

MASTER'S THESIS IN LIBRARY AND INFORMATION SCIENCE  
SWEDISH SCHOOL OF LIBRARY AND INFORMATION SCIENCE  
2021

# Science communication on Twitter

An analysis of vocabulary and content

Görel Sundström



UNIVERSITY  
OF BORÅS

© Görel Sundström

Partial or full copying and distribution of the material in this thesis  
without permission is forbidden.

Title: Science communication on Twitter: An analysis of vocabulary and content

Author: Görel Sundström

Completed: 2021

Abstract:

Twitter is one platform where scientists can communicate their research results, both among each other and to a wider audience. This master thesis investigates to what extent, and by which means, tweets with scientific content invite the general public to engage in the topics. The four different topics analysed in this study were: *C.elegans/Neuromyelitis*, *Staphylococcus*, *mRNA expression* and *Species diversity/Phylogenetic tree*. Several methods were used to analyse these datasets, such as identification of jargon, content analysis and word frequencies, analysed within the metadiscourse framework stance and engagement. All in order to detect any intentions of communication outside the academic circle. It was possible to detect communicative and descriptive content in two of the topics, *mRNA expression* and *Species diversity/Phylogenetic tree*. The vocabulary was analysed in both of these topics, detecting a high frequency of reader-mentions and markers for novelty, something that has been seen in other kinds of media producing popular science. However, for most tweets with scientific content the main receivers seem to be other researchers in the same fields. Tweets containing links to scientific articles predominantly contain only the title of the article. One prominent aspect of Twitter is its changing nature. This can be seen in this study where tweets from the topics *Staphylococcus* and *Species diversity/Phylogenetic tree* had links to news media. If the datasets were collected today, tweets from the topic *mRNA expression* would probably also display this pattern.

Keywords: science communication, Twitter, vocabulary, jargon, content analysis

- *In bocca al lupo.*
- *Crepi il lupo.*

High Garda,  
The Great Library



## Table of Contents

1. Introduction.....	5
1.1 Social media, dissemination and miscommunication.....	6
1.2 Research gap, aim and research questions.....	8
1.3 Outline.....	9
2. Literature review.....	11
2.1 Researching Twitter.....	11
2.1.1 Collect data.....	11
2.1.2 Analyse data.....	12
2.1.3 Problems and obstacles.....	14
2.2 Researchers on Twitter.....	15
2.2.1 Twitter users and interdisciplinary differences.....	15
2.2.2 Twitter usage.....	16
2.2.3 Altmetrics and Twitter.....	17
2.3 Dissemination, vocabulary and social media.....	19
2.3.1 Research dissemination and communication.....	19
2.3.2 Vocabulary and jargon.....	20
2.3.3 Research communication on social media.....	21
3. Theory.....	23
3.1 A brief overview of metadiscourse.....	23
3.2 Stance and engagement.....	24
4. Material and methods.....	27
4.1 Data collection.....	28
4.2 Datasets.....	28
4.2.1 <i>Caenorhabditis elegans</i> / <i>Neuromyelitis</i> .....	29
4.2.2 Staphylococcus.....	29
4.2.3 mRNA expression.....	30
4.2.4 Species diversity/phylogenetic tree.....	31
4.3 Methods.....	31
4.3.1 Descriptive statistics and pre-processing.....	32
4.3.2 Word frequency analysis.....	33
4.3.3 Jargon detector.....	33
4.3.4 Content analysis.....	34
4.3.5 Vocabulary analysis.....	35
4.4 Ethical considerations.....	37
5. Result and analysis.....	39
5.1 Descriptive statistics and pre-processing.....	39
5.2 Word frequency analysis.....	42
5.3 Jargon detector.....	45
5.4 Content analysis.....	47
5.4.1 Type of source linked to.....	47
5.4.2 Content of tweets.....	49
5.5 Vocabulary analysis.....	53
6. Discussion.....	57
6.1 Representativity, generalisation, and limitations.....	57
6.2 Communication on Twitter.....	59
6.3 Conclusions and further studies.....	63
References.....	67
Appendix.....	75

## List of Figures

Figure 1: Schematic phylogenetic tree. Proposed divergence times in million of years marked with triangles.....	31
Figure 2: Flow-chart with the methods and the order they are used in.....	32
Figure 3: Distribution of tweets, with and without URLs, in the investigated topics.....	39
Figure 4: Percentage of tweets that are RTs (retweets), in all topics.....	40
Figure 5: Proportion of mentions, @, calculated against total number of tweets.....	41
Figure 6: <i>C. elegans</i> /Neuromyelitis, word frequency comparison.....	42
Figure 7: Staphylococcus, word frequency comparison.....	43
Figure 8: mRNA expression, word frequency comparison.....	44
Figure 9: Species diversity/Phylogenetic trees, word frequency comparison. ....	45
Figure 10: Distribution of the types of sources the tweets links to in the different topics.....	48
Figure 11: Content of tweets in the topic <i>C. elegans</i> /Neuromyelitis, n=20....	50
Figure 12: Content of tweets in the topic Staphylococcus, n=79.....	51
Figure 13: Content of tweets in the topic mRNA expression, n=269.....	52
Figure 14: Content of tweets in the topic Species diversity/Phylogenetic tree, n=184.....	53

## Index of Tables

Table 1: Types and examples of “stance”.....	36
Table 2: Types and examples of "engagement".....	36
Table 3: Number and percentage of words classified as common, mid or rare in the different topics.....	46
Table 4: Log-likelihood for mid- and rare-frequency words.....	46
Table 5: Topic comparison, markers of stance.....	54
Table 6: Topic comparison, markers of engagement.....	55
Table 7: Topic comparison, markers of novelty.....	56

## 1. Introduction

One can call it science communication, research dissemination or popular science, but it all comes down to one thing – results from the academia needs to be presented to, and be a part of, society as a whole. In Sweden this is called "den tredje uppgiften" and regulated by the law as one of the main tasks for the university, besides teaching and research (SFS (1992:1434)). This mission is similar, but not equivalent to, what is called the third mission or the third stream internationally. In this international perspective the focus is on how universities and other academic institutions can contribute to regional development by encouraging entrepreneurship and commercialising of scientific research. Education and interaction with the society is however also considered a part, albeit smaller, of the third mission also outside Sweden (Compagnucci & Spigarelli, 2020).

Several variants of how this interaction between society and the academia can occur have been formulated in ways such as *public communication*, *public consultation* and *public participation*. In the first concept is the information going from the scientists to the public, in the second are the public giving feedback, while the latter aims at dialogue. This type of continuing exchange is in line with the current view of research communication and dissemination occurring on social media (Hargittai, Füchslin, & Schäfer, 2018).

The commentary functions included in social media are seen as a way to interact with a wider audience. The general public is often considered the target when writing science-related blogs (Sugimoto, Work, Larivière, & Haustein, 2017) and one interesting question in this context is if the vocabulary used matches the intended reader. It is at the same time important not to underestimate the readers and remember that followers of scientific blogs and Twitter accounts are already interested in these subjects. Côté and Darling (2018) describes it as an extension of the echo-chamber that Twitter can be described as, the tweets reach a public interested in scientific topics, but Twitter is not automatically a way to get a broad dissemination out in to society. Nevertheless, the size and diversity of the public is often larger than for a scientific article.

Studies of communication and its relation to a new information society, digital transition and digital literacy is something that are covered by the wide subject of library and information science. A model often used is one of a communication chain, where it is possible to study and describe the process

from creation of information to its disposal. Included in this model are the steps dissemination, organisation, indexing, storage and use, (Bawden & Robinson, 2015) of which the first is of particular interest for this project. Also the last concept, use, is interesting since it connects to another area of Library and information sciences, namely bibliometrics and altmetrics.

Several studies have investigated what effect communicating research via social media actually have on the number of citations received for an article, and the results shows marginal or slightly positive correlations (Costas, Zahedi, & Wouters, 2015; Thelwall, Haustein, Larivière, & Sugimoto, 2013). However, regardless of what outcome this type of science communication and research dissemination has on research impact it is part of today's society, and altmetrics is a complement to the more traditional bibliometric measurements. Altmetrics has been described as “research indicators based on social media activity” (Sugimoto et al., 2017) and also as a metric for the societal impact of research (Díaz-Faes, Bowman, & Costas, 2019; Tahamtan & Bornmann, 2020).

Key terms for research dissemination via Twitter or other social media are communication and interaction. The importance of these concepts can not be underestimated while aiming at increasing the public awareness and understanding in science (Álvarez-Bornstein & Montesi, 2019; Su, Scheufele, Bell, Brossard, & Xenos, 2017). This master thesis will investigate to what extent, and by which means, tweets with scientific content invites the general public to engage in the topics.

## 1.1 Social media, dissemination and miscommunication

In order to awake an interest and engagement in science and science-related questions it is necessary to provide meeting places where interaction can take place. Results, values, and knowledge need to be discussed and explained, and different forms of social media have been identified as such places (Büchi, 2017 and references therein; Della Giusta, Jaworska, & Vukadinović Greetham, 2021; Freddi, 2020). Twitter is one social media platform and its relation to research dissemination has been described as follows: “The possibility that scholars can push their research out, rather than hope that it is pulled in, holds the potential for scholars to draw wide attention to their research.” (Klar, Krupnikov, Ryan, Searles, & Shmargad, 2020). An aspect often lifted in this discussion is the importance of correct disseminate results, with an appropriate vocabulary and without relying on the journalistic touch (Kuehne & Olden, 2015).

One question is of course if the material presented in tweets is picked up by a wider audience or if it stays in a limited bubble, sometimes formulated as the contrast between inreach – preaching to the choir, and outreach –

singing from the rooftops (Côté & Darling, 2018). Many publications are mentioned on Twitter, and the fact that it is a platform with many non-academic users indicates that this is a place where such interaction can occur (Mohammadi, Thelwall, Kwasny, & Holmes, 2018). It should however be noted that voices have been raised, claiming that Twitter might have evolved into an informal discussion forum among colleagues in the academia (Sugimoto et al., 2017).

Twitter is sometimes considered a hybrid form between a recommendation service and a networking platform (Schmitt & Jäschke, 2017) and the tone and language used have changed throughout the years, from formal to informal (Della Giusta et al., 2021). One problem while studying any aspects of social media is its changing nature. In the year 2012 12% of the articles published in PLoS were included in at least one tweet, and 3 years later (2015) the same figure was 53% (Sugimoto et al., 2017). This has been highlighted as one confounding factor while interpreting altmetrics scores, but it also shows the development of a practice.

If this outreach process and dissemination of research is going to be successful the scientific jargon needs to be abandoned. Usage of jargon has been shown to diminish the ability to process information as well as change how the information is perceived (Bullock, Colón Amill, Shulman, & Dixon, 2019). Studies like this have led to the development of tools for both jargon detection and jargon quantification in order to assist the researchers in their communication (Rakedzon, Segev, Chapnik, Yosef, & Baram-Tsabari, 2017; Willoughby, Johnson, & Serman, 2020). It has also been suggested that the format of Twitter, with the limited number (280) of characters available, invites and forces the users to express themselves carefully and clearly (Denia, 2020). This would indicate that if the purpose of scientific tweets is research dissemination, care would be taken to not include jargon. Previous studies have shown that the spontaneous and informal element present while writing science blogs are one part of their success (Freddi, 2020), something that also might strengthen this speculation. To what extent Twitter, as a social medium, really contains a low amount of jargon in an attempt to reach out to a wider audience is however unknown and needs to be investigated.

