid19_ant	179	Accuracy > Mistranslation	Open	Minor			2023-02-18 16:15:11.443
3	en-ro_covid19_ant	191	Accuracy > Misunderstanding	Open	Minor			2023-02-18 16:30:25.048
3	en-ro_covid19_ant	191	Accuracy > Misunderstanding	Open	Minor	medicii rezidenți	rezidenți într-o unitate de îngrijire pe termen lung	2023-02-18 16:33:53.318
3	en-ro_covid19_ant	196	Accuracy > Ambiguous	Open	Minor			2023-02-18 16:35:38.513
3	en-ro_covid19_ant	196	Terminology > Inconsistent	Open	Minor	locuitorii		2023-02-18 16:35:21.451
3	en-ro_covid19_ant	198	Language > Grammar	Open	Minor			2023-02-18 16:36:09.605
3	en-ro_covid19_ant	199	Accuracy > Addition	Open	Minor	unitatea dumneavoastră sanitară		2023-02-18 16:37:03.461

Figure 1: The Quality Assessment report

It should be mentioned that columns that were not relevant for this study were hidden. The most important columns in the activity report generated from Trados Studio can be seen in the table above. These criteria include the activity ID, which denotes the number of the session in which the annotations were carried out, the activity name, which is the name of the file that was analysed in Trados Studio, and the segment ID. Furthermore, the table contains information about the name of the quality metric, its status, severity, and the specific content to which the annotation was made, including the comments made by the user. The comments that were made in this case were mostly improvement suggestions for the errors that were encountered during the analysis.

The encountered errors are illustrated in the following table and are sorted according to their name and the number of encountered errors of their type:

Table 1: Quality metrics – types of encountered errors

Type of error	**Number of encountered errors**
Accuracy > Addition	61
Accuracy > Ambiguous	20
Accuracy > Improper (exact)	16
Accuracy > Mistranslation	93
Accuracy > Misunderstanding	3
Accuracy > Omission	34
Accuracy > Untranslated	7
Language > Grammar	11
Language > Punctuation	7
Language > Spelling	18
Others > Repeat	2
Style > Grammar	14
Style > Overly literal	3
Style > Unidiomatic	1
Terminology > Inconsistent	29

Based on the data provided, there was a total of 319 errors encountered in the 1034 corpus segments, which can be classified into different categories as follows: accuracy errors which are errors that are related to the precision and correctness of the translated text. In total, 234 errors of this type were identified during this experiment, which include additions, ambiguities, improperness errors, mistranslations, misunderstandings, omissions, and untranslated parts. As per the explanations provided in the Qualitivity plug-in in Trados Studio, additions refer to unnecessary elements that are not present in the source text but are included in the translation. Ambiguity-related errors occur when the translation of a clear source segment is uncertain or vague, while improperness errors occur when a segment is poorly translated or not appropriate for the context. Mistranslations can arise from an incorrect interpretation of the source text, whereas misunderstandings can occur when technical concepts are misconstrued. An omission occurs when essential elements of the source text are absent in the translation. However, when the entire source text is not translated, this type of error can be categorized as "untranslated", according to the TAUS DQF standard.

Language errors are errors related to the language used in the translated text and include grammar, punctuation and spelling mistakes. A total of 36 language errors were identified, with grammar errors typically occurring due to noncompliance with the rules of the target language in terms of syntax. Similarly, punctuation errors result from noncompliance with the target language rules or the style guide in terms of punctuation. Spelling errors can also arise from incorrect usage of capital and lowercase letters.

A total of 2 repetition errors were identified, referring to errors that were encountered multiple times throughout the translation. Style errors are errors related to the style and tone of the translated text and include grammar mistakes, instances of overly literal translations, as well as unidiomatic translations. A total of 18 such style errors were identified in this case. Grammar errors that fall under the style category usually arise due to awkward syntax in the translated text. Style errors also include overly literal translations, labelled as "overly literal", as well as unidiomatic use of the target language, labelled as "unidiomatic".

Lastly, terminology errors are errors related to inconsistency errors in terms of the terminology used in the translated text. A total of 29 such terminology errors were identified during the process of analysis.

To assess the alignment accuracy of the selected bilingual corpora that have been automatically aligned and to evaluate the consistency of the terminology used in translation, it is essential to examine instances of some of the most significant error categories. Among these categories, addition errors are some of the most commonly observed. The table below illustrates four relevant examples belonging to this error category.

Table 2: Examples of additions

	Source text (EN)	**Target text (RO)**	**EN translation of the highlighted text**
Accuracy > Addition	Antimicrobial resistance is the ability of a microorganism (e.g., a bacterium, a virus) to resist the action of an antimicrobial agent.	Rezistenţa antimicrobiană este capacitatea unui microorganism (de exemplu o bacterie, un virus **sau un parazit, precum parazitul malariei*) de a rezista acţiunii unui agent antimicrobian.	* or a parasite, like the malaria parasite
	A factsheet on antibiotic resistance in primary care provides primary care prescribers with EU and national data on the latest trends.	O fişă informativă despre rezistenţa la antibiotice în asistenţa medicală primară le oferă medicilor date naţionale şi la nivelul UE privind cele mai recente tendinţe **utile atunci când prescriu antibiotice.*	* useful when prescribing antibiotics
	76% of hospitals reported the availability of antimicrobial use guidelines.	76 % din spitale au raportat disponibilitatea orientărilor privind utilizarea antimicrobienelor, **iar 54 % au raportat că personalul*	* and 54% reported that staff have a certain amount of time dedicated to promoting the

		dispune o anumită perioadă de timp dedicată acțiunilor de promovare a utilizării adecvate a antimicrobienelor.	appropriate use of antimicrobials.
	ECDC is looking forward to collaborating with stakeholders at pan-European level and welcomes dialogues with organizations throughout the world running awareness campaigns on prudent use of antibiotics.	ECDC intenționează să colaboreze cu părțile interesate și cu organizațiile care susțin sau derulează campanii de informare privind utilizarea prudentă a antibioticelor **în domeniul sănătății oamenilor și animalelor.*	*in human and animal health.

The table presented above contains four columns, one which illustrates the error type – addition in this case -, one for the segment in the source language (EN), one for the segment in the target language (RO) and a gloss of what the highlighted text means in English so that the readers who do not speak Romanian can get the same idea. In the column related to the target text, the highlighted parts aim to demonstrate the elements that were added to the translation but are, however, absent from the source text. Whenever additions occur, one characteristic that is always noticeable is the length of the segment in the target language, which is visibly longer in comparison to the segment in the source language. During the analysis of the automatically-aligned bilingual corpora, a total of 61 addition errors were detected.

Mistakes associated with ambiguity do not necessarily imply that the translation is incorrect; they merely suggest that the phrasing chosen by the translator may not be the most suitable and can lead to confusion among readers. Here are some examples of translation errors related to ambiguity that were identified:

Table 3: Examples of errors related to ambiguity

	Source text (EN)	Target text (RO)	EN translation of the highlighted text
Accuracy > Ambiguous	In France and Belgium, the decrease was attributed to national action including a yearly, nationwide public campaign on the prudent use of antibiotics;	În Franţa şi Belgia, reducerea a fost atribuită măsurilor naţionale **şi anume o campanie anuală*, naţională, publică privind utilizarea prudentă a antibioticelor;	*and namely an annual campaign
	Follow us on Twitter, use #EAAD or take part in our twitter chats on EAAD: http://www.twitter.com/EAAD_EU	Urmăriți-ne pe Twitter, **trimiteți mesaje la #EAAD* sau participați la discuțiile noastre despre EAAD: http://www.twitter.com/EAAD_EU	* send messages to #EAAD
	Organise and promote educational events, courses and meetings together with hospital administrators to strengthen infection prevention and control activities among all healthcare professionals (e.g. hand hygiene, contact precautions, active screening cultures, and environmental cleaning)	Organizați și promovați evenimente educaționale, cursuri și întâlniri împreună cu administratorii de spital pentru a consolida activitățile de prevenire și control al infecțiilor în rândul tuturor profesioniștilor din domeniul sănătății (de exemplu, igiena mâinilor, precauții de contact, depistarea activa prin screening a germenilor cu risc epidemiologic crescut și **curățarea mediului*)	*cleaning of the environment

The initial translation could benefit from increased clarity by rephrasing to: "În Franţa şi Belgia, reducerea a fost atribuită măsurilor naţionale, printre care se numără și o companie anuală, naţională, publică privind utilizarea prudentă a antibioticelor; ". A suggested revision for the second example could be: "Urmăriți-ne pe Twitter, folosiți #EAAD sau participați la discuțiile noastre despre EAAD pe Twitter: http://www.twitter.com/EAAD_EU". In the third example, "curățare" is a general word and could lead to misinterpretations in this context. Therefore, "decontaminarea mediului" would have been a better choice of words in the medical context.