To summarise, social media is considered a place where interaction with the general public can take place, the phenomenon is dynamic in its nature and up-to-date analyses are always necessary. The problem is that the outreach process might not be successful if the vocabulary used is not adapted for a non-specialist audience. If the message in the tweet is not understandable for its intended audience it will not be transmitted correctly and no real research dissemination have occurred. Notions like this can also undermine the value of the metrics measuring this kind of impact.

## 1.2 Research gap, aim and research questions

The aim of this study is to investigate how scientific content are communicated on the social media platform Twitter, and to explore what kind of features are used in order to invite the general public to engage in the topics.

Twitter is one of the platforms available for researchers and scientists to disseminate their results and communicate with lay-people. It is a platform for communication where recommendations and linking are prominent features, the latter especially for tweets with scientific content (Büchi, 2017). Often it is only the title of the scientific article that makes up the content of tweets linking to these sources, something that has been noticed in earlier studies (Thelwall 2013). A similar analysis will be conducted in this project; the previous studies are however not a hindrance since one of the most distinctive features of social media is its changing nature, and the research dissemination practice is no exception. It has been suggested in earlier research that proficiency and domain knowledge in the topics tweeted about can be an advantage in this kind of analysis (Büchi, 2017; Holmberg & Thelwall, 2014), and since I have a PhD from the Faculty of Medicine with focus on gene and genome evolution the datasets chosen for this study will be familiar to me.

Comparisons of the content of tweets are not often conducted in Twitter research, and to my knowledge there exist no studies describing interdisciplinary differences in the practice of what additional content is included while tweeting links to scientific articles. Other research have shown that almost half of embedded links are never clicked on (Fang, Costas, Tian, Wang, & Wouters, 2021), but this without any discussion about how the content of the tweet, besides the links themselves, influenced the tendency to click.

Some studies also define scholarly or scientific tweets as only those linking to a research article (Joubert & Costas, 2019). This limits the tweets used in the analysis in contrast to this investigation where all tweets containing specific scientific query terms are included.

One way to investigate attempts to interaction with the readers is by analysing the vocabulary. This has been done for academic blogs where interdisciplinary differences were detected (Zou & Hyland, 2020). I have also found one study where detection and discussion of metadiscourse markers in tweets occurred, however, the Twitter data used in that study covered tweets from two specific academic conferences and not research dissemination and scientific tweets (Luzón & Albero-Posac, 2020).

Taken together, a combination of content and vocabulary analyses will hopefully be able to answer the following research questions.

*RQ1 How prevalent is the use of scientific jargon in tweets with scientific content?*

*RQ2 What kind of information do the tweets with embedded links and scientific content contain?*

*RQ3 Using Twitter data as an example of research communication, what differences are possible to see in the vocabulary used between scientific topics?*

### 1.3 Outline

This master thesis includes a literature review divided into three parts giving an overview on how to research Twitter, what researchers themselves do on Twitter, and how vocabulary usage in science communication is studied. The following section describes in more depth the concept of metadiscourse, with focus on the stance and engagement model used in this study. The material and methods section gives both a description of the different datasets used in the analyses, and explanations of the methods used. The results from each method are thereafter presented and analysed in the following section. The master thesis ends with a discussion where the results are the main focus but both the methods and limitations are discussed as well as conclusions and future research.



## 2. Literature review

Research covering Twitter, research communication and vocabulary use are large fields. This literature review focuses on three main areas: the first touches upon how to research Twitter, including examples of different approaches that can be taken while studying, as well as known methodological problems and other obstacles important to take into consideration. The second part includes an overview of what is known about what researchers do on Twitter, how they act, why they are there and what the benefits are. A brief overview of the concept of altmetrics is included in connection to this. The third part examines some literature regarding research dissemination with focus on information written and spread by the researchers themselves via blogs, popular science articles, or Twitter, and mainly covers articles that address the use of adjusted vocabulary in research dissemination.

### 2.1 Researching Twitter

Twitter has been a research subject for some years now, but a large part of the research done still includes method development as a more or less explicit part of the aim of the study. The main considerations in this respect involve what can be studied, what are the best ways to analyse the data, and if it is possible to get representative samples and thereafter formulate generalised conclusions.

#### 2.1.1 *Collect data*

From a technical point of view most studies use the Twitter API in order to get access to the tweets (Gaffney & Puschmann, 2013), how to select which tweets to use in the analysis differs however from study to study. The method chosen depends on the objective of the study in question, and can have its starting point in the tweets themselves or the users of Twitter.

For the latter approach it is possible to start by identifying a handful of users, and thereafter collect the followers and the followers' followers. This has been done while for example investigating the differences in tweeting practice between research disciplines (Holmberg & Thelwall, 2014) and also in order to examine who is reading tweets with scientific content (Côté & Darling, 2018). A similar approach was used in a study by Schmitt and Jäschke (2017) where attendances to computer scientist conferences were used in order to collect a selection of Twitter users from that profession. It is

also possible to analyse the tweets, their reception and responses from just one user, which has been done for the prominent Twitter-profile Neil deGrasse Tyson, an American astrophysicist (Denia, 2020).

The other main avenue in order to collect data for Twitter research is to focus on the content and use specific keywords to select the tweets to study. These keywords can be retrieved in different ways, one is to utilise the concept of hashtags and use them as hook to collect the tweets of interest, while for example investigating the Twitter-use at academic conferences (Luzón & Albero-Posac, 2020) and how a specific scientific topic are discussed (Haunschild, Bornmann, Potnis, & Tahamtan, 2021; Haunschild, Leydesdorff, & Bornmann, 2020).

Another way is to extract keywords of interest from other sources and then use them as query-terms while retrieving tweets. These external sources can be from the academic world, such as topics discussed in a restricted selection of research papers (Haunschild et al., 2020) or other lists of both keywords and key terms used in specific disciplines (Della Giusta et al., 2021). It is also possible to collect words present in contemporary online news, and use those as query-terms on Twitter, which has been done in order to investigate how current issues are discussed in social media (Büchi, 2017).

In order to get a fuller picture of conversations on Twitter it is also possible, and necessary to, collect data from both hashtags and users simultaneously. Using this method facilitates the study of conversations as well as subconversations and follow-on conversations (Lorentzen & Nolin, 2017).

The service altmetric.com (altmetric.com, n.d.) collects altmetric information and via that route is it possible to retrieve tweets mentioning scientific articles and thereafter analyse both the users behind the accounts as well as the tweets themselves (Didegah, Mejlgaard, & Sørensen, 2018; Joubert & Costas, 2019; Mohammadi et al., 2018). A version of this is not to use the altmetric.com but instead select a number of articles, collect the URLs for each article and thereafter use these URLs as query terms while extracting tweets for analysis (Klar et al., 2020).

### *2.1.2 Analyse data.*

Beside the different ways to collect data there are also several ways to analyse data, as well as other ways to study Twitter besides the tweets themselves.

The Twitter users, their behaviour and reasons for using the platform have been studied in surveys. Questionnaires have been aimed at followers of active researchers (Álvarez-Bornstein & Montesi, 2019), also Twitter-users that have shared links to scientific articles have been studied (Mohammadi et al., 2018). Questions about Twitter usage in relation to scientific content have

been part of larger panel studies investigating social media usage in general, and this data can and have been extracted and analysed separately (Hargittai et al., 2018). Related to these surveys are studies investigating the digital behaviour of the users by investigating to what extent links embedded in tweets are actually clicked on, and thereby giving indications on what content awakes interest enough to continue reading somewhere else (Fang et al., 2021).

Analyses of the content of tweets can and have been investigated in several ways and with different focus. These content analyses can be performed by manual coding of selected samples of tweets and the general message – what the tweets are communicating is noted (Holmberg & Thelwall, 2014). Other studies use the same manual process but focus on how the message is communicated instead, for instance by observing the use of emojis and emoticons (Luzón & Albero-Posac, 2020). Hashtags also have been the focus in content analyses; due to the special structure it is possible to automatically extract them from larger datasets (Holmberg & Thelwall, 2014). The latter types of analyses are connected to vocabulary analyses where different linguistic aspects of the tweets are studied. It is also common to use a more automatic approach as the first step in vocabulary analyses in order to get the frequencies of the most commonly used words. Studies following this strategy have investigated the difference in vocabulary between research disciplines (Della Giusta et al., 2021), as well as concluding that emotional content in tweets gives more responses from the followers (Denia, 2020). The latter study used the Tidyverse packages (Silge & Robinson, 2017; Wickham et al., 2019) further described in the methods section. Large scale frequency analyses have also been used as a tool while comparing co-occurrence of words in online news with those describing the same features on Twitter (Büchi, 2017). More descriptions and references to vocabulary analyses are available in the theory section below.

Another popular way to analyse Twitter data is by doing network analyses, which has been done on topic level where often differences in keyword network structure between research articles and tweets are investigated (Haunschild et al., 2021, Haunschild et al., 2020). Similar comparisons have been performed between news articles and tweets (Büchi, 2017). It is also possible to use networks in order to describe Twitter on other levels, both when describing the relation between words within tweets (Della Giusta et al., 2021), and networks of the mentions in a dataset (Büchi, 2017).

These networks of mentions are connected to what is described in the beginning of this section, different methods used to study the users of Twitter. Beside the mentions, retweets and URLs are also investigated in several studies in an attempt to describe datasets of tweets and get a clearer picture of

both content, users and user practices (Büchi, 2017; Didegah et al., 2018; Fang et al., 2021; Holmberg & Thelwall, 2014).

### *2.1.3 Problems and obstacles*

The main problem when it comes to Twitter research is questions about if a sample can be considered representative and what the possibilities of generalisation is. The problem with these concepts also occurs on different levels.

The first can be a question about geography and if it is possible to draw conclusions about Twitter usage and behaviour of a large population. While discussing this in a more world wide view it is important to remember that Twitter is illegal in some countries such as China, and other similar platforms are used instead (Mohammadi et al., 2018). There exists, however, studies that try to include a geographic component in the research, in order to map possible differences in usage from different parts of the world, and the geolocation can be done using different methods. It is possible to get a geographic location from the altmetric.com service, which was utilised in a study identifying the regional pattern of Twitter users in Africa (Joubert & Costas, 2019). It is noted in the article that it is only a little more than half of the Twitter users that have an assigned geolocation in that database. Other studies have pointed out that Twitter itself offers a “superficial distinction between countries” which limits opportunity to generalisation in such questions (Denia, 2020).

Another problem related to the discussion on the possibility to draw conclusions from these kind of datasets is the question about who is tweeting. This is something that is continuously discussed in the literature and will be touched upon throughout this study, however some general issues will be mentioned here. The first is how much influence the so-called bots, automatic social media accounts, have on both the discussions occurring on Twitter and the impact they have on the research results. The conclusion drawn from two recent studies investigating different fields of research is that the bots are a large part of the research dissemination process. However, their tweeting practices are not different than those of humans, so their influence of what is distributed is limited (Didegah et al., 2018; Haunschild et al., 2021). Related to this is the common practice of using Twitter as a way to promote one’s own article. This kind of tweeting is conducted by both the authors, journals and publishing houses (Álvarez-Bornstein & Montesi, 2019; Klar et al., 2020). Scientific articles that only are tweeted about once have therefore been considered noise in some studies, and been excluded for the analyses (Haunschild et al., 2021; Haunschild et al., 2020).

Twitter is a social media platform and analysis of single tweets will always only get part of the full picture – the replies, retweets and conversation are

often missed while using keyword and hashtag queries. This is a problem of incompleteness that has its base in the reliance on the API provided by Twitter, which includes restrictions of how much data can be collected. There also exist concerns that these constraints might in a way control the research questions asked in the field (Gaffney & Puschmann, 2013). A general awareness and discussion of what is captured and not captured in the datasets is therefore recommended in order to put the spotlight on these problems (Lorentzen & Nolin, 2017).

A final troublesome feature while working with data from Twitter or other social media is its changing nature. This issue is partly because it is a relatively new phenomenon; tweeting about research and linking to scientific articles this way is still a growing practice in some areas. One example mentioned in the introduction is the percentage of articles published by PLoS (a publisher focusing on Open Access publication) that have been tweeted have increase from 12% to 53% between 2012 and 2015 (Sugimoto et al., 2017). This has implication for the evaluation of altmetrics methods, see section 2.2.3 for more background and description, and the problems with altmetric measurements and older articles articles have been discussed (Thelwall et al., 2013). It is not only what is posted that is changing with time, how the content is presented has also changed. A transition from a more formal style inspired by journalistic texts, to a more spontaneously and relaxed tone can be seen for some disciplines (Della Giusta et al., 2021).

## 2.2 Researchers on Twitter

The researchers present on Twitter, together with their activities, have been studied and analysed in many aspects. This part of the literature review will discuss who the users are, what they are doing on Twitter, and why they are active on the platform. The answer to the latter is often related to their professional roles as researchers and writers of articles. Therefore an overview of altmetrics, its importance and relation to Twitter is also included.

### *2.2.1 Twitter users and interdisciplinary differences*

One way of investigating the use of Twitter in the research community is to map and characterise the researchers that utilise Twitter, who is tweeting, how they are describing themselves, how many followers they have and what they are tweeting about. This data can then be compared to the scientific community in other countries, regions, or research disciplines. Or it can be used in order to approximate the potential influence the owners of the accounts can have (Joubert & Costas, 2019). The mapping itself has been done by collecting data from the information announced in the biographical description on the Twitter account (Joubert & Costas, 2019), or by conducting a more in-depth study using a survey (Mohammadi et al., 2018).

Also more personal reflections and comments have been given by researcher regarding Twitter usage (Britton, Jackson, & Wade, 2019).

World-wide comparisons have shown that North America and Western Europe dominates the Twitter scene when it comes to the origin of tweets that concern research articles, a conclusion that also is in line with the number of articles produced (Joubert & Costas, 2019). Besides these geographical difference the usage pattern between disciplines and research areas has been investigated. In one study the number of retweets was used as an indicator of scientific communication on Twitter and it was concluded that users from the field of biochemistry were most prone to promote and share research using that function (Holmberg & Thelwall, 2014). The concept of retweeting have however also been described as mechanical and as a sign of lack of intent to conversation and communication (Robinson-Garcia, Costas, Isett, Melkers, & Hicks, 2017). Interdisciplinary differences have also been investigated by looking at the type of vocabulary used in the tweets. There it was noted that scientists used a more informal language and style, including usage of multimedia features, compared to economists (Della Giusta et al., 2021). Differences between scholars from different parts of academia in regard to their use Twitter have also been detected, this by investigating the use of keywords and hashtags in searches as well as how they find new accounts to follow (Mohammadi et al., 2018). The activity on Twitter can also been measured by counting the number of clicks on embedded links. This has been done as a comparison between disciplines, where the clicks on links to Web of Science listed publications were investigated. It was seen that Social Sciences and Humanities, Biomedical and Health Sciences, together with Life and Earth Sciences, had more clicks than the STEM-disciplines investigated (Fang et al., 2021).

Related to interdisciplinary differences is if, and how, interdisciplinary communication occur. This kind of communication is considered generally rare (Ke, Ahn, & Sugimoto, 2017), however, recent survey investigations have concluded that knowledge transfer between disciplines do occur on Twitter (Mohammadi et al., 2018) and this might therefore be a growing practice.

### *2.2.2 Twitter usage*

One specific aspect of researchers' activity on Twitter are questions regarding the reasons why they are using the platform. Despite the fact that Twitter is part of what is generally known as social media, many researchers use the platform in their work. Mappings of the times when tweets are posted have shown that computer scientists use it during workdays (Schmitt & Jäschke, 2017). The borders between personal and professional is nevertheless, like in other social media, somewhat diffuse where engagement from the researcher is based on personal interest (Büchi, 2017; Sugimoto et al., 2017). Twitter has

been considered as being a possible help during all stages of the development of an article, including as a way to connect to old collaborators or find new ones, finding inspiration and methods for analyses in the form of articles – and finally as a tool to disseminate the final product (Britton et al., 2019; Côté & Darling, 2018; Klar et al., 2020).

It is not only collaborations for specific research projects that can be initiated via Twitter. The platform is also seen as a way to find a new employment (Álvarez-Bornstein & Montesi, 2019), and also as a tool for recruiting both students and staff (Britton et al., 2019). Twitter is considered a place suitable to promote oneself and one's work on, as well as building up, and communicate within, a network of colleagues in an easy and informal way. One of the advantages with Twitter often mentioned is this type of communication between peers, something that is considered to partly reduce the needs of conferences (Britton et al., 2019; Mohammadi et al., 2018). The conferences are however not forgotten on Twitter, a specific branch of Twitter research seems to be about the use of Twitter at academic conferences (Luzón & Albero-Posac, 2020 and references therein). These studies generally show that Twitter can be used as an announcement board in order to promote the conference, as a way to publish last minute changes as well as live-tweeting the talks on the conference. This area of Twitter research also includes studies showing how material tweeted with the conference participants as the intended audience can be spread and disseminated to a broader audience and the general public (Letierce, Passant, Breslin, & Decker, 2010). Based on the vocabulary used, the predominating results are however that Twitter is a communication tool for the group attending the conference, or the colleges unable to attend.

The possibility to get up-to-date information via Twitter is not limited to the data originating from conferences. That the platform facilitates easy sharing of articles, with and without comments, is something appreciated by the users both within and outside the academia (Álvarez-Bornstein & Montesi, 2019; Mohammadi et al., 2018). More about how Twitter is used as a tool for research dissemination will be discussed in section 2.3 below, but its recommendations and comments opens the scientific discussion to a wider audience (Büchi, 2017).

### *2.2.3 Altmetrics and Twitter*

Altmetrics has been suggested to be both a complement and a replacement for the traditional bibliometrics. The concept has its own manifesto where it is described as a faster and more inclusive way to both recommend and interact with scientific results (Priem, Taraborelli, Groth, & Neylon, 2010). These types of communication take place on different platforms and via features available on Internet, such as blogging, bookmarking,

recommendations, data sharing, comments and other similar activities, for extensive reviews see Sugimoto et al. (2017) and Tahamtan and Bornmann (2020). Altmetrics measurements can be in the form of views, downloads, mentions and similar actions. The service altmetrics.com calculates an altmetric score for scientific publications in an attempt to summarise the social media attention (Costas et al., 2015). It has been suggested that in order for any altmetric measures to work it is necessary to include some sort of field-normalisation, this since both the general interest and size of the academic audience varies between disciplines (Tahamtan & Bornmann, 2020). Also time has been highlighted as a factor that influences the altmetric score; most articles getting this kind of attention are published the last ten years (Costas et al., 2015).

The connection to bibliometrics have, as already mentioned in the introduction, resulted in several studies investigating the presumed connection between altmetrics score and number of citations (Costas et al., 2015; Thelwall et al., 2013) The changing nature of social media and time as an important factor for altmetrics are acknowledged problems, and it has also been suggested that older articles, published pre-Twitter should be compensated in some (yet unknown) ways when ranked on search pages (Thelwall et al., 2013). With the caveat that the study presenting the statistics, Costas et al. (2015), is on the older side, Twitter is the social media platform that generates the largest amount of data while calculating altmetrics scores. This means that it is more common that an article is mentioned on Twitter than on Facebook.

One part of altmetrics that it is regarded more as a score of popularity and actuality than impact, can be seen as a negative aspect, but also highlighted as an opportunity to use it as a measurement of interaction with the general public. Altmetrics will then be one of the metrics to use in order to estimate the societal impact of science, something sought after from funding agencies (Tahamtan & Bornmann, 2020).

Twitter was however in one study concluded to be of little value while measuring societal impact, in contrast to Wikipedia mentions that was concluded to be more relevant (Bornmann, Haunschild, & Adams, 2018). Also in other studies it has been shown that scientific articles are mainly shared by, and among, researchers (Tahamtan & Bornmann, 2020). This is in line with results showing that platforms like Mendeley and Figshare, that sometimes are included under the social media umbrella, have a clear academic user focus and are not suitable in order to measure research dissemination (Álvarez-Bornstein & Montesi, 2019).

## 2.3 Dissemination, vocabulary and social media

Research dissemination and communication occurring via social media in general, and on Twitter specifically, have been studied for several disciplines. Often the focus is on polarised topics full of opinions such as vaccines (Radzikowski et al., 2016) or climate-change (Moernaut, Mast, Temmerman, & Broersma, 2020). These studies often include network and argumentation analyses of Twitter-conversations, due to the dichotomized nature of the topics. Those are large fields in themselves and since neither these topics or methods are used in this study it will not be included and discussed in this part of the literature review. Instead a more general pattern of online research communication will be given, including examples of how dialogue and discussion can be achieved, but not covering the aspect of controversies and debates. This section will start with an overview of the concept of research dissemination and communication, followed by a section discussing the importance of awareness of the presence and problems with jargon and specialised vocabulary in communication. The third part will cover literature focusing on advantages and disadvantages with using social media as a tool for research communication.

### *2.3.1 Research dissemination and communication*

As already mentioned in the introduction is it considered important to see research communication as part of a dialogue, something that also requires an adaptation of vocabulary in order to increase the awareness, interest and enjoyment of science within the general public (Burns, O'Connor, & Stocklmayer, 2003). The concept of seeing dissemination as an exchange has however not always been the predominant view; the audience of lay-people has long time been seen as a group lacking both knowledge of and interest in the subject. Focus was at that time on one-way communication and education (for a review see Hargittai et al., 2018). This change of view of what research dissemination is and should be also includes one aspect that it is important to see popularisation of science as an adaptation of the results. The result should be changed and presented in a way that makes them suitable in a different context, not just as a simplification. Publication in media that enable comments and other informal communication is considered to facilitate this process (Burns et al., 2003; Luzón, 2013).

The emergence of social media has changed the practice of research communication, and the focus on informality, spontaneity and interaction seems to be key features (Della Giusta et al., 2021; Freddi, 2020). It has also been shown that science bloggers are utilising the linking possibilities embedded in the media in order to add extra information and explanations to the texts. These explanations can be both links to Wikipedia articles that

explains basic concepts, as well as to other scientific articles in support of opinions and claims written in the text (Luzón, 2013).

### 2.3.2 Vocabulary and jargon

A specialised vocabulary is sometimes called *jargon*, *technical words*, *academic words* or *low-frequency vocabulary*. What defines and distinguishes these concepts are discussed in different research areas, including linguistics and language teaching with focus on English for academic or special purposes. While attempting to create a universal academic vocabulary, in order to help students in their reading and writing, it was possible to see interdisciplinary differences in both choice of words and slightly different meaning of the same word. With this lack of possibility to generalisation instead the importance of context and the choice of suitable words for the intended public was stressed (Hyland, 2007).