Improperness errors can occur when a segment is poorly translated or not appropriate for the context. This can happen when certain terms are not translated with their corresponding specialised term or when a more contextually appropriate translation is available, but not used in the translation, as illustrated in the examples below:

Table 4: Examples of errors related to improperness

	Source text (EN)	**Target text (RO)**
Accuracy > Improper (exact)	These positive experiences from some EU Member States are the background for the European Antibiotic Awareness Day, a campaign to reduce use of antibiotics in situations where they are not necessary, for example for viral infections such as colds and influenza.	Aceste experienţe pozitive ale unor state membre UE sunt fundamentul pentru Ziua Europeană a Informării despre Antibiotice, o campanie pentru reducerea utilizării antibioticelor atunci când nu sunt necesare, de exemplu în infecţii virale ca răcelile şi *gripa*.
	The EARSS Annual Report 2007, as well as national data, indicated decreasing resistance trends in Streptococcus pneumoniae, a bacteria commonly responsible for infections in outpatients, in particular children;	Raportul Anual 2007 EARSS, precum şi datele naţionale, au indicat tendinţe de scădere a rezistenţei pentru Streptococcus pneumoniae, bacteria responsabilă de

		infecţii la pacienţii din *ambulator*, în special la copii;
	Surveillance data show that antimicrobial resistance is a growing public health problem in European hospitals and communities.	Datele rezultate din supraveghere arată că rezistenţa antimicrobiană este o problemă de sănătatea publică în creştere în spitalele europene şi **mediul extraspitalicesc*. * out-of-hospital environment

In the first example, the more appropriate translation in the given context would be "influența", the Romanian term corresponding to "influenza" from the source text. The translator, however, opted to translate this term as "gripă", which is a broader, more common term. "Influența" is a subcategory of this term and, therefore, a more precise translation, more appropriate for the medical field. The second example illustrates an instance where the highlighted word is wrongly used, as "ambulator" does not exist in the Romanian language. In the last example, the word "communities" from the source text is improperly translated as "mediul extraspitalicesc". A better choice of words would have been "comunități" in this case. This is however a subjective remark, as the translator could have opted for this wording to highlight even more the difference between hospital and community, or out-of-hospital settings. It is evident that the translation in question is rather a free translation, not necessarily adhering to the source text, albeit still effectively conveying the intended message.

The table below exemplifies some of the mistranslations encountered in the analysed text. It should be mentioned though, that most of the mistranslations identified are, in fact, the result of misalignments.

Table 5: Examples of mistranslations

	Source text (EN)	**Target text (RO)**	**EN translation of the highlighted text**
	Key messages for hospital prescribers.	Mesaje-cheie pentru **farmaciștii din spital.*	* hospital pharmacists
Accuracy > Mistranslation	While people die in developing countries because they lack access to the correct antimicrobial treatment, antimicrobial resistance resulting from inappropriate use is causing concern in every continent;	Dacă în ţările în curs de dezvoltare încă mor oameni din cauza lipsei de acces la tratamente cu antibiotice corespunzătoare, **rezistenţa la antibiotice* rezultată din utilizarea incorectă a acestora devine o sursă de îngrijorare pe toate continentele;	* antibiotic resistance
	Page 1 of 5 // 50 Results found	Pagina **1 din 4 // 40 Rezultate găsite*	*1 of 4 // 40 Results found

The aforementioned examples exhibit different degrees of mistranslations. The first and second examples are clear instances of mistranslations. According to IATE[18], the multilingual terminology database created and managed by the institutions of the European Union, the appropriate translation for "antimicrobial resistance" into Romanian would be "rezistență la antimicrobiene" and that of "hospital prescribers" would be "medici prescriptori din spital". With respect to the third example, it is, however, uncertain whether this is a misalignment or an adaptation to the layout and number of pages required for the translation. Therefore, it may not necessarily be a mistranslation.

[18] https://iate.europa.eu/home

Another type of error commonly identified throughout the analysis of the translation is omission. The following table presents examples of segments from the source text that were not included in the translation.

Table 6: Examples of omissions

	Source text (EN)	**Target text (RO)**	**EN translation of the highlighted text**
Accuracy > Omission	A total of 8.9 million HAIs were estimated to occur each year in *European* hospitals and long-term care facilities.	Se estimează că, anual, se înregistrează 8,9 milioane de cazuri de IAAM-uri în spitale și în unitățile de îngrijiri pe termen lung.	A total of 8.9 million HAIs were estimated to occur each year in hospitals and long-term care facilities.
	Overview of the data on antimicrobial resistance in Europe released in 2014	Rezistența antimicrobiană în Europa: 2014	Antimicrobial resistance in Europe released in 2014
	ECDC is looking forward to collaborating with stakeholders at *pan-European level* and welcomes dialogues with organizations throughout the world running awareness campaigns on prudent use of antibiotics.	ECDC intenționează să colaboreze cu părțile interesate și cu organizațiile care susțin sau derulează campanii de informare privind utilizarea prudentă a antibioticelor în domeniul sănătății oamenilor și animalelor.	ECDC is looking forward to collaborating with stakeholders and welcomes dialogues with organizations throughout the world running awareness campaigns on

			prudent use of antibiotics.
	if you have lung disease (e.g. chronic bronchitis, emphysema, chronic obstructive pulmonary disease);	dacă aveți o boală de plămâni (de exemplu, bronșită, emfizem, bronhopneumopatie obstructivă cronică);	if you have lung disease (e.g. bronchitis, emphysema, chronic obstructive pulmonary disease);

The table above shows instances of words in the source text were not included in their corresponding translation into Romanian. Such omissions can lead to misunderstandings and misinterpretations. For instance, the absence of the term "European" in the first example could lead the reader to believe that the reported number of 8.9 million cases of HAIs is a global figure.

Another type of error that was identified in the given translation is represented by inconsistencies in the use of terminology. This error category is, in fact, the starting point for answering one of the research questions addressed by this study, regarding the consistency of EN-RO Covid-19 terminology used in the chosen bilingual corpora. The small number of terminology inconsistencies that were identified during the analysis, namely 29, indicated the fact that the medical terminology used in the translation of the chosen text is mostly consistent. The following examples, taken from the analysed corpus, show instances of inconsistencies in the translation of the same English terms.

Table 7: Examples of inconsistencies

	Source text (EN)	**Target text (RO)**
	Prudent use of *antimicrobials* (i.e. only when needed, with the correct dose, at correct	Utilizarea prudentă a *antibioticelor* (adică doar atunci când este necesar, în doză corectă, la intervale de

	dose intervals and for a correct duration);	dozare corecte şi pentru o durată corectă);
	Antimicrobials are medicinal products that kill or stop the growth of living microorganisms and include among others:	*Antimicrobienele* sunt produse medicamentoase care omoară sau împiedică creşterea microorganismelor şi cuprind printre altele:
Terminology > Inconsistent	Since October 2008, it is possible for patients in the UK who are *asymptomatic* but who have a diagnosed genital infection with Chlamydia to obtain from pharmacists (without a prescription) a single dose of the antibiotic azithromycin, representing a complete treatment course.	Începând cu octombrie 2008, pacienţii din Regatul Unit care au o infecţie genitală cu Chlamydia diagnosticată, dar *fără simptome*, pot obţine de la farmacişti (fără reţetă) o singură doză de azitromicină, care reprezintă un ciclu complet de tratament.
	Avoiding diagnostic tests and cultures in *asymptomatic* patients	Evitarea testelor de diagnosticare şi a culturilor la pacienţii *asimptomatici*
	E. coli causes *urinary tract* and more serious infections and is one of the most common bacteria causing infections.	E. coli cauzează *infecţii ale căilor urinare* şi alte infecţii mai grave şi este una dintre cele mai des întâlnite bacterii care cauzează infecţii.
	On arrival at the mainland, he immediately went to a large hospital in Rome where he had a urine culture and clinical examination, which	După ce a sosit pe continent, s-a dus imediat la un spital mare din Roma, în care i s-a efectuat urocultura şi examenul clinic, acestea

	confirmed he had a complicated *urinary tract infection*.	confirmând că prezenta o *infecţie complicată a tractului urinar*.

Various other error types were identified in the translation as well, including spelling mistakes, incorrect punctuation, and grammar errors. For example, the name of the country "Cipru" was articulated with the definite article as "Ciprul". Another instance of grammar mistake encountered during the analysis is "împotriva malariei şi a altor infecţii datorate paraziţilor". The preposition "datorate" is used here incorrectly and should be replaced by "cauzate de". Additionally, some of the segments were translated overly literal and contained syntax that is awkward for the Romanian language. For instance, "Prioritise antibiotic stewardship and infection prevention and control policies" was translated into Romanian as "Stabiliți prioritatea politicilor privind prevenirea și controlul infecțiilor și gestionarea antibioticelor". A more appropriate choice of words would have been "Prioritizați politicile privind prevenirea și controlul infecțiilor și gestionarea antibioticelor".

Furthermore, some of the translated segments contained repetitions of the same errors, such as the word "ambulator", which was also mentioned under the category of "improper" translations. The correct Romanian word is, in fact, "ambulatoriu". These errors could lead to misunderstandings and misinterpretations.

6.2. Corpus alignment evaluation

In order to address the first research question concerning the alignment accuracy of the selected bilingual corpora, the number of misaligned segments had to first be determined out of a total of 1034 segments. Therefore, apart from the errors listed above, that were pointed out by using the quality metrics in the Qulitivity plug-in, the analysis also included the identification instances of misalignments, which occur when the translation of a specific unit of text, such as a sentence or phrase, does not accurately convey the intended meaning of the original text. These errors were marked as "mistranslations" and in order to indicate these mistranslations, the status of each misaligned segment was changed in Trados Studio to "translation rejected". Examples of mistranslations encountered during the analysis are presented in the table below.

Table 8: Examples of misalignments

	Source text (EN)	**Target text (RO)**	**EN translation of the target text**
Mistranslations / Misalignments	The main objective of EPA-UNEPSA is to encourage scientific co-operation between not-for-profit National European Paediatric Societies/Associations and between European paediatricians working in primary, secondary and tertiary paediatric care in Europe, in order to promote child health and comprehensive paediatric care.	ECDC colaborează îndeaproape cu Comisia Europeană și cu alte agenții ale UE, cu părțile interesate și cu organizațiile care desfășoară campanii de sensibilizare cu privire la utilizarea prudentă a antibioticelor.	ECDC works closely with the European Commission and other EU agencies, stakeholders and organisations campaigning to raise awareness on the prudent use of antibiotics.
	The poster contains a key message that is deemed relevant overall for the hospital or the healthcare setting, followed by a set of actions that the hospital prescriber can/should/must implement to ensure that antibiotics remain effective.	Lista de verificare este un mic card care poare fi ţinut în buzunarul halatului şi poate fi consultat ca memento privind prescrierea prudentă a antibioticelor.	The checklist is a small card that can be kept in your coat pocket and can be consulted as a reminder on prudent antibiotic prescribing.