In relation to this several studies have analysed the frequencies of different words in the English language and divided them into groups necessary for comprehension at different levels. It should be noted that there is a difference between understanding written and spoken language. It is estimated that one is required to understand around the 3000 most common word families for spoken English, while knowledge about up to 8000 word families might be necessary for a written text such as a newspaper. This has led to a division of word families into high-frequency, mid-frequency and low-frequency (Schmitt & Schmitt, 2014). These kind of frequency calculations are the base for tools like the jargon-detectors described in the method section below (Rakedzon et al., 2017; Willoughby et al., 2020), and it is important to note that the mid-frequency words are often overlapping with what in other studies is called the academic vocabulary.

The importance of a vocabulary relevant for the context can not be underestimated. Studies have shown that scientific jargon limits the possibility to understand the described data (Bullock et al., 2019).

Besides staying away from jargon there are also other linguistic features one can use in order to be inclusive. As mentioned above openness and two-way communication are important factors, and those can be achieved by choice of words (Freddi, 2020; Luzón, 2013). What factors in a text that can be markers for interaction with the readers and by that promote proximity are studied within the metadiscourse framework, something that will be further described in the theory section. One main common feature in research communication is the focus on the novelty of the findings, many texts starts with terms such as *new*, *recent*, *just published*, something they have in common with text written by journalists (Hyland, 2010).

### *2.3.3 Research communication on social media*

Scientists often mention one main asset that Twitter has when it comes to research communication, that it enables broadcasting of the research results by the scientists themselves (Klar et al., 2020). The possibility of communication with the readers and an interesting general public without a journalistic filter is seen as one advantage (Peters, Dunwoody, Allgaier, Lo, & Brossard, 2014). It should be noted that network analyses have shown that despite these opportunities news media is present on the platforms, and the information structure is similar to traditional ones with few central players (Büchi, 2017).

Survey studies have revealed that a majority of the respondents agreed with the statement that tweeting an article is part of its dissemination process, even if it is not proven to what degree a general public reads such tweets (Mohammadi et al., 2018). Other studies have shown that it is mostly other scientists, mainly from the same discipline, that follow scientists accounts. One interesting finding was however that the type of followers depends on the amount of followers. Accounts with many followers attract an audience from all parts of the community, including media and decision-makers (Côté & Darling, 2018).

The content of a tweet is also something that influences how attractive a subject is for the general public. Tweets with a scientific content combined with an emotional component, or tweets that can be related to current social and political issues, generate both more likes and retweets in comparison to tweets lacking such aspects (Denia, 2020). This focus on emotions on Twitter was also visible in a vocabulary comparison between how scientific topics are discussed on Twitter and online news respectively (Büchi, 2017). Analyses of how often links embedded in tweets are actually clicked on further highlight the fact that Twitter contributes to the online visibility of scientific articles but that in many case it does not evoke enough interest to continue reading (Fang et al., 2021).

It has also been shown that the keywords added to scientific articles can influence the possibility for the article to be tweeted about, articles with general keyword had more tweets than those with jargon ones (Haunschild, Leydesdorff, Bornmann, Hellsten, & Marx, 2019).



### 3. Theory

Lay summaries are discussed by Kuehne and Olden (2015) as one way to invite the community to a public dialogue about results, and this practice would require that the scientists looked at their work from the outside and adjust their vocabulary thereafter. This has later been tested using the De-Jargoniser tool, which showed that the amount of rare words, considered jargon, was still high in the texts that are aimed at a general audience (Rakedzon et al., 2017). Despite of this drawback, the suggestion to introduce lay summaries points to an assumption that also written text has an audience, and that we address different audiences in different ways. One way to approach this phenomena is to use the framework *metadiscourse*. Hyland (2005a) describes the rationale behind the development of metadiscourse in the late 1950s as “a way of understanding language in use, representing a writer’s or speaker’s attempts to guide a receiver’s perception of a text.” (Hyland, 2005a, Section 1.1 , para 1).

#### 3.1 A brief overview of metadiscourse

Metadiscourse is today considered the main approach while studying texts written by academics and other specialists, and a corpus is often used in the analyses, this compared to other related theoretical frameworks like metapragmatics that use ethnographic and sociolinguistic methods (Hyland, 2017). Despite, or maybe because of, its popularity the definition of where the borders around what really is metadiscourse are debated. An overview of different theories within the metadiscourse framework can be found in Ädel (2006, chapter 7). It has been possible to distinguish two different traditions – the broad and the narrow definition of metadiscourse, where the latter is sometimes called *metatext* instead of metadiscourse and covers only the textual function of the text, while the broader also include an interpersonal function. These two views are sometimes called the interactive (integrative) and the reflective (non-integrative) model (Ädel, 2006; Ädel & Mauranen, 2010) where the former focuses on the interaction between the writer and the reader. Not surprisingly the validity of such distinctions within metadiscourse research tradition has also been discussed (Hyland, 2017).

The concept metadiscourse can be used to describe both the framework and methods used to study a text, as well as being used to label the aspects of the text, the specific words and phrases that signal interaction with a presumed

reader. The use of the term in the latter sense is something that differs between the interactive and reflective model mentioned above, what kind of expressions and wording that is considered marking a proper metadiscourse event within a text (Ädel & Mauranen, 2010). In this context metadiscourse can be described as follows: “the linguistic material which does not add propositional information but which signals the presence of an author”<sup>1</sup>.

Although often large-scale, frequency-based analysis are conducted, it is necessary to remember the importance of the context of the word, and this at different levels. Metadiscourse is a framework that studies the relations between the writer and its reader in a specific context; the writer is aware and familiar with the audience while formulating the sentences (Hyland, 2005b). Context is present in a direct way in the analyses where specific words have different meaning depending on the neighbouring words, highlighting the importance of looking behind the word frequency lists (Hyland, 2017).

### 3.2 Stance and engagement

Following the broader, interactive approach it is possible to identify and classify words and phrases into metadiscourse groups that guide the readers through the text in different ways. One such group, is as example, “the text connectives” including sentence elements such as: *however, first, as noted earlier, in the next section*. According to Vande Kopple, (1985) these connectives are one of seven types of metadiscourse that can be found in a text. Not all of them will be discussed in this study it; instead focus on the stance and engagement model presented by Hyland (2005b) as an attempt to identify and classify signs of acknowledgement of the readers in academic text, and develop a tool suitable to study such interaction. The following paragraphs are all based on the article by Hyland (2005b), if no other source is indicated.

As the name implies the model consist of two parts, the *stance* in which the author is presenting oneself and declaring competence, and the *engagement* where the reader is recognised. The metadiscourse markers for stance can be divided into four group: *hedges, boosters, attitude markers* and *self-mention*. The hedges have been described as a way to recognise that other opinions in the subject also exist – and also inviting others into the discussion. Expressions showing that the presented results are based on reasoning and not direct facts are often used, such as *suggest, appear, indicate, and likely*. Hedges have been seen to be common in all disciplines, but are more frequent in what Hyland calls the soft disciplines. The next group, the boosters can be seen as the opposite to the hedges, the writers are displaying

---

1 This quote has been attributed to Vande Kopple (1985) in several instances, it is however not possible to find this exact phrasing in that article. Nevertheless is the sentence informative and therefore used here.

confidence and certainty, and uses words like *demonstrate*, *prove*, *clearly*, and *must*. The attitude markers are placed in the text in order to signal a more emotionally component. When these kinds of expressions are used a connection with the readers is created. Scientific results can be described as *remarkable* or *logical*, and a course of events can be labelled as *unfortunately* or *hopefully*. The final group of words describing stance is self-mention, and here it is possible to detect clear interdisciplinary differences. In line with respective academic tradition are social sciences and the humanities more frequently using the first person pronouns compared to science. If the self-mention classification is extended to also include *we*, *us*, and *our*, it is also possible to detect this practice with the STEM-disciplines. While analysing these kinds of texts it can sometimes however be difficult to distinguish a *we* as in “we in the research group” from a *we* as the writer and the readers. The latter is namely one of the five groups describing how engagement can be expressed in a text.

The engagement part in the model pinpoint the features included in a text in order to evoke engagement and interaction with the readers. Hyland distinguish five ways that this can be achieved in; *reader pronouns*, *personal asides*, *appeals to shared knowledge*, *directives*, and *questions*. The reader pronouns can both be an inclusive *we* as described above, or by directly addressing the reader by using *you/your*, something not common in academic text except for specific disciplines but more frequent when communicating popular science (Zou & Hyland, 2019). Personal asides are comments included in the text about what just been written, these kinds of expressions are however rare in both academic articles as well as in blogs (Zou & Hyland, 2019). The concept appeals to shared knowledge are more common even if disciplines within the STEM-sciences rarely express it explicitly. Words expressing this shared knowledge can for example be *conventional*, *established*, *traditional*, and *familiar* (Hyland & Jiang, 2016). In the directives group directions are given to the reader, and these can occur at different levels. They can be directions on a physical level like laboratory instructions, or guidance of the reader through different sections of the text. Also more cognitive directions helping the reader to follow a line of reasoning can be present. This last way of giving direction is the most common and markers of this in a text can be *note*, *consider*, *suppose*, and *compare*. The final group of markers used to create a connection with the readers are to include questions, sometimes rhetorical, in the text. This practice has decreased in the social science but more than doubled in biological papers (Hyland & Jiang, 2016).

Taken together the stance and engagement model is one interesting approach that can be used while investigating communication between author and reader in written text.



## 4. Material and methods

As described in the literature review the use of Twitter can be analysed in several ways and originate from interest in both the tweets themselves and the users. In this study both quantitative and qualitative methods were used while analysing a large dataset of tweets from topics connected to what sometimes is called the Life Sciences. Biomedical and health-related sciences are often pinpointed as specific areas where science can have an impact on people's life (Tahamtan & Bornmann, 2020), and are therefore both interesting and suitable to study in the context of research dissemination and communication. Different aspects of health, for example digital health care and health informatics, have also before been of interest and discussed in papers from the field of Library and information sciences (Haunschild et al., 2020; Larivière, Sugimoto, & Cronin, 2012).

The collection of tweets analysed in this study are parts of datasets collected by the consortium Data4Impact (Data4Impact, n.d.). The choice to use an already existing dataset can in some aspects be a limitation, since it is not possible to use the optimal approach based on the research questions to collect the data. The data was however retrieved by a research consortium, as described in section 4.1, and has been used in other research project – so there are no concerns regarding the validity of the approach used in collecting the tweets analysed in this study.

Combinations of manual and automatic classifications are common while investigating social media and its role in research dissemination (Denia, 2020; Freddi, 2020; Hyland, 2010; Zou & Hyland, 2019, 2020). I chose to follow this tradition, and a combination of word frequency analysis and a more qualitative approach was used in this study. The results were analysed and discussed in relation to linguistic models. In addition to this a quantitative analysis of tweets with embedded links was conducted, together with a content analysis of webpages the links pointed at and the content of the tweets themselves. This approach is also in line with previous research (Nelhans & Lorentzen, 2016; Thelwall et al., 2013). However, even if the same methods as these were utilized the content analysis itself was performed using an inductive approach. In this inductive approach, sometimes called a qualitative content analysis and in contrast to a deductive, is the coding scheme developed during the coding procedure (White & Marsh, 2006).

All methods used will be described in more details in the sections below.