	Things to do to keep antibiotics working.	Trebuie să vă îngrijiți, nu să luați antibiotice	You need to take care of yourself, not antibiotics
	Documenting indication, drug choice, dose, route of administration and duration of treatment in the patient chart leads to better use of antibiotics.	O prezentare care poate fi utilizată pentru sesiuni de formare în spitale, în care sunt detaliate principalele aspecte ale rezistenţei la antibiotice şi modul în care poate fi îmbunătăţită utilizarea antibioticelor.	A presentation that can be used for training sessions in hospitals, detailing the main aspects of antibiotic resistance and how antibiotic use can be improved.

Moreover, it should be noted that the order of the segments in the chosen bilingual corpora was entirely aleatory, lacking logical sequence. The accuracy of the text segmentation was partly compromised, leading to the identification of 100 misaligned segments out of the 1034 total segments. To determine the alignment accuracy of the selected bilingual corpora, it was necessary to employ a performance metric that assesses the proportion of correctly classified instances within a dataset, commonly known as accuracy. This was calculated by utilizing the formula of correctly predicted data points (934) divided by the total number of data points (1034), multiplied by 100. As a result, the alignment accuracy of the chosen corpora was found to be 90.32%. The high percentage thus obtained indicated that the automatic alignment was greatly successful.

6.3. Automatic term extraction evaluation

This study focuses not only on the alignment accuracy of the chosen bilingual corpora, but also on the accuracy of the automatic term extraction, from the chosen bilingual corpora, in contrast to the extraction performed manually. For this experiment, the automatic term extraction was performed using the Sketch Engine, while the manual term extraction was performed by analysing the text and manually adding any term that was thought to be relevant for the medical field to a beforehand created term base.

As the bilingual term extraction feature was not available in the Sketch Engine for the combination of languages English-Romanian, the term extraction was performed separately for the English source text, as well as for the Romanian translation. As explained in the previous chapter, the lists of extracted term candidates were then merged into single Excel files, one for the English terms, containing both single-word term candidates as well as multiword term candidates, and one for the Romanian term candidates. From the total term candidates produced by the Sketch Engine, only a subset of term candidates, namely 1200 English term candidates and 500 Romanian term candidates were then further compared to the terms that had already been extracted manually and added to the term base. It has to be mentioned, though, that for the purpose of this study, only the automatically extracted term candidates that were not overlapping with the ones added manually to the term base were analysed. The term base comprises of a total of 355 terms that were manually extracted, while the Sketch Engine yielded a total of 2000 term candidates per language, leading to a general total of 4000 automatically extracted term candidates, for both English and Romanian. For the part of this experiment dealing with the English term candidates, out of the 2000 automatically extracted term candidates, only 210 overlapped with the manually extracted terms (total of 355 terms), leaving a total of 1790 term candidates that were proposed by the Sketch Engine, but were left out during the manual term extraction. The overlapping terms are to be found scattered all over the list of term candidates proposed by the Sketch Engine, although, they are by far more frequent in the first half of the list. The exact numbers were illustrated below for a better understanding as follows:

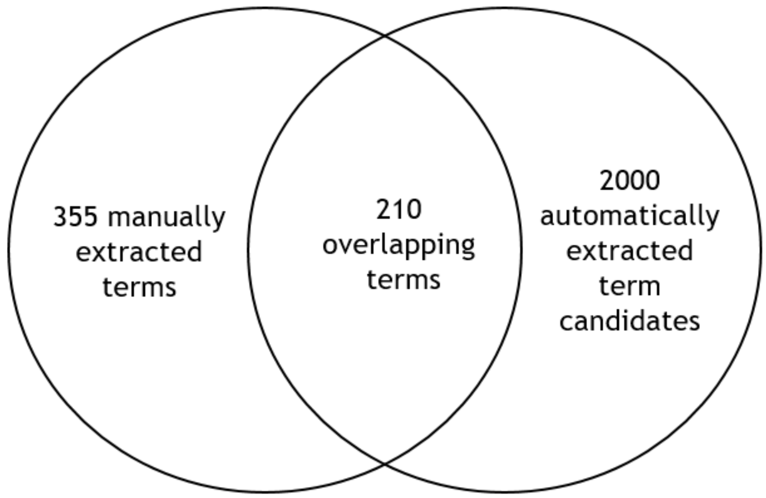

Figure 2: Overlapping English terms

After eliminating the overlapping terms, it was observed that the number of terms proposed by the Sketch Engine added up to 1790 terms (these were not included in the manually created term base). Out of these 1790 potential terms, only the first 1200, arranged alphabetically, were chosen for further analysis. The automatically extracted terms were classified into five categories: "not a term", "not a medical term", " similar to term already in the term base", "wrong", and "overlooked during the manual term extraction". For a better understanding of the results, the following pie chart was created with the exact percentages attributed to each category.

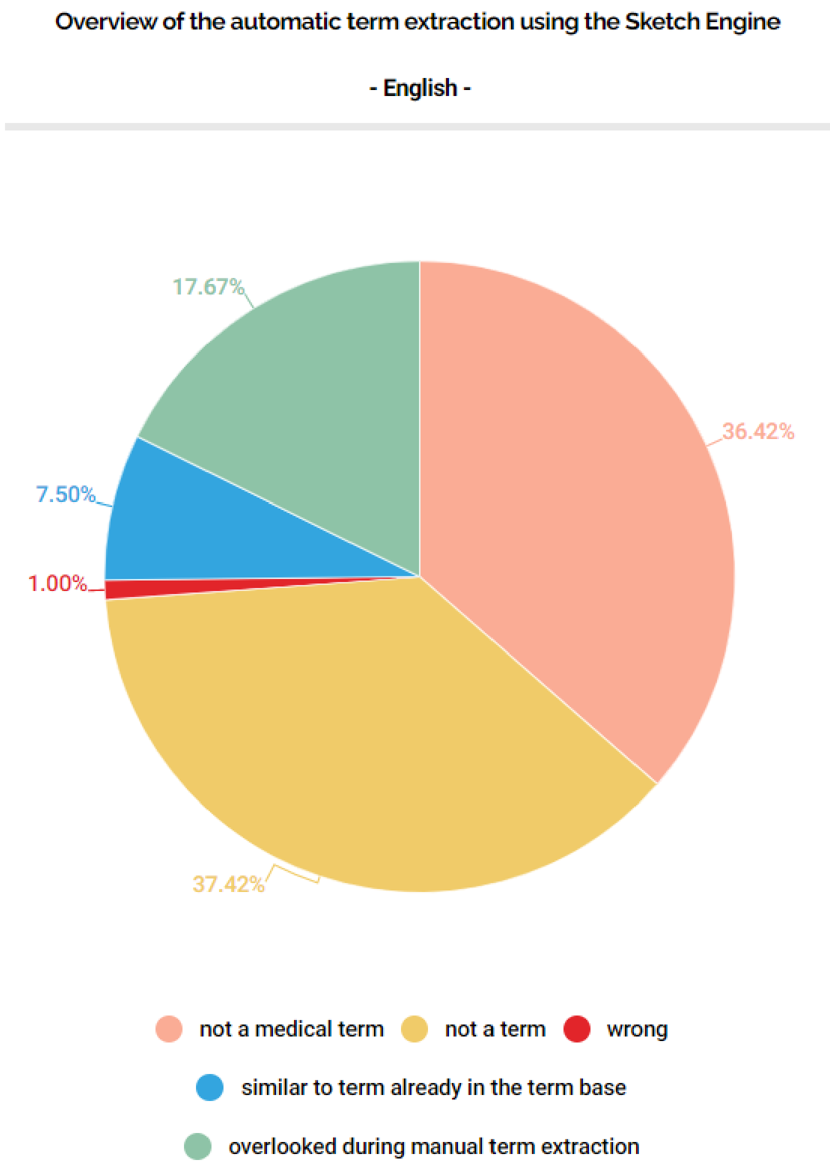

Figure 3: Pie chart representation of the categories identified during the analysis of the automatically extracted English term candidates

As it can be observed from the pie chart above, the highest percentage, 37.42 % (449), was obtained by the category "not a term", which is marked yellow on the pie chart. In this category were included all words extracted by the Sketch Engine and presented as potential terms by the software. These, however, did not fulfill the necessary criteria and did not qualify as terms. Here are some examples: "account local antimicrobial susceptibility", "antibiotic consumption in many hospitals", "antibiotic in such settings", "antibiotic use in the emergency", "antibiotic with other people", "decrease in both consumption", "example for viral infections", "example of such bacteria", "given day", "measure for antibiotic prescriptions" etc.

The second highest percentage, 36.42% (437), was attributed to the category "not a medical term" which is marked orange on the pie chart. This category encompasses all words automatically extracted by the Sketch Engine, that fulfill the necessary criteria to qualify as terms. However, these terms do not necessarily belong to the medical field, as the name of the category suggests. The following terms are examples belonging to this category: "adaptation", "adherence", "authority", "awareness campaign", "brother-in-law", "campaign organiser", "characteristic", "core task", "decontamination of the environment", " area of specialisation" etc.

The third category, "overlooked during the manual term extraction" represents 17.67% (212) of the total number of 1200 automatically extracted term candidates that were further analysed in this study. This category, coloured green on the pie chart, includes the terms that were overlooked during the manual term extraction and are, thus, not part of the term base. These are terms, that the Sketch Engine identified correctly and that, in fact, belong to the medical field. Here are some examples: "allergic", "alleviate", "health threat", "health authority", "hospital infection", "hospital infectious disease specialist", "incidence of healthcare-associated infections", "infection acquired in hospital", "infection cure rate", "intensive care unit".

As it can be observed in the screenshot below, which compares the English term candidates proposed by the Sketch Engine in their original order, based on their general score, with the terms in the category "overlooked during manual term extraction", most of these terms are not at the top of the list (at least among the first 40 term candidates), so are not among the term candidates with the highest scores. The "Duplicate Value" feature in Excel was utilized to identify terms that are overlapping, between the initial two categories proposed by the Sketch Engine, namely single-word and multi-word terms, and the third category which comprises terms proposed by the Sketch Engine but were omitted during manual term extraction.

1	THE SKETCH ENGINE - single-words	THE SKETCH ENGINE - multi-word terms	Overlooked during manual term extraction
2	antibiotic	antibiotic use	acid
3	ecdc	antibiotic stewardship	acinetobacter
4	prescriber	antibiotic resistance	acquire
5	healthcare-associated	expert consensus	acquired resistance
6	antimicrobial	infection prevention	action of an antibiotic
7	antibiotic-resistant	healthcare-associated infection	action of an antimicrobial agent
8	eaad	use of antibiotics	acute care
9	multidrug-resistant	antibiotic treatment	acute care hospital
10	carbapenem	antimicrobial resistance	adaptation of the microorganism
11	prophylaxis	stewardship programme	adequate infection prevention
12	mrsa	antibiotic stewardship programme	adverse effect of antibiotic therapy
13	prudent	key message	allergic
14	antimicrobial-resistant	prudent antibiotic use	alleviate
15	stewardship	antibiotic awareness	alternative antibiotic therapy
16	self-medication	antibiotic stewardship team	antibacterial agent
17	microorganism	toolkit material	antibiotic
18	last-line	european antibiotic awareness	antibiotic consumption in a population
19	esbl	stewardship team	antibiotic consumption in humans
20	factsheet	antibiotic-resistant bacterium	antibiotic cost
21	fluoroquinolone	resistant bacterium	antibiotic dose
22	klebsiella	prudent use	antibiotic dose for inpatients
23	coli	information material	antibiotic for a previous illness
24	microbiologist	healthcare setting	antibiotic for urinary tract infections
25	infection	antibiotic therapy	antibiotic formulary
26	infographic	hospital prescriber	antibiotic guideline for common infections
27	toolkit	antibiotic guideline	antibiotic prescribing practice
28	bacterium	prudent use of antibiotics	antibiotic resistance in hospital settings
29	resistant	european antibiotic awareness day	antibiotic resistance in primary care
30	bloodstream	antibiotic awareness day	antibiotic resistance in the hospital
31	escherichia	antibiotic consumption	antibiotic resistance in the microbiota
32	pneumoniae	primary care prescriber	antibiotic resistance pattern
33	urinary	care prescriber	antibiotic resistance prevalence rate
34	staphylococcus	awareness day	antibiotic resistance rate
35	formulary	antibiotic prescribe	antibiotic resistance-bacterium combination
36	food-producing	infectious disease specialist	antibiotic stewardship in healthcare settings
37	broad-spectrum	other healthcare setting	antibiotic stewardship measure
38	aureus	urinary tract infection	antibiotic stock
39	evidence-based	disease specialist	antibiotic therapy for individual patients
40	cephalosporin	long-term care facility	antibiotic treatment duration

Figure 4: Comparison between the term candidates proposed by the Sketch Engine in their original order, according to their score, and the terms in the category "overlooked during manual term extraction"

Among the automatically extracted analysed in this study there were also a series of terms that even though had not been included in the term base in the form proposed by the Sketch Engine, were very similar to those already in the term base, most of them being derived words that differed in terms of the grammatical number (singular or plural) or in the number of words embedded in the term. Therefore, these terms cannot be considered as having been overlooked. On the pie chart, this category is marked blue and comprises of 7.5% (90) of the analysed terms. To better illustrate this the table below was created and contains examples of terms that were included in the term base compared to the derived versions found by the Sketch Engine.

Table 9: Examples from the category "similar to term already in the term base" (English)

From the term base	**From the Sketch Engine**
antibiotic resistance	antibiotic-resistant
antimicrobial resistance	antimicrobial-resistant
asymptomatic patient	asymptomatic
hospital stay	hospital stays
invasive procedure	invasive

Based on the aforementioned examples, it can be discerned that some of the terms identified by the Sketch Engine and further analysed in this study were very similar to those that had been manually identified and subsequently incorporated into the term base.

The final category, coloured red on the pie chart, encompasses all words that were erroneously identified by the Sketch Engine. This includes words that could not be located within the source text or were identified in a form that did not correspond to that of the source text. Additionally, the Sketch Engine sometimes proposed terms that were grammatically incorrect, such as when "belgium" was written in lowercase. Another instance is the proposed term "healthcare set", which could not be found in the source text. The only similar term that was identified in the source text was "healthcare setting", indicating that the term proposed by the Sketch Engine either omitted or truncated a part of the original term. This category represents a mere fraction of the total of analysed terms, amounting to only 1% (12).

Regarding the Romanian part of the experiment, only 96 term candidates proposed by the Sketch Engine overlapped with those that had manually been extracted, as demonstrated below:

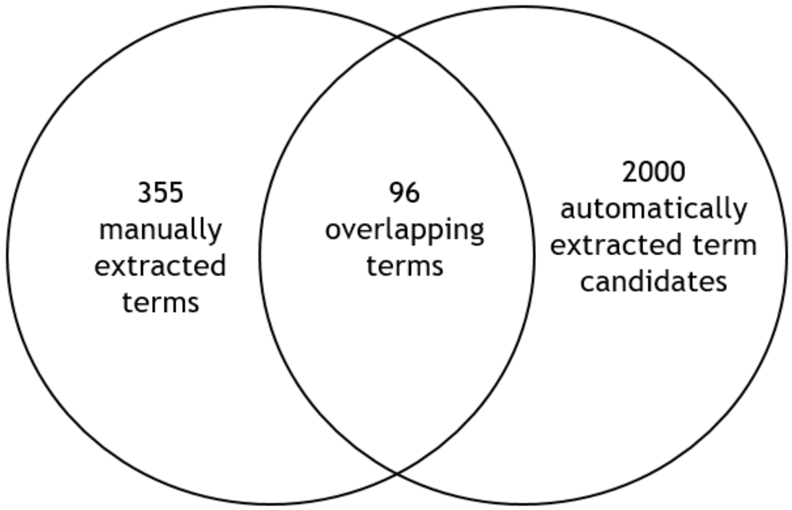

Figure 5: Overlapping Romanian terms

Again, while the term base comprises of a total of 355 terms and their translations, which were manually extracted from the chosen bilingual corpora, the Sketch Engine, on the other hand proposed a total of 2000 Romanian term candidates, which is the maximum number of term candidates allowed by the Sketch Engine. After removing the overlapping terms (96), a total of 1904 term candidates remained. Out of these 1904 term candidates, the first 500 were selected, in alphabetical order, for further analysis. The following pie chart presents the same categories previously illustrated, applied in this case to the Romanian automatically extracted term candidates.

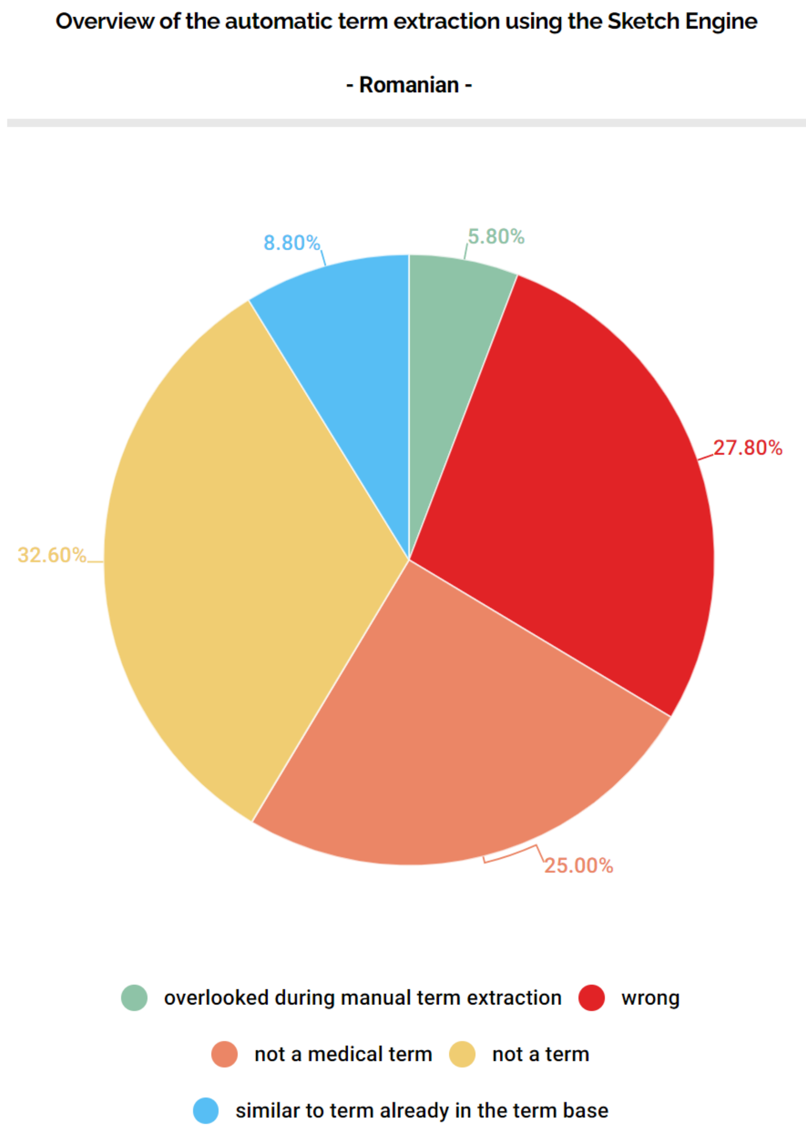

Figure 6: Pie chart representation of the categories identified during the analysis of the automatically extracted Romanian term candidates

Upon initial inspection, the percentage differences between the present pie chart and the preceding one appear to be unnoticeable, at least in the case of the "not a term", "not a medical term" and "similar to term already in the term base" categories, where the percentages are very similar. Nevertheless, a clear contrast can be observed between the percentages of the "wrong" categories from this and the previous pie chart representation. However, it should be kept in mind, that the number of Romanian term candidates that were further analysed for the purpose of this experiment is also much lower than that of the English term candidates analysed in this research.