## 4.1 Data collection

The datasets used in this study were generated by the Data4Impact consortium in order to be analysed as part of their research projects. The main aim for this consortium was to investigate the impact of research within different parts of society. This with the background that large amounts of money are distributed to researchers via the EU H2020 programme, and a Big Data-approach such as Data4Impact could be one way of monitoring and estimating the effects of this research funding. By using different types of data already available on the web it was considered to be possible to perform such evaluation at a large scale (Pukelis & Stanciauskas, 2018). Impact of the EU-funded research was investigated in three dimensions: the economic, the academic and the societal, and all within research areas connected to Health, Demographic Change and Well-being Societal Challenge (Feidenheimer, Frietsch, Schubert, & Neuhäusler, 2018).

The tweets within the datasets were collected based on keyword queries; these keywords were generated by extracting terms from finalised EU projects. These terms were then grouped together based on co-localization in the texts in order to generate the queries for each topic. For a detailed description of this selection method see Nelhans, Papageorgiou, Pukelis, and Demiros, (2020) and references therein. A full list of query-terms used for the topics investigated in this study is available in appendix table A1. The tweets were collected between February 2<sup>nd</sup> and March 3<sup>rd</sup> 2019, and the query-terms can occur in the tweets themselves as well as in the URL or the metadata.

## 4.2 Datasets

Data4Impact collected tweets from several scientific, health related topics within the Life Sciences. Scientists have in earlier studies displayed tweeting practices including elements of an informal tone and other signs of attempts to outreach (Della Giusta et al., 2021), which are examples of features interesting for this study.

Four subjects were selected for analyses in this study, two of them contain scientific names, one of the model organism *Caenorhabditis elegans*, and one of the bacteria *Staphylococcus*. The third topic is also of a more specialist type, mRNA expression, while the fourth consists of words such as species diversity and phylogenetic tree – a vocabulary that might indicate a collection of tweets aiming at a more general public. This sampling was done with the aim to generate a good mix of topics and a base for interesting analyses and results.

The Twitter data is collected in areas I am more or less familiar with from my previous professional career. This is an asset in this study, and might also introduce bias in some instances. Hopefully an extensive and transparent

description of the methods used will help to avoid any problem in that regard. The potential advantages of having knowledge about the scientific topics investigated have been discussed in earlier research, in the context of content analysis of tweet from different disciplines (Holmberg & Thelwall, 2014), as well as while using co-occurrence of words in order to identify the subject of the tweets (Büchi, 2017). Based on my experience and expertise I have the possibility to describe the chosen topics and speculate about what kind of results they may give, concerning the topics' inherent attraction and availability to a general public.

#### 4.2.1 *Caenorhabditis elegans/Neuromyelitis*

As mentioned above *C. elegans* is a model organism, it is a nematode – a transparent, one millimetre long, wormlike creature with short generation time and suitable for, among other things, investigating organ development. Research using this organism was rewarded with the Nobel Prize in Physiology or Medicine in 2002 (NobelPrize.org, 2002). Neuromyelitis optica, the other query term for this topic, is a rare neurodegenerative autoimmune disorder that affect the optic nerves, and is also called Devic's disease (NHS, 2020).

The fact that this disease is paired with the query term *C. elegans* indicates that some of the finalised EU projects (used as a corpus while extracting the query terms, see section 4.1), have used the model organism while studying the disease. It is however important to remember that the query terms are formulated with “OR”-statements, co-localization of the terms within the tweets was not required. This sampling method can therefore result in tweets only containing the *C. elegans* term, without connection with the health aspect and the general public might therefore not have been the targeted audience.

#### 4.2.2 *Staphylococcus*

*Staphylococcus aureus* is a bacterium, pathogenic to humans and causes a variety of diseases, including bacteria in the bloodstream, pneumonia and skin infections. It is estimated that up to 30% of the human population are carriers of *S. aureus*, primarily in the nasal regions. The bacteria is also known to be antibiotic resistant, in those cases often called MRSA, methicillin-resistant *S. aureus* (Gnanamani, Hariharan, & Paul-Satyaseela, 2017).

Taken together this suggests that tweets regarding this topic could be of interest for a wider audience, opening up for the possibility of a varied vocabulary despite the scientific nature of the query term.

### 4.2.3 mRNA expression

mRNA is one part of the transcription/translation mechanism and process that takes place in our cells, the other two main players are DNA and protein. It is possible to use a digital library as an analogy for how mRNA is related to DNA and protein. The cell nucleus can be seen as a library where all the information is stored as DNA, each book or journal as its own piece of DNA. When a user borrows a digital book, journal, or article it is the copies of the original file that are sent out from the library. In the case of the nucleus this is the mRNA, short for messenger RNA. It can be seen as a blueprint which describes how the finished product, the protein, should look like. Continuing with the digital library metaphor the protein is the result if the user chose to print the borrowed book.

The expression part of mRNA expression refers to how many copies that leave the library at a specific time point. A popular item, or course literature, might be borrowed 30 times during a couple of days, while a more obscure article only once. This is also true for the cell, some proteins, and therefore their mRNAs are in more demand than others, depending on what occurs in and around the cell.

Other query terms used while generating the dataset for this topic included variants of the word “splicing”. Splicing can be described as when only some chapters of the book are interesting. One user might download chapter 1, 2, and 5 and at the same time another user wants 2, 5, and 8. This book has been alternatively spliced and the same is true for mRNA, it is not always the full-length mRNA that is of interest and only part of it is expressed.

This is one example of how a subject can change over time. At the time of writing this master thesis in the spring of 2021 mRNA based vaccines against COVID-19 have been developed (Bettini & Locci, 2021; Jackson, Kester, Casimiro, Gurunathan, & DeRosa, 2020). This has resulted in an increased interest in, and maybe also increased knowledge of, mRNA and its expression. The general interest for this process was however not that large in the spring of 2019 when the tweets analysed in this study were collected.

#### 4.2.4 Species diversity/phylogenetic tree

This topic might seem diverse at a first glance, with query terms such as *phylogenetic tree*, *species diversity*, and *evolutionary history* and variations thereof. However, working in these areas it is obvious how often the concepts co-occur.

A phylogenetic tree is a way to illustrate relationships, where the genes, species or lineages are the leaves in the end of the branches, shown in figure 1 (adapted from Sundström, 2010). The tree is a hierarchical representation, comparable to an XML-file with parents, children and siblings, but are often visualising the evolutionary history of the items.

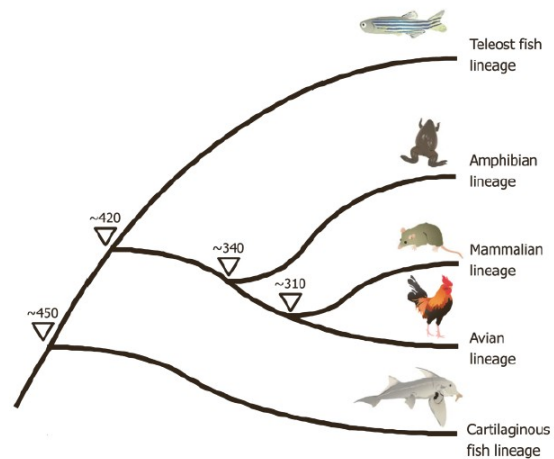


Figure 1: Schematic phylogenetic tree. Proposed divergence times in million of years marked with triangles.

The connection between species diversity and evolutionary history is discussed in many contexts, among others how the present and historic climate change influence species diversity (Theodoridis et al., 2020; Wilsey, 2020). How this process in turn will effect the human population (Considine, Siddique, & Foyer, 2017) gives this topic an actuality and relevance for the general public. This together with a general interest in fossils, dinosaurs and other things related to evolutionary history open the possibility that this topic contains a wide variety of tweets aimed at a diverse audience.

#### 4.3 Methods

The flow-chart below summarises the different analyses made with this dataset, each method will be described in more detailed below. The bottom right box indicate how the data will be analysed and discussed in the chosen linguistic theoretical framework.

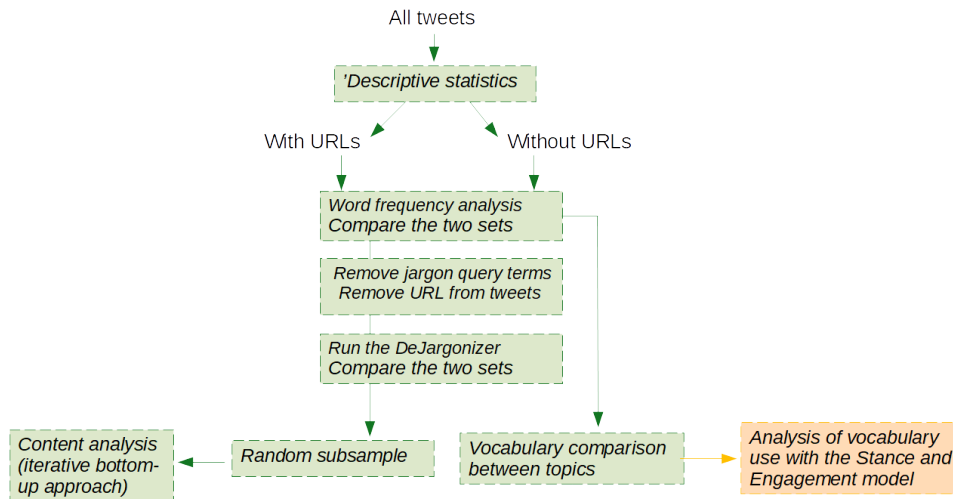


Figure 2: Flow-chart with the methods and the order they are used in.

The statistical analysis method Log-likelihood was applied throughout this study while conducting comparisons between different datasets and subsets. This method is suitable to use while comparing corpora (Rayson & Garside, 2000), and has been used in studies comparing use of engagement markers in academic blogs (Zou & Hyland, 2020). In this master thesis a log-likelihood calculator was used (Rayson, n.d.), which specific significance levels and thresholds used for this study will be discussed together with the results.

#### 4.3.1 Descriptive statistics and pre-processing

This descriptive step, including taking note of how many tweets were retweets, contained URLs or mentions, served as a way to get an insight in the characteristics of data available. It was also performed as an aid while separating the data in subsets for further analyses.

The division of tweets into those with and without URLs was done by sorting them according to the “expanded\_urls”-column in the spreadsheet containing the raw data. The communicative role of Twitter has been discussed by Büchi (2017). That study included descriptive statistics concerning number of retweets and mentions in order to analyse this phenomena. Inspired by this study these concepts were also included in this project. Using the spreadsheet sorting tool it was possible to note both the number of retweets, defined as tweets starting with RT, and number of mentions – the tweets containing an @.

The tweets were collected using sets of query terms as described above in section 4.2. These terms were checked for jargon content by using the DeJargonizer (the method described in more detail in 4.3.2 below) in the first of

two preprocessing steps. Presence of one or several query terms in the tweets were a requirement in order to be included in the dataset, and the query terms for each topic can contain one or several words classified as jargon. By removing these terms in this pre-processing step the differences, or noise, between the sets were levelled out before the comparison between the topics.

In the second pre-processing step the URLs were removed from the tweets, this due to the fact that pilot runs showed that “https” were marked as jargon. Keeping this term would disable the possibilities to compare the jargon content between tweets with and without URLs.

I did not remove the retweets from the datasets, since this study is supposed to give an overview of how the specific topic is presented and represented on Twitter. Keeping the retweets has implications; some tweets will be repeated many times and could therefore dominate the subject, and by that also the vocabulary. However, it is important to remember that this is a snapshot of how the specific topic was ventilated on Twitter during the month the tweets were collected in 2019 and a set collected any other month would probably give different results. By keeping the retweets a more complete picture of what was discussed will be captured, in comparison to the more cropped picture one would get if each tweet only occurred once.