Following are some instances of term candidates that fall under the category "wrong": "acest bacterie rezistent", "adecvat al antibiotic", "al antibiotic reduce eficacitate", "al piele și țesut", "al utilizare antibiotic", "antibiotic sine număra următoarea", "antibiotic-resistant", "apariție bacterie rezistent", "attributable", "boală infecțios din spital". The majority of these exemplars are composed of word pairings recommended by the Sketch Engine as terms. Nonetheless, these word amalgamations do not satisfy the necessary criteria to warrant their classification as terms. Rather, they appear to be a mere assemblage of words, lacking any cohesive or "unithood" properties, at first sight. As an example, the word pairing "boală infecțios" is syntactically incorrect, as "boală" is a feminine noun while "infecțios" is a masculine adjective. The appropriate pairing of words to form a valid term would have been "boală infecțioasă". However, upon closer examination, it becomes apparent that the Sketch Engine software applies lemmatization to all multi-word term candidates extracted automatically from the Romanian corpora.

It is also noteworthy that some of the term candidates automatically extracted by the Sketch Engine for the Romanian part of this experiment were, in fact, in English: "against", "attributable", "analysis". These were also placed under the category "wrong".

The proportion of the "not a term" category is comparable to that of the previous pie chart which presented the outcomes of the automatic extraction of English terms; as mentioned before, this category refers to word combinations that, even though are correct from a grammatical point of view, do not constitute an actual term. e.g.: "atunci când prescrie", "bacterian și nu", "astfel de bacterie", "antibiotic și aspect important", "boală care apărea iarna", "care prescrie antibiotic pentru", "coleg mai experimentat", "comun și pentru profilaxie", "control al bolilor", "cu privire la utilizare".

The percentage of the "not a medical term" category is lower for the Romanian term candidates than for the English ones. Accordingly, 25% (125) of the total of 500 analysed term candidates, which were selected automatically by the Sketch Engine, fall outside the medical domain. To the category "not a medical term" belong terms such as: "cumnat", "curățare", "călătorie", "campanie", "barcă", "axă", "aspect important legat de", " cât mai curând posibil", "alternativă", "alimentară".

The percentage of the category "similar to term already in the term base" is comparable to that of the same category from the pie chart illustrating the English terms. This category was implemented to mark those terms that were extracted by the Sketch Engine, but were already present in the term base, albeit in the form of derived words or word combinations that differed in terms of the grammatical number (singular or plural) or in the number of words embedded in the term. Examples can be found in the table below:

Table 10: Examples from the category "similar to term already in the term base" (Romanian)

From the term base	**From the Sketch Engine**
angină pectorală	angină
antibiotic	antibiotice
medicament antiviral	antiviral
asimptomatic	asimptomatici
aztreonam	aztreonamul

The proportion of Romanian terms that were "overlooked during manual term extraction" is smaller than that in the pie chart representing English terms. Among the Romanian terms extracted automatically, this category only constitutes 5.8% (29) of the total number of analysed terms (500). Following are some examples of terms that belong to this category: "bacterian", "anaerob", "antibiotic pentru infecție", "carbapenemaza", "cardiac".

After analysing both pie charts, it can be concluded that Sketch Engine's automatic term extraction has higher accuracy for English terms than for Romanian terms. The "wrong" category for English term candidates was only 1% (12), while the same category representing Romanian term candidates scored 27.8% (139). The size of the samples should, however, be kept in mind, since the number of English term candidates analysed in this study was considerably higher than that of Romanian term candidates. Additionally, the first pie chart shows that the number of terms overlooked during manual extraction was higher among the English term candidates.

However, the categories "not a term", "not a medical term", and "similar to term already in the term base" were similar in both charts.

The main objective of this study was to evaluate the effectiveness of manual and automatic term extraction methods that had been used for identifying relevant medical terms in bilingual corpora. Upon comparing the results of both methods, it was observed that the Sketch Engine outperformed the manual process in terms of the number of words extracted. However, not all automatically extracted words were identified as terms. To determine the effectiveness of the two methods, it is necessary to analyse the "overlooked during manual term extraction" category, which is one of the accuracy indicators of the automatic term extraction performed by the Sketch Engine.

In the section of the experiment that concentrates on the English term candidates proposed by the Sketch engine, the "overlooked during manual term extraction" category accounted for 17.67% (212) out of 1200 analysed potential terms. This means that a total of 212 valid terms failed to be identified during the manual extraction process. If added to the terms correctly identified by the Sketch Engine which also overlap with the ones identified manually (210), a total of 422 correctly identified terms is obtained. This is a higher number than the gold standard of 355 terms. In other words, the automatic term extraction method was able to identify more relevant medical terms than the gold standard, indicating its effectiveness in identifying important terms.

Regarding the automatically extracted Romanian term candidates, only 29 out of 500 analysed potential terms fell under the category mentioned above. Additionally, there were 96 terms overlapping with the manually extracted ones, resulting in a total of 126 correctly identified terms.

The number of correctly identified terms for English was higher than for Romanian, but so was the sample of analysed English term candidates. Therefore, this information should be regarded with caution. Moreover, many of the automatically identified terms, especially those consisting of several words, cannot be utilized further as they have undergone a lemmatization process by the Sketch Engine, resulting in their appearance in dictionary form. This limitation may affect the usefulness of the automatically extracted terms in Romanian.

Although manual terminology extraction is generally considered more precise and accurate, it can be time-consuming and subject to human bias. Automatic terminology extraction, on the other hand, relies on computer-based methods to quickly identify term candidates but may overlook the nuances of language or specific contextual information of the terms. Therefore, the Sketch Engine may be a very good, helpful tool when it comes to extracting English terminology, but the term candidates proposed by the software still require the confirmation and validation of a domain expert, as computers struggle with the complexities of language and meaning (Kris Heylen & Dirk De Hertog 2015).

6.4. Precision and Recall of the automatic term extraction

The third research question of this study, which aims to determine the precision and recall of the automatic term extraction tool against the gold standard produced through manual human terminology extraction, was answered based on all the collected data and information. As explained in the chapter regarding automatic term extraction, precision refers to the number of correctly identified terms from the list of term candidates, while recall, on the other hand, measures the proportion of terms identified out of all valid terms present in the specialised corpus, and can be calculated if an exhaustive gold-standard glossary is available. However, recall cannot be determined in the case of manual term extraction (Kris Heylen & Dirk De Hertog 2015:203).

In other words, precision can be conceived as the measure of the software's ability to accurately identify terms. It quantifies the proportion of terms identified by the software that are indeed valid terms. On the other hand, recall represents the number of terms that Sketch Engine successfully identified from the total pool of valid terms available (including those identified manually). If the software fails to recognize certain terms they are automatically assigned to the non-term group, and they result in false negatives. Thus, recall is the ratio of correctly identified terms to the total number of terms available in the specialised corpus.

In order to determine precision and recall, the automatically extracted term candidates had to be divided into two groups: terms and non-terms. Therefore, the terms that fell under the categories "not a term", "not a medical term", "wrong", which were explained previously, were considered as non-terms.

For a better overview and for calculating the precision and recall we are going to look at the categories that were identified during the validation process of the English terms and the number of terms that were attributed to each category. As already mentioned, only the first alphabetically arranged 1200 term candidates proposed by the Sketch Engine were analysed for validation. Here are the exact numbers obtained after the validation process:

Table 11: Number of terms in each of the identified categories (English)

Category	**Number of words in that category**
not a term	449
not a medical term	437
wrong	12
similar to term already in the term base	90
overlooked during manual extraction	212

The three values required to calculate precision and recall are: true positives (TP), false positives (FP), and false negatives (FN). TP is the number of true positives, i.e., the number of instances that are correctly predicted as positive. FP is the number of false positives, i.e., the number of instances that are incorrectly predicted as positive. FN is the number of false negatives, i.e., the number of instances that are incorrectly predicted as negative.

As previously explained, there were 210 terms that were both manually and automatically identified. Therefore, the number of terms that the Sketch Engine failed to identify, 145, was calculated by subtracting the overlapping terms (210) from the total number of manually extracted terms (355). In other words, 145 of the 355 terms that were manually identified, failed to be identified by the Sketch Engine as well.

Considering the information contained in the table above one should be able to determine the following values:

a) TP: the number of terms in the categories "overlooked during manual term extraction" (212) + "similar to term already in the term base" (90) + overlapping terms (210) = 512;

b) FP: the number of terms in the categories "not a term" (449) + "not a medical term" (437) + "wrong" (12) = 898;
c) FN: the number of terms that were not identified by the Sketch Engine, but are, however, part of the manually created term base, i.e. the total number of manually identified terms (355) – overlapping terms (210) = 145.

The precision of the automatic term extraction tool can be calculated by dividing the number of true positives (TP) by the sum of TP and false positives (FP), as expressed by the following formula: TP / (TP + FP) = 0.36 or 36%. This indicates that, for the English data I analysed, the Sketch Engine accurately predicted a word or a phrase as a valid term only 36% of the time.

On the other hand, recall can be calculated by dividing the TP by the sum of TP and false negatives (FN), as represented by the formula: TP / (TP + FN) = 0.77 or 77%. In other words, for the English data I analysed, the Sketch Engine correctly identified 77% of all actual terms present in the specialised corpus.

In conclusion, based on the provided information, the automatic term extraction tool utilized in the study has a precision of 36%, which means that the tool accurately predicts a word as being a valid term only 36% of the time. However, the recall value is 77%, indicating that the tool correctly identifies 77% of all actual terms present in the specialised corpus. The results suggest that while the tool is effective in identifying valid terms, it may produce a high number of false positives. Therefore, it may be necessary to refine the tool to improve its precision.

Regarding the 500 Romanian terms proposed for analysis out of the 2000 automatically extracted by the Sketch Engine, the numbers are as follows:

Table 12: Number of terms in each of the identified categories (Romanian)

Category	**Number of words**
not a term	163
not a medical term	125
wrong	139
similar to term already in the term base	44
overlooked during manual extraction	29

Applying the same methodology as explained earlier, it was determined that the total number of true positives (TP) is 169, the number of false positives (FP) is 427, and the number of false negatives (FN) is 259 for the manually extracted Romanian terms. Based on these values, precision and recall were calculated, revealing that the precision for automatically extracted Romanian terms is 28%, while recall is 39%. These values indicate a reduced performance of the Sketch Engine with concern to the Romanian term candidates as compared to the English ones. However, it should be noted that the Sketch Engine performed a lemmatization process, which may have caused many term candidates to be labelled as "wrong" due to a lack of coherence between the words within the potential term.

7. Conclusions

The COVID-19 pandemic has led to a vast amount of research conducted in various forms, including research papers, public information articles, and press releases, which have been translated into multiple languages, including English and Romanian. Computer systems are being used to automatically extract relevant data from these documents, align and organize it in a standardized format to make it easier to analyse and share. However, there are challenges such as inaccurate automatic alignment between English and Romanian sentences and a shortage of COVID-19 glossaries containing terms translated from English into Romanian, which can negatively impact further research that relies on these bilingual corpora as reference materials.

Consequently, in an attempt to address these challenges, this Master's Dissertation investigated the following research questions:

RQ1: What was the alignment accuracy of the automatically-aligned chosen bilingual corpora?
RQ2: How consistently was EN-RO Covid-19 terminology used in the chosen bilingual corpora?
RQ3: What was the precision and recall of the automatic term extraction tool the Sketch Engine against a gold standard produced through manual human terminology extraction?

To be able to answer the research questions above, the study involves, first and foremost, the creation of a gold-standard glossary in the English-Romanian language combination, which is based on the selected bilingual resources. The glossary represents an accurate and reliable source of terminology and constitutes a significant contribution to the domain of translation studies. Secondly, the study provides a comprehensive evaluation of both manual and automatic term extraction techniques, leveraging the sophisticated features of the Sketch Engine tool. This aspect of the research is important for understanding the strengths and limitations of different methods of term extraction and their impact on translation quality. Finally, the research investigates the quality of various automatically-aligned EN-RO Covid-19 resources, offering insights into the accuracy, consistency, and reliability of such resources in translation practice. In summary, the outcomes of this research have broad implications for translation studies and can serve as a starting point for the development of more effective and reliable translation tools and practices.

The empirical research relied, therefore, on manually processing the chosen bilingual corpora using Trados Studio and the Qualitivity plug-in for annotation purposes to answer the first research question. The Multiterm integration into Trados Studio was utilized to answer the second research question and to create the gold standard for the third research question which was answered using the Sketch Engine.

To address the first research question on the alignment accuracy of the selected bilingual corpora, the "Covid-19 antibiotic dataset", the Qualitivity plug-in in Trados Studio was utilized to annotate the encountered errors. This plug-in generated a comprehensive activities report that detailed the identified errors, their severity, and related content based on the TAUS DQF standard. The analysis aimed to identify various errors such as misalignments, mistranslations, omissions, additions, terminology errors, punctuation errors, spelling mistakes, and other related issues. The activities report enabled the monitoring and tracking of encountered errors.

Based on the activities report, a total of 319 errors were identified from a sample of 1034 segments that were analysed in Trados Studio. However, it should be noted that this number did not necessarily correspond to 319 individual segments since, in some cases, the same segment was associated with multiple types of errors. Therefore, the total number of segments that contained errors was 259. These errors were then classified into 15 different categories, based on the type of error. The analysis identified 61 instances of additions, where words or even sentences that were not present in the source text were added to the translation. Ambiguities, which occurred when the translation of a clear source segment was uncertain or vague, were encountered 20 times throughout the translation. There were 16 instances of segments that were poorly translated or not appropriate for the context. The number of mistranslations was higher than that of other categories, adding up to 93, while the number of misunderstandings was very low, at only 3. There was also a moderate number of instances (34) where essential elements of the source text were absent from the translation, and only 7 instances where the source text was left completely untranslated. Spelling mistakes were encountered 18 times, and there were a few instances of grammatical mistakes (14). The number of errors related to an overly literal translation was very low, at 3, and there were also instances where the style was unidiomatic (1) or where the same mistake was repeated multiple times (2).

However, the most important category for this experiment was related to inconsistencies in the use of terminology throughout the text, with 29 inconsistency errors identified, representing 9.09% of the total errors. This indicates that the terminology was used consistently in almost 91% of the translations. These findings provide the answer to the second research question regarding the consistency of COVID-19 terminology used in the selected bilingual corpora. The study found that the use of medical terminology was generally consistent, with only a small percentage of errors in the use of specific terms.

Furthermore, a comprehensive analysis was conducted to determine the number of misaligned segments from a total of 1034, in order to be able to determine the alignment accuracy of the automatically aligned bilingual corpora. The annotation process involved the use of different quality metrics available in the Qualitivity plug-in for Trados Studio. The purpose of using these quality metrics was, as mentioned before, to identify various errors such as mistranslations, omissions, additions, spelling mistakes, and terminology inconsistencies. With the intention of answering the first research question, the analysis aimed to identify instances of misalignment, which refers to discrepancies in accurately conveying the intended meaning of the original text during the translation of a specific unit of text, such as a sentence or phrase.

To highlight these errors, the status of each misaligned segment was changed to "translation rejected" in Trados Studio. The analysis revealed that out of 1034 segments, 100 were labelled as "translation rejected", resulting in an alignment accuracy of 90.32%. This means that 934 out of 1034 segments were aligned accurately, while 100 segments were misaligned. Nevertheless, these 100 instances of misalignment, that were marked as "translation rejected", should not be confused with the total number of errors (319) discussed under the subchapter "Quality evaluation of translation errors". Only segments where misalignments were encountered were marked as "translation rejected", while those containing merely errors of accuracy, language, style or terminology were not marked as "translation rejected" and are not relevant for the research question regarding the alignment accuracy. In addition, it is worth noting that the segmentation of the chosen bilingual corpora did not have a logical sequence, with the order of the segments being completely random. This lack of order compromised the accuracy of the text segmentation and contributed to the identification of 100 misaligned segments out of the total 1034 segments.

While automatic alignment proved to be a highly effective method for aligning bilingual corpora, there were still instances of misalignment, which may have been caused by various factors such as layout-related issues or the different structures of the texts. Despite this, the analysis indicates that automatic alignment can be a valuable tool when working with large amounts of data, and it can significantly increase efficiency and accuracy in the alignment process.

Another focus point for this master's thesis was assessing the accuracy of the automatic term extraction from chosen bilingual corpora, compared to manual term extraction. The automatic extraction was performed using the Sketch Engine. However, the bilingual term extraction feature was not available for the chosen language combination (EN-RO), and the terms were extracted separately, for each language. Samples from the term candidates proposed by the Sketch Engine were then compared to the ones manually extracted that had been added to the term base. The analysis revealed that out of the 2000 automatically extracted English term candidates, only 210 overlapped with the manually extracted terms. Further analysis of a sample of 1200 of the automatically extracted terms showed that the majority of these term candidates fell under the "not a term" or "not a medical term" categories. The analysis also identified a category of terms that were "overlooked during manual extraction", but were, however, identified correctly by the Sketch Engine. Additionally, it was found that some of the automatically extracted terms were similar to those already present in the term base but differed in grammatical number or the number of words. A small group of term candidates that were incorrect in some aspects was also identified. Five categories were thus identified based on the characteristics of the proposed terms which were then grouped accordingly.

The same process was carried out on the Romanian term candidates as well. Similar to the representation of the English term candidates, each of the five categories identified here was then expressed as a percentage utilizing a pie chart. Similarly to the English terms, the majority of the Romanian terms were classified into the categories of "not a term" and "not a medical term". Additionally, a significant number of terms were categorized as "wrong". However, the number of terms that were "overlooked during manual extraction" was lower compared to the English terms. Furthermore, the category "similar to term already in the term base" also had a lower number of terms.