#### *4.3.2 Word frequency analysis*

The vocabulary between tweets containing and lacking URLs, within each topic, was compared using a word frequency analysis. The analysis was conducted using R, and utilizing several packages from Tidyverse (Wickham et al., 2019). The procedure from the book *Text mining with R: A tidy approach* (Silge & Robinson, 2017) was followed with some adjustments. In this study for example all retweets were kept throughout the analysis. Before the calculation of word frequencies some special characters were removed together with stop words, and the dataset was tokenised using a tokeniser specialised on Twitter-data (Mullen, 2016). This adaptation of the tokeniser for tweets includes that hashtags and usernames, words preceded by an @-sign, are preserved. The mentions, the usernames, were not of interest in this study, and would only introduce unnecessary noise in the data, and were therefore filtered out in a secondary step.

#### *4.3.3 Jargon detector*

Jargon detectors are based on a principle of comparing the input text with a reference corpus. This results in that the choice of datasets to be included in the corpus is important and will influence the outcome of the analysis. The tool used in this study, the De-Jargonizer, had its corpus constructed by using the content from around 250 000 articles published on the BBC sites, with the addition of the American spelling of the words. Using this strategy the

developers get a contemporary list of words that also includes the scientific content presented on BBC (Rakedzon et al., 2017). The corpus is also continuously updated and the user interface allows for choosing a corpus from a time period appropriate for the text to be analysed.

The words in the corpus are divided into three classes: high-frequency that occur more than 1000 times in the corpus as a whole, a mid-frequency class with words appearing between 50 and 1000 times, and the rare-frequency words occurring less than 50 times that are classified as jargon (Rakedzon et al., 2017). However, as mentioned in the literature review, it is important to remember that many words in the mid-frequency class are considered belonging to an academic vocabulary in other studies (Schmitt & Schmitt, 2014) and therefore are of interest in this study. In connection to this it is important to remember that it is single words that are analysed and not phrases, which is something discussed by Willoughby, Johnson, and Sterman (2020) together with the importance of putting the word into correct context. Words like *vacuum* have different meaning if it is part of the phrase vacuum cleaner or in a scientific article. There exist jargon detection tools where it is possible to specify and flag phrases as jargon in the settings (Willoughby et al., 2020). However, this requires a familiarity with the type of text and its vocabulary, and not possible to apply on this diverse dataset.

#### 4.3.4 Content analysis

Ten percent of the total of number of tweets with URLs were chosen from each topic for further analyses. They were randomly selected using a random number generator based on atmospheric noise (Haahr, 2021), and thereafter coded manually in two different aspects using an open, inductive coding approach. The URLs were opened in a browser, and for any dead links a new link was randomly drawn, this under the assumption that all links were working at the time of the writing of the tweets. It should be noted that some tweets contain two or more URLs; when the input spreadsheet for this dataset was constructed only one of the URLs was parsed into a separate column. This column was used in order to collect the URLs, however both URLs are available in the full text of the tweet. The second included URL links in several cases to another tweet. This might be the result of using Twitter's relatively new quote feature. This is another type of interaction, other than the retweets, although not investigated in this study.

The first part of the coding noted the type of source the URL was pointing towards, using an iterative bottom-up approach, inspired by the method described in Nelhans & Lorentzen (2016). The main procedure was to keep the coding at a more detailed level at first and then summarise these into broader categories which are presented as the results.

The second aspect coded was in order to depict the content of the tweet. The same inductive coding, without pre-defined labels was conducted also in this phase, and just as in the first part an iterative element was included. When a “new” label appeared in any dataset, the previous ones were rechecked, with focus on the additions. This whole two-layered process of coding is similar to the one used by Holmberg and Thelwall (2014) in a study investigating interdisciplinary differences in the use of Twitter.

In order to check for intra-rater reliability ten percent of all coded tweets were analysed a second time one month later. The results from the first and second coding events were compared and the Intraclass Correlation Coefficient (ICC) was calculated as a measurement of the reliability. Following the guidelines presented by Koo and Li (2016) the two-way mixed effect model with single measurement and absolute agreement was calculated using the ICC function as implemented in the *psych* package in R (Revelle, 2021). The two-way mixed model only evaluates the reliability of the specific rater(s) that performed the coding in the current study, and the results can not be transferred to other raters of the same dataset (Koo & Li, 2016).

#### 4.3.5 Vocabulary analysis

Even though word frequency analysis adjusts for the size of the document only the vocabularies of the two largest topics, *mRNA expression* and *Species diversity/Phylogenetic trees*, were compared in this part. As illustrated in figure 2, section 4.3, tweets without any URLs were selected for this analysis.

The datasets were analysed in the same way as described in 4.3.2, by using the Tidyverse package in R, with the exception that stop-words were not removed. This since many of the words signalling stance and engagement, and are the focus of this analysis, are considered stop-words. In order to compare between the topics as well as with previous research the frequencies were calculated as by 1000 words.

Previous studies have used the *Stance and engagement* model while analysing both academic text as well as blogs from different research disciplines (Zou & Hyland, 2019; Zou & Hyland, 2020) As described in section 3, Theory, the *Stance and engagement* model consists of two parts, the *stance* where the authors of the text put themselves and their opinion into the text. The ways to express *stance* is something that has been shown to differ between texts aimed at a professional or a more general public. Hedges have been found present in both article and blogs, but are more frequently used in blogs, something that has been interpreted as a response to the presence of commenting fields directly below the post. The possibilities to discuss the topic directly changed the way the results were presented (Zou &

Hyland, 2019). Table 1 (and Appendix table A2) below contains a selection of expressions of *stance* detected and analysed in Hyland, (2005b) and Zou & Hyland (2019) combined with the listed stance features in appendix B in McGrath & Kuteeva, (2012). The words in these lists also present in the topics *mRNA expression* and *Species diversity/Phylogenetic trees* were placed in table 1 and the frequency of each type were calculated.

*Table 1: Types and examples of “stance”.*

<b>Type</b>	<b>Example</b>
Hedges	Suggests, possible, likely, indicate, apparently, appear, knowledge, perhaps
Boosters	Extremely, obviously, demonstrate, prove, certain, sure, must, clearly, indeed, claim
Attitude markers	Promising, agree, disagree, important, interesting
Self-mention	I, we, our, us

The other part of the model, the *engagement*, refers to markers where the authors connects, engage, the reader of the text. To include a reader mention is the most common way to make this connecting; those mentions are more common in academic blogs than in articles, and also more common in the non-STEM disciplines (Zou & Hyland, 2019; Zou & Hyland, 2020). This recognition of the reader can take place in different ways, and markers for those can be categories as shown in the table below, summarized from Hyland, (2005b), Zou & Hyland (2019), McGrath & Kuteeva, (2012) and appendix 2 in Hyland and Jiang (2016), and selected based on occurrence the same way as for the different types of stance. Since the list in the appendix is so extensive only terms with five or more occurrences together in the two topics were retained for analysis.

*Table 2: Types and examples of "engagement".*

<b>Type</b>	<b>Example</b>
Reader mentions	we/our/us, you/your
Directives (to instruct the reader)	Should, go, analyse, must, consider, demonstrate, consider, add *
Questions	?
Shared knowledge	Usually, apparently, common, obviously, traditional, integrate, routinely

\*these are the most common words detected in the datasets, a full list for all words used as markers for engagement is available in table A3 in the Appendix.

For the type *questions* the frequency of the number of tweets with question-marks was calculated, instead of word frequency. Besides the types listed in table 2 the type *Personal asides* is included as markers for engagement. This type was excluded from this analysis, due to the fact that a specific content analysis would be required to investigate these aspects which was unfortunately not doable within the time-frame for the study.

In relation to the stance and engagement model is also the concept of creating proximity to readers, and especially the detected differences while addressing a professional or general audience. The second part of the quote below describes an interesting aspect that was further analysed in this study.

While it embraces the notion of interpersonality, proximity is a slightly wider idea as it not only includes how writers manage themselves and their interactions with others, but also the ways ideational material, what the text is ‘about’, is presented for a particular audience. (Hyland, 2010, p 117)

On common denominator for both the scientific articles and the more popular texts were the importance highlighting novelty (Hyland, 2010). Based on these results, the following terms were investigated and compared between the two topics: *novel/novelty*, *advanced/advancing*, *important/importance*, *recent/recently*, *new/newest*, *discover/discovering/discovery* (these words are also a listed in table A4 in the appendix).

#### 4.4 Ethical considerations

As mentioned earlier the dataset used in this study is part of a larger collection of Twitter data selected, retrieved, used and archived within the frames and ethical guidelines for the research consortium Data4Impact. I have only access to the parts of the information required for this limited study, something that follows recommendations regarding how to handle Twitter data in order to diminished the ethical implications (Ahmed, Bath, & Demartini, 2017).

Nevertheless, I have in this study excluded the usernames in the examples, this based on research showing that even if most Twitter-users feel indifferent or positive with the fact that their tweets can be used in research, most feel unease that their username should be shown in order to exemplify findings (Fiesler & Proferes, 2018). This kind of anonymisation is not always practised, as example are the usernames included in a studies by Luzón and Albero-Posac (2020) and Klar et al. (2020). That latter analysis was however investigating the use of Twitter while promoting research articles, a subject where the participants have used Twitter in outreach activities and a wide dispersion of the tweets was the original intent. The examples used in my

study are also from tweets with a similar intent, to spread information regarding recently published articles, courses and cures. All examples were carefully chosen in order to be informative, but not intrusive or compromising to the authors in any way. However since I have not followed conversations or performed argumentation analysis this have not been a problem. This links to a continuing and old discussion about if data published on the Internet should be consider to reside in the public domain (Pace & Livingston, 2005). It has also been highlighted that the Twitter user agreement states that information can be used of third parties, something that has been used as an argument to bypass the practice of informed consent. Due to the large datasets often analysed in these type of studies is it also often impossible to achieve the informed consent (Ahmed et al., 2017).

It should be noted that it is only via the texts in the tweets themselves that I as a researcher have access to Twitter usernames. The usernames are only presents in the mentions, texts where a user are addressed by the @-sign. Usernames for the authors of the tweets are not available in the dataset, all personal information was encrypted before the data was available for me and each tweet and user have an anonymous ID-number instead.

Both older and newer studies have concluded that using data from the Internet require some ethical discussion. If a general principle of protecting the integrity of the participants are followed this kind of data can be used in research.

## 5. Result and analysis

### 5.1 Descriptive statistics and pre-processing

The four chosen topics display slightly different characteristics; how the number of tweets within each subset differs is visible in Figure 3. The topic *C. elegans/Neuromyelitis* clearly contains the smallest number of tweets and also has a larger proportion of tweets with URLs in comparison to the others.

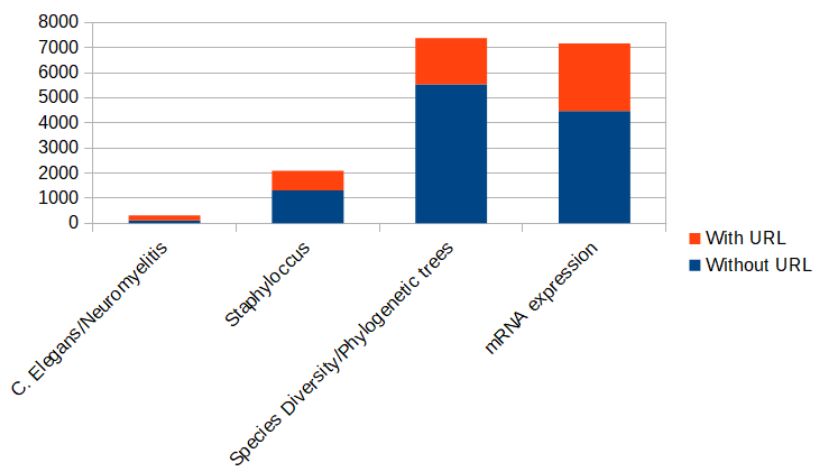


Figure 3: Distribution of tweets, with and without URLs, in the investigated topics.

The topic *C. elegans/Neuromyelitis* is constructed by retrieving tweets with the, at the first glance, disparate query terms: "caenorhabditis elegans" OR "nematode elegans" OR "neuromyelitis optica". The frequency of these words within the tweets were investigated together with the prevalence of co-occurrence between them. The term "neuromyelitis optica" occurs in 36% of the tweets without URLs, and in 32% of those containing URLs. However, the term never coincides with any of the terms describing the nematode. This gives that the topic called *C. elegans/Neuromyelitis* in this project probably could be considered two separate topics. It will nevertheless stay as one topic due to the fact that the tweets were retrieved as part of another project, with different objectives. The topic *Species diversity/Phylogenetic trees* is also a combination of what can be query terms without connection to each other. But as described in the methods section 4.2.2 these terms are interlinked and can be, and are often used in the same kind of scientific research.

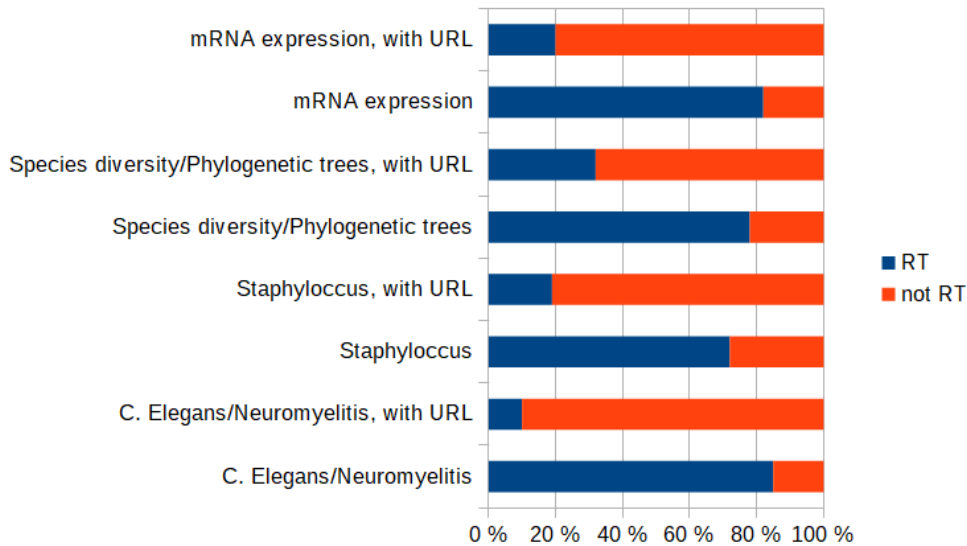


Figure 4: Percentage of tweets that are RTs (retweets), in all topics.

The number of retweets, RT in each topic, for both the tweets with and without links, were also calculated and an interesting pattern was revealed and is depicted in figure 4. Despite the intuitive thought that Twitter users identify an interesting link and retweets it in order to spread it, the picture seems to be the opposite.

It is only between 20-30 percent of the tweets with URLs that are a retweet, while retweets make up around 80 percent of tweets lacking embedded links. It should be noted that it is the functional links that count as an URL in this dataset. If the link-address in the tweet is partial and broken the tweet will be categorised as “without URL”. However, for these four topics only around 10 percent of the retweeted tweets without URLs contained partial links, and this phenomenon does not explain the difference in number of retweets.

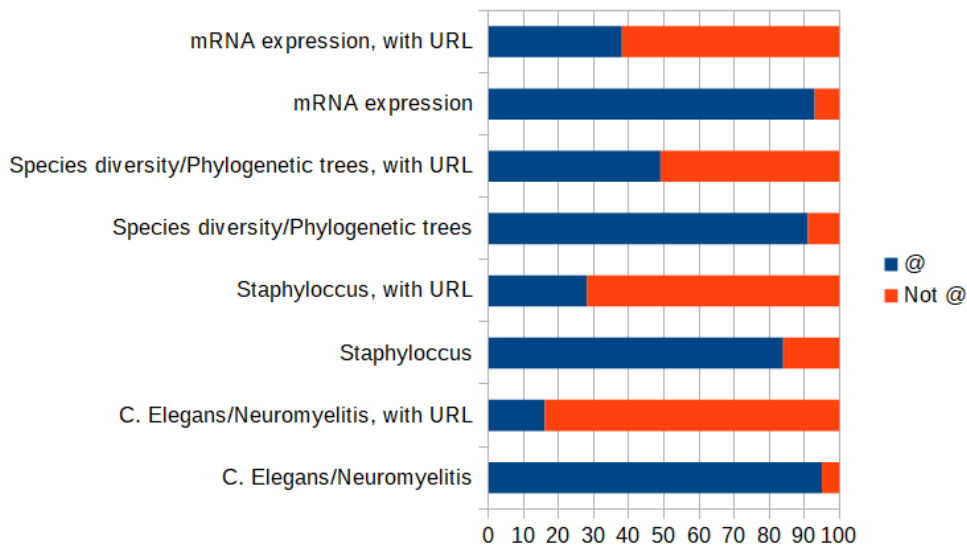


Figure 5: Proportion of mentions, @, calculated against total number of tweets.

The patterns for tweets with mentions follows the same pattern as for the retweets, which is not surprising since they are often used in combination. A retweet is often followed by a mention that gives the source of the original tweet. One tweet can also contain more than one mention, which is visible comparing figure 4 with figure 5, the mentions occur more frequently than retweets. The mentions can also occur without a retweet something that might also contribute to this number.

The first pre-processing step included using the De-Jargonizer in order to detect jargon among the query terms. All terms for the two numerically smaller topics *C. elegans/Neuromyelitis* and *Staphylococcus* were marked as rare: **caenorhabditis** **nematode** **elegans** **neuromyelitis** **optica** and **staphylococcus** respectively.

The tweets for the topic *mRNA expression* were collected using several query terms, for a full list see appendix table A1. As described in the method section these terms consist of variants of the same terms. Below is the result from the De-Jargonizer analysis of the non-redundant set of query terms. Words in the mid-range are in italics and rare in bold.

**mrna** expression levels *decay* **splice** site **splicing** *alternatively spliced* *alternative variant variants*, pre stability *transcripts messenger rna* binding *gene protein*

Note that it is not possible to use phrases as input in the analysis, which means that despite the fact that some of the query terms were combined as phrases, like “mrna expression”, the words will be considered as single

words. As has been discussed before, it is important to remember that this analysis is not taking the context into consideration when classifying the words as common, mid, or rare. The word “expression” is considered a common word but in the context of mRNA it probably had been more correct to place it in the mid or rare category.

The same caveats are of course present while creating the non-redundant set of query terms for the *Species diversity/Phylogenetic tree* topic. The result after the content has been run in the De-jargonizer is shown below, word in the mid-range in italics and rare in bold.

species *genus* **phylogenetic** tree trees *analyses* analysis  
relationships diversity related distribution *richness* *identification*  
specific **genera** *evolutionary* history

Comparing these results it is possible to see that the two topics with larger number of tweets, *mRNA expression* and *Species diversity/Phylogenetic trees*, contain less jargon in their query terms. This strengthens the decision to remove the query terms from the tweets before topic comparisons.

## 5.2 Word frequency analysis

The word frequency analyses can be seen as a mid-step between descriptive statistics and a more in-depth analysis. The figures below show the similarities and differences in word use between tweets with and without URLs for each subject. Unfortunately it is not possible to print all words in the comparisons in the figures, since they then will become unreadable. The picture makes it however possible to get an overview of the data.

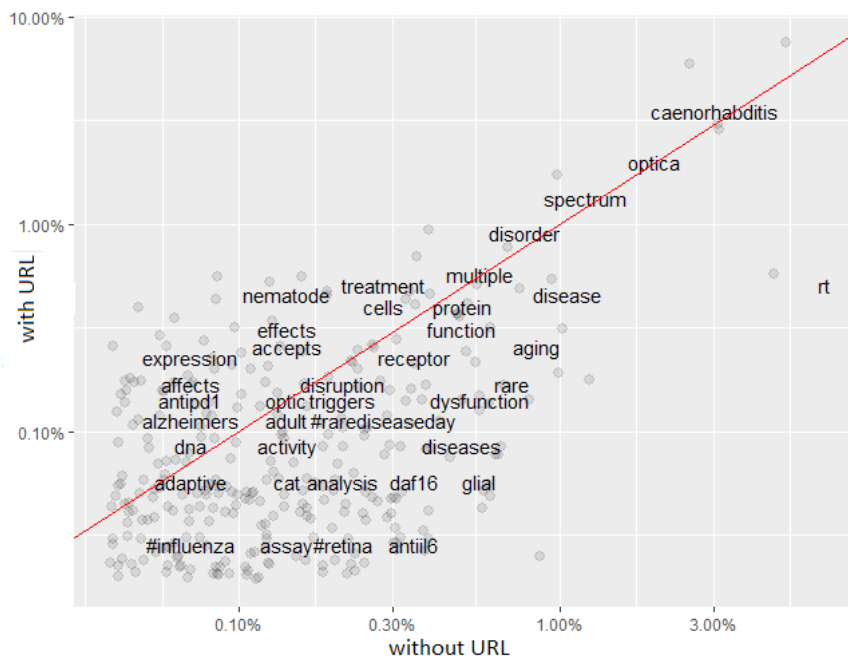


Figure 6: *C. elegans/Neuromyelitis*, word frequency comparison.

While interpreting figure 6 which displays the difference in word frequencies for the topic *C.elegans/Neuromyelitis*, it is important to remember that this topic consists of a low number of tweets and that the distribution between tweets with and without URLs is skewed, with more tweets containing links than not. The words along the red line are as common in both categories, and within these are the query-terms *caenorhabditis* and *optica*. In general most words are present in similar frequencies in tweets with and without links. It is also possible to detect the word “rt” in the middle to the right, again illustrating that the practice of retweeting is more common for tweets without URLs.