Therefore, after analysing both pie charts, it can be concluded that the accuracy of the automatic term extraction performed by the Sketch Engine was higher for the English terms than for the Romanian ones, but this results should be, however, regarded with caution, since the analysed samples differ in size. The percentage of the "wrong" category was also much lower (only 1%) for the English part of the study, compared to the 27.8% which was scored by the Romanian term candidates. In other words, a significant portion of the Romanian medical terms proposed by the Sketch Engine were grammatically or syntactically wrong. The Sketch Engine underwent a process of lemmatization, which resulted in terms made up of words that did not agree in number, gender, and case. As a result, many of the term candidates proposed by the Sketch Engine were made up of mismatched words. Therefore, the proposed Romanian term candidates cannot be further used in their current form. For instance, many of the term candidates had a grammatically incorrect construction, such as a feminine noun followed by a masculine adjective. On the other hand, the category representing the terms that were overlooked during manual extraction was larger in the pie chart representing English terms, emphasizing the higher accuracy of the automatic term extraction for English. It should also be mentioned here, that as a non-native English speaker, things are sometimes bound to be overlooked. Additionally, the category representing the terms that were overlooked during manual extraction was bigger in the first pie chart representing English terms, highlighting, again, the fact that the accuracy of the automatic term extraction was higher for the English language. The other three categories, "not a term", "not a medical term", and "similar to term already in the term base", were similar in percentages in both pie charts.

Upon further examining the generated pie charts, on the one hand, it can be inferred that the automatic term extraction performed poorly. With concern to the English term candidates, the categories which scored the highest percentages were "not a term" and "not a medical term". The same goes for the Romanian language as well, since the categories "not a term" and "not a medical term" were the most dominant together with the category "wrong", which scored a significantly high percentage as well. Therefore, the majority of the term candidates proposed by the Sketch Engine were, in fact, not terms as they did not fulfil the necessary criteria.

On the other hand, when comparing the total number of the English term candidates proposed by the Sketch Engine and the terms obtained in the process of manual extraction, the total number of valid terms identified automatically is higher than the number of terms identified manually. The number of terms identified manually amounts to 355, while the total number of valid terms identified by the Sketch Engine adds up to 512 terms. Specifically, while the manual process identified 355 terms, the Sketch Engine identified a total of 512 valid terms after excluding categories such as "not a term", "not a medical term", and "wrong", and focusing only on the categories "overlooked during manual term extraction", "similar to term in the term base", and the overlapping terms. It should be kept in mind that this study only analysed 1200 of the 2000 term candidates automatically detected by the Sketch Engine, so the number of correctly identified terms would have likely been higher if all 2000 were analysed and it should also be taken into consideration that the Sketch Engine imposes a limitation of 2000 automatically extracted terms per language

Similar to the findings for the English part of the experiment, the Sketch Engine identified a lower number of valid Romanian terms compared to the number of term candidates categorized as "not a term", "not a medical term", and "wrong" as well. The total number of correctly identified Romanian terms was only 73, but this result should be interpreted with caution since only a small sample of 500 words was analysed. When adding the 73 Romanian valid terms identified by the Sketch Engine to the 96 terms identified by both manual and automatic methods (overlapping terms), the total number of automatically identified valid terms reaches 169 out of the 500 analysed in this experiment. It is worth noting that 800 English term candidates and 1500 Romanian term candidates were excluded from this experiment.

In conclusion, the task of terminology extraction presents a trade-off between accuracy and efficiency. While manual extraction is deemed more precise, it is also a time-consuming and subjective process. On the other hand, automated extraction, as exemplified by the Sketch Engine, rapidly generates comprehensive lists of term candidates by relying solely on the text corpus. Nonetheless, the nuances of language and context are lost in this approach, requiring the confirmation of a domain expert to ensure the validity of the final term list. The Sketch Engine's main advantage is its speed, saving valuable time. Therefore, the Sketch Engine is an advantageous tool, as it can quickly generate thorough lists of term candidates in just a few seconds, compared to the five-day manual term extraction process used in this study. This saved time can be redirected, for example, towards other practical applications such as terminography.

which utilizes the list of extracted terms to create specialised dictionaries or term bases. This list also plays a crucial role in the translation process, serving as an ad-hoc glossary that helps identify unfamiliar words requiring clarification, thus ensuring consistency throughout the translation. Nevertheless, the lists generated by the Sketch Engine should always be filtered and refined by a domain specialist. In essence, the Sketch Engine serves as an excellent starting point in the terminology extraction process and should be employed in tandem with expert analysis to achieve the most precise and effective results.

The last research question referred to the precision and recall of the automatic extraction tool, the Sketch Engine, against a gold standard, human terminology extraction. After calculating the precision and recall for the automatically extracted English term candidates the following were concluded: precision sums up to 36% or, in other words, when the software predicts a word is an actual term, it is correct 36% of the time and recall sums up to 77%, or, in other words, the Sketch Engine correctly identifies 77% of all actual terms.

Precision and recall were both calculated for the Romanian term candidates as well, revealing that the precision for automatically extracted terms candidates is 28%, while recall is 39%. These values indicate a weaker performance of the Sketch Engine for the Romanian language compared to English. While these rates are not perfect, they still demonstrate the Sketch Engine's potential for saving time in the extraction process and serving as a valuable starting point.

Based on the information provided, it can be concluded that the Sketch Engine's performance for English term extraction is better than for Romanian term extraction, also bearing in mind the higher number of English term candidates that were analysed in this experiment. The precision and recall rates for English are higher at 36% and 77%, respectively, compared to the rates for Romanian at 28% and 39%, respectively. However, even with these low rates, the Sketch Engine can still save time in the extraction process and serve as a valuable starting point for other applications. Overall, while the Sketch Engine's performance may not be perfect, it still shows potential as a tool for term extraction in both English and Romanian languages.

All things considered, the Sketch Engine is a useful tool for automatic terminology extraction, providing an efficient way of identifying term candidates. These, however, need to be further refined by a domain expert, since the accuracy of the extracted terms requires improvement. The balance between efficiency and accuracy remains an important aspect of the terminology extraction process, and the Sketch Engine offers a valuable solution to this dilemma. The Sketch Engine serves as an excellent starting point in the terminology extraction process and should be employed in tandem with expert analysis to achieve the most precise and effective results. By employing expert analysis to validate and refine its results, the Sketch Engine's potential to revolutionize terminology extraction in various fields becomes even more apparent. Although the Sketch Engine's performance isn't flawless, it demonstrates potential as a term extraction tool for both English and Romanian languages.

8. References

Alberts, M. (2011). Lexicography versus Terminography. *Lexikos*, *11*(0), Article 0. https://doi.org/10.5788/11-0-840

Astrakhantsev, N. A., Fedorenko, D. G., & Turdakov, D. Yu. (2015). Methods for automatic term recognition in domain-specific text collections: A survey. *Programming and Computer Software*, *41*(6), 336–349. https://doi.org/10.1134/S036176881506002X

Bai, Y., Yao, L., Wei, T., Tian, F., Jin, D.-Y., Chen, L., & Wang, M. (2020). Presumed Asymptomatic Carrier Transmission of COVID-19. *JAMA*, *323*(14), Article 14. https://doi.org/10.1001/jama.2020.2565

Baisa, V., Ulipová, B., & Cukr, M. (2015). *Bilingual Terminology Extraction in Sketch Engine*. 61–67.

Baker, M., & Saldanha, G. (Eds.). (2019). *Routledge Encyclopedia of Translation Studies* (3rd ed.). Routledge. https://doi.org/10.4324/9781315678627

Boulanger, J.-C. (1992). SAGER, Juan C. (1990): A Practical Course in Terminology Processing, Amsterdam-Philadelphia, John Benjamins Publishing Company, XI + 254 p. *Meta: Journal des traducteurs*, *37*(3), 543. https://doi.org/10.7202/002118ar

Bowker, L. (2015). Terminology and translation. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 304–323). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.16ter5

Cabré Castellví, M. T. (2003). Theories of terminology: Their description, prescription and explanation. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *9*(2), 163–199. https://doi.org/10.1075/term.9.2.03cab

Cabré Castellví, M. T., Condamines, A., & Ibekwe-SanJuan, F. (2007). Introduction:

Application-driven terminology engineering. In F. Ibekwe-SanJuan, A. Condamines, & M. T. Cabré Castellví (Eds.), *Benjamins Current Topics* (Vol. 2, pp. 1–17). John Benjamins Publishing Company. https://doi.org/10.1075/bct.2.03cab

Cabré Castellví, M. T., & Sager, J. C. (1998). Preface: Towards a theory of terminology. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *5*(1), 1–3. https://doi.org/10.1075/term.5.1.02cab

Cerrella Bauer, S. (2015). Managing terminology projects: Concepts, tools and methods. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 324–340). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.17man1

Chan, S. (Ed.). (2015). *Routledge encyclopedia of translation technology*. Routledge, Taylor & Francis Group.

Chen, N., Zhou, M., Dong, X., Qu, J., Gong, F., Han, Y., Qiu, Y., Wang, J., Liu, Y., Wei, Y., Xia, J., Yu, T., Zhang, X., & Zhang, L. (2020). Epidemiological and clinical characteristics of 99 cases of 2019 novel coronavirus pneumonia in Wuhan, China: A descriptive study. *The Lancet*, *395*(10223), Article 10223. https://doi.org/10.1016/S0140-6736(20)30211-7

Daille, B., Gaussier, É., & Langé, J.-M. (1994). Towards automatic extraction of monolingual and bilingual terminology. *Proceedings of the 15th Conference on Computational Linguistics* -, *1*, 515. https://doi.org/10.3115/991886.991975

Depecker, L. (2015). How to build terminology science? In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 34–44). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.03how1

DIN 2342:2011-08, *Begriffe der Terminologielehre*. (n.d.). Beuth Verlag GmbH. https://doi.org/10.31030/1759531

Dobrina, C. (2015). Getting to the core of a terminological project. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 180–199). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.get1

Ettman, C. K., Abdalla, S. M., Cohen, G. H., Sampson, L., Vivier, P. M., & Galea, S. (2020). Prevalence of Depression Symptoms in US Adults Before and During the COVID-19 Pandemic. *JAMA Network Open*, *3*(9), e2019686. https://doi.org/10.1001/jamanetworkopen.2020.19686

European Foundation for the Improvement of Living and Working Conditions. (2021). *COVID-19: Implications for employment and working life*. Publications Office. https://data.europa.eu/doi/10.2806/160624

Faber Benítez, P. (2009). The cognitive shift in terminology and specialized translation. *MonTi: Monografías de Traducción e Interpretación*, *1*, Article 1. https://doi.org/10.6035/MonTI.2009.1.5

Faber, P. (2015). Frames as a framework for terminology. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 14–33). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.02fra1

Felber, H., & Budin, G. (1989). *Terminologie in Theorie und Praxis*. G. Narr.

Görög, A. (2014). Dynamic Quality Framework: Quantifying and benchmarking quality. *Tradumàtica: Tecnologies de La Traducció*, *12*, 443. https://doi.org/10.5565/rev/tradumatica.66

Guan, W., Ni, Z., Hu, Y., Liang, W., Ou, C., He, J., Liu, L., Shan, H., Lei, C., Hui, D. S. C., Du, B., Li, L., Zeng, G., Yuen, K.-Y., Chen, R., Tang, C., Wang, T., Chen, P., Xiang, J., … Zhong, N. (2020). *Clinical characteristics of 2019 novel coronavirus infection in China* [Preprint]. Respiratory Medicine. https://doi.org/10.1101/2020.02.06.20020974

Haas, W. (1973). John Lyons' 'Introduction to theoretical linguistics.' *Journal of Linguistics*, *9*(1), 71–113. https://doi.org/10.1017/S0022226700003558

Hamed, M. A. (2020). An overview on COVID-19: Reality and expectation. *Bulletin of the National Research Centre*, *44*(1), Article 1. https://doi.org/10.1186/s42269-020-00341-9

Heylen, K., & De Hertog, D. (2015). Automatic Term Extraction. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 203–221). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.11aut1

Humbley, J. (2018). 22. Term Formation And Neology. In J. Humbley, G. Budin, & C. Laurén (Eds.), *Languages for Special Purposes* (pp. 437–452). De Gruyter. https://doi.org/10.1515/9783110228014-022

Ivanović, T., Stanković, R., Todorović, B. Š., & Krstev, C. (2022). Corpus-based bilingual terminology extraction in the power engineering domain. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *28*(2), 228–263. https://doi.org/10.1075/term.20038.iva

Kageura, K. (2012). *The Quantitative Analysis of the Dynamics and Structure of Terminologies* (Vol. 15). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.15

Kang, S., Peng, W., Zhu, Y., Lu, S., Zhou, M., Lin, W., Wu, W., Huang, S., Jiang, L., Luo, X., & Deng, M. (2020). Recent progress in understanding 2019 novel coronavirus (SARS-CoV-2) associated with human respiratory disease: Detection, mechanisms and treatment. *International Journal of Antimicrobial Agents*, *55*(5), Article 5. https://doi.org/10.1016/j.ijantimicag.2020.105950

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kockaert, H., & Steurs, F. (Eds.). (2015). *Handbook of terminology*. John Benjamins Publishing Company.

Löckinger, G., Kockaert, H. J., & Budin, G. (2015). Intensional definitions. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 60–81). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.05int1

Lyons, J. (1968). *Introduction to Theoretical Linguistics* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139165570

Melby, A. K. (2012). Terminology in the age of multilingual corpora. *The Journal of Specialised Translation*, *18*, 7–28.

Morin, E., & Hazem, A. (2016). Exploiting unbalanced specialized comparable corpora for bilingual lexicon extraction. *Natural Language Engineering*, *22*(4), 575–601. https://doi.org/10.1017/S1351324916000140

Nilsson, H. (2015). Enumerations count: Extensional and partitive definitions. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 82–100). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.06enu1

Nuţu, M. (2021). Deep Learning Approach for Automatic Romanian Lemmatization. *Procedia Computer Science*, *192*, 49–58. https://doi.org/10.1016/j.procs.2021.08.006

Oeser, E., & Budin, G. (1998). Grundlagen der Terminologiewissenschaft. In L. Hoffmann, H. Kalverkämper, H. E. Wiegand, C. Galinski, & W. Hüllen (Eds.), *Fachsprachen* (pp. 2171–2183). De Gruyter. https://doi.org/10.1515/9783110158847.2.24.2171

Oliver, A. (2017). A system for terminology extraction and translation equivalent detection in real time: Efficient use of statistical machine translation phrase tables. *Machine Translation*, *31*(3), Article 3. https://doi.org/10.1007/s10590-017-9201-7

Ovallath, S. (2020). *Covid-19*. Nova Medicine & Health.

Pandey, A., Nikam, A. N., Shreya, A. B., Mutalik, S. P., Gopalan, D., Kulkarni, S., Padya, B. S., Fernandes, G., Mutalik, S., & Prassl, R. (2020). Potential therapeutic targets for combating SARS-CoV-2: Drug repurposing, clinical trials and recent advancements. *Life Sciences*, *256*, 117883. https://doi.org/10.1016/j.lfs.2020.117883

Pandey, S., Yadav, B., Pandey, A., Tripathi, T., Khawary, M., Kant, S., & Tripathi, D. (2020). Lessons from SARS-CoV-2 Pandemic: Evolution, Disease Dynamics and Future. *Biology*, *9*(6), Article 6. https://doi.org/10.3390/biology9060141

Pavel, S., Nolet, D., & Leonhardt, C. (2001). *Précis de terminologie*. Travaux publics et services gouvernementaux Terminologie et normalisation, bureau de la traduction.

Pazienza, M. T., Pennacchiotti, M., & Zanzotto, F. M. (2005). Terminology Extraction: An Analysis of Linguistic and Statistical Approaches. In S. Sirmakessis (Ed.), *Knowledge Mining* (Vol. 185, pp. 255–279). Springer-Verlag. https://doi.org/10.1007/3-540-32394-5_20

Prather, K. A., Wang, C. C., & Schooley, R. T. (2020). Reducing transmission of SARS-CoV-2. *Science*, *368*(6498), Article 6498. https://doi.org/10.1126/science.abc6197

Reynolds, P. (2015). Machine translation, translation memory and terminology management. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 276–287). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.14mac1

Roelcke, T. (2018). 25.Technical Terminology. In J. Humbley, G. Budin, & C. Laurén (Eds.), *Languages for Special Purposes* (pp. 489–508). De Gruyter. https://doi.org/10.1515/9783110228014-025

Sager, J. C. (1990). *A Practical Course in Terminology Processing*. John Benjamins Publishing Company. https://doi.org/10.1075/z.44

Santos, C., & Costa, R. (2015). Domain specificity: Semasiological and onomasiological knowledge representation. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 153–179). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.09dom1

Sarvepalli, D. (2020). Coronavirus Disease 2019: A Comprehensive Review of Etiology, Pathogenesis, Diagnosis, and Ongoing Clinical Trials. *Cureus*. https://doi.org/10.7759/cureus.8076

Steurs, F., De Wachter, K., & De Malsche, E. (2015). Terminology tools. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 222–249). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.12ter3

Temmerman, R. (2000). *Towards New Ways of Terminology Description: The sociocognitive approach* (Vol. 3). John Benjamins Publishing Company. https://doi.org/10.1075/tlrp.3

ten Hacken, P. (2015). Terms and specialized vocabulary: Taming the prototypes. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 3–13). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.ter1

Thurmair, G. (2003). *Making Term Extraction Tools Usable*. Joint Conference of the 8th Workshop of the European Association for Machine Translation and the 4th Controlled Language Applications Workshop, Dublin.

Vo, C., Cao, T., Truong, N., Ngo, T., & Bui, D. (2022). Automatic medical term extraction from Vietnamese clinical texts. *Terminology. International Journal of Theoretical and Applied Issues in Specialized Communication*, *28*(2), 299–327. https://doi.org/10.1075/term.20037.vo

Warburton, K. (2015). Managing terminology in commercial environments. In H. J. Kockaert & F. Steurs (Eds.), *Handbook of Terminology* (Vol. 1, pp. 360–392). John Benjamins Publishing Company. https://doi.org/10.1075/hot.1.19man2

Wiese, I. (2018). 27. Terminology Work In Different Domains: Medical Terminology. In J. Humbley, G. Budin, & C. Laurén (Eds.), *Languages for Special Purposes* (pp. 522–534). De Gruyter. https://doi.org/10.1515/9783110228014-027

8.1. Internet sources

Qualitivity. RWS Group, 2023, https://community.rws.com/product-groups/trados-portfolio/rwsappstore/w/wiki/2251/qualitivity.

SDL Trados Studio. SDL Trados Studio, 2023, https://www.trados.com/products/trados-studio/.

Sketch Engine. Sketch Engine, 2023, https://www.sketchengine.eu/.

Sketch Engine. "How does it work?" Sketch Engine Terms. Sketch Engine, 2023, https://terms.sketchengine.eu/how-does-it-work/.