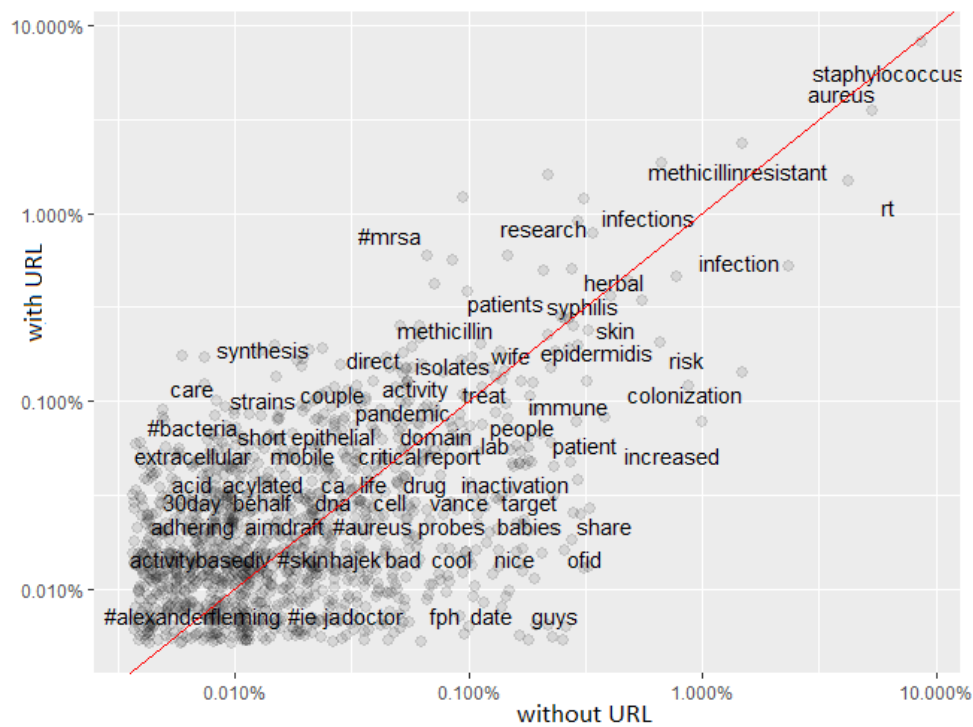


Figure 7: *Staphylococcus*, word frequency comparison

A similar pattern is visible for the topic *Staphylococcus* in figure 7. The query-term is present in both, since this also is a requirement for inclusion, and the retweets are more common in tweets without links. It is interesting to see that the hashtag *#mrsa* together with the terms *methicillinresistant* and *methicillin* are more common for tweets with links. Other words that generally can be considered as “scientific” are also more common in those tweets. It is nevertheless important to not only look at the frequencies but also at the tweets behind. The word *guys* has a higher frequency in the subset without links, however, all the tweets containing the word are retweets from a tweet that had a link, and the links are incomplete for the retweets.





Table 3: Number and percentage of words classified as common, mid or rare in the different topics.

	<b><i>C.elegans/Neuromyelitis</i></b>				<b><i>Staphylococcus</i></b>			
	Without URL		With URL		Without URL		With URL	
	No.	%	No.	%	No.	%	No.	%
common	1314	69	2217	63	15491	68	8816	60
mid	292	15	662	19	3173	14	2651	18
rare	312	16	663	19	4050	18	3148	22

	<b>mRNA expression</b>				<b>Species diversity/ Phylogenetic trees</b>			
	Without URL		With URL		Without URL		With URL	
	No.	%	No.	%	No.	%	No.	%
common	66101	71	35274	65	97362	79	27415	75
mid	16266	17	11082	20	18028	15	6220	17
rare	10659	11	7733	14	8105	7	3163	9

Results of the calculation of the Log-likelihood are presented in table 4. It is calculated as a pairwise comparison of both mid- and rare-frequency words, between tweets containing and not containing URLs within each topic.

Table 4: Log-likelihood for mid- and rare-frequency words

	<b>Log-likelihood</b>	
	<b><i>mid</i></b>	<b><i>rare</i></b>
<i>C. elegans/Neuromyelitis</i>	8.74	4.25
<i>Staphylococcus</i>	97.50	62.67
<i>mRNA expression</i>	164.87	216.35
<i>Species diversity/Phylogenetic trees</i>	97.00	158.90

The null hypothesis is that there are no differences between the different types of classes. The log-likelihood value should be compared to a critical value for the selected threshold of significance. The critical value 3,84 corresponds to the 5% level, and 6,63 to the 1% level. The results show that the differences seen both in the number and percentage of mid- and rare-frequency words between tweets with and without embedded links are statistically significant. This is true also for the topic *Species*

*diversity/Phylogentic trees*, despite the fact that this topic, compared to the other topics analysed in this study, clearly contains a lower amount of rare words.

Despite the discovered differences between the two types of tweets is the overall picture that all tweets about these topics use a vocabulary that can be considered full of rare words, and that at a level considered not understandable for the general public.

## 5.4 Content analysis

This content analysis was performed in two parts, first to identify which *type* of source the link was directing the reader to. The second part was conducted in order to put a label on the *content* in the tweet itself. Intra-rater reliability was measured using the Intraclass Correlation Coefficient (ICC) and the calculations were performed for each part separately. The *type* aspect received an ICC of 0,92 with 95% confidence interval between 0,88 and 0,95 and thereby classified as having good to excellent reliability, based on the interpretations suggested by Koo and Li, 2016. The corresponding values for the content aspect were an ICC of 0,78 with 95% confidence interval between 0,68-0,86, and should be considered to have moderate to good reliability.

### 5.4.1 Type of source linked to

Using the iterative bottom-up approach it was possible to divide the sources linked to into six broad types or categories: scientific article, science news, science related, news, social media and other. *Scientific article* is easy to distinguish while the category *Science news* includes, beside journalistic science news, also science blogs and news for medical professionals. *Science related* describes resource platforms for learning about science at all levels, including websites from university departments and conferences. This category focuses on education and does not include companies that sells laboratory equipment or medicine. *News* can be both newspapers and blogs, without an outspoken science focus. *Social media* is more straightforward with Facebook, Instagram and Twitter. Some tweets in the dataset also contain links to Youtube, the classification of these depends on who posted the video. The tweets often link to educational videos about the topic in question, and are therefore placed in the science related category. The *Other* category contains links to charity foundations, organisations, company webpages and web discussion forums, the latter without any clear scientific connection.

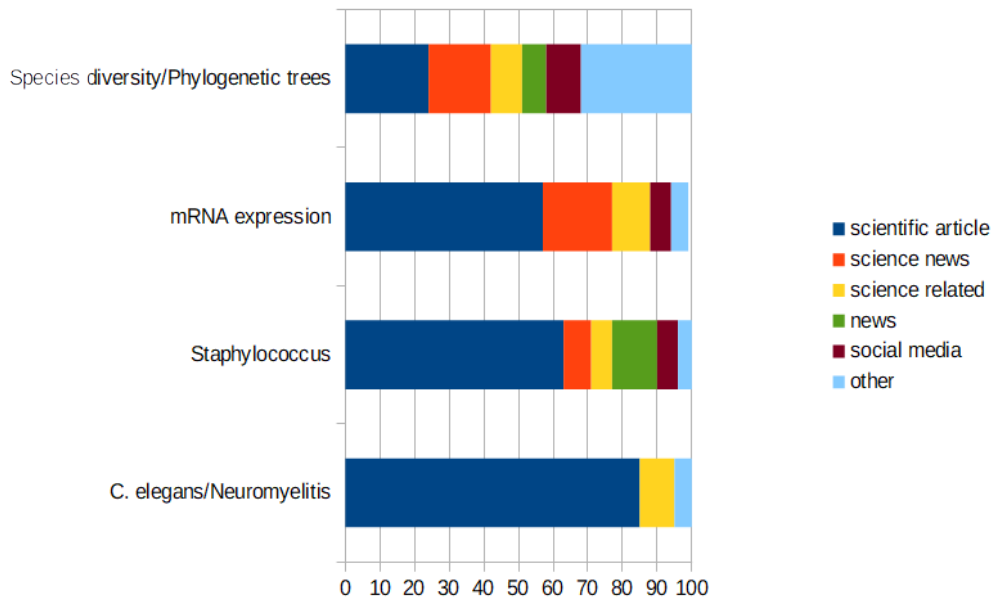


Figure 10: Distribution of the types of sources the tweets links to in the different topics.

Figure 10 displays the distributions of types of sources linked to in the different topics. These distributions are visualised as percentage of total number of investigated tweets in each topic. Note that the number of analysed tweets differs between topics:  $n=184$  for *Species diversity/Phylogenetic trees*;  $n=269$  for *mRNA expression*;  $n=79$  for *Staphylococcus*; and  $n=20$  for *C.elegans/Neuromyelitis*.

*Scientific article* is the most prevalent category, except for the topic *Species diversity/Phylogenetic trees*; in this topic most of the tweets instead point at sources in the category *Other*. This deviation from the other topics can be explained by a number of retweets within that category. One tweet is from the World economic forum regarding how a decrease in biodiversity threatens future food production. The full URL includes the query term “species diversity”, as shown in (1) and the tweet itself is formulated as in (2).

- (1) [https://www.weforum.org/agenda/2019/02/future-of-food-under-severe-threat-as-species-diversity-disappears-un?utm\\_source=Facebook%20Videos&utm\\_medium=Facebook%20Videos&utm\\_campaign=Facebook%20Video%20Blogs](https://www.weforum.org/agenda/2019/02/future-of-food-under-severe-threat-as-species-diversity-disappears-un?utm_source=Facebook%20Videos&utm_medium=Facebook%20Videos&utm_campaign=Facebook%20Video%20Blogs)
- (2) RT @wef: Wildlife: ↓ Biodiversity: ↓ Human population: ↑   
Read more: <https://t.co/s4te2PY6lB> #environment #sustainability  
<https://t...>

In total 184 tweets (10%) were analysed in this topic and 31 of them were retweets of either example (2) or a related tweet, from the same source and with the same message in a different wording. The category *Other* in this

topic also contains examples of retweets that do not include “RT” in the text and is instead just a copy and paste of original text and link in a new tweet. This means that these tweets are invisible in the descriptive statistic presented above in section 5.1, but possible to detect while manually analysing a subset like this. Together this formal and informal retweeting practice adds up to more than half of the numbers of tweets in the *Other* category in this topic.

Overall the distribution pattern is similar for all the chosen topics, it is however only two topics that seem to be interesting enough to be covered in more traditional news media – and therefore gives the possibility to link to such a source, namely *Species diversity/Phylogenetic trees* and *Staphylococcus*.

#### 5.4.2 Content of tweets

As seen in figure 10 *scientific article* is the dominating type in almost all topics. Within that type the *title* of the article is the most common content of the tweet. Using the title as the tweet content is also the most common way to refer to links in the news and science related categories. However, the title of a blog post or a newspaper article is not always the same as the title on a scientific article. This needs to be taken into consideration while interpreting the results.

Tweets with a content described as *title+comment* is often a promotion of an article, the tweet constructed after a formula similar to: “Our latest article is now out + title + link”, see example (3) below with the title of the article in italics.

- (3) New Preprint w/ @XXXX, X. XXXXX and @XXXXX: *Stochastic gene expression influences the selection of antibiotic resistance mutations* <https://t.co/DptIbMcR5V>

A tweet with a *descriptive* content can both be a short summary of the content on the linked page, and a statement/description of something where the link is supporting/expanding the content in the tweet, as exemplified in (4).

- (4) A new CRISPR system, known as Mobile-CRISPRi, works by binding to DNA and blocking other proteins from gaining access and activating transcription, thereby reducing gene expression and protein synthesis. <https://t.co/mYstjUAhe6>

The content *communication* includes conversations together with invitations/information about upcoming seminars and conferences, as well as commercial texts. Example (5) and (6) show two variants.

- (5) Are complex variants giving you a complex? Then join us for an educational webinar in which we demonstrate strategies for analyzing gene fusions, splice-site mutations, and co-occurring variants within the context of somatic cancer. Register here: <https://t.co/JKAI3p18Hq>
- (6) Eliminate staphylococcus totally from your system. For more info and order placing Call or Whatsapp: +2349065449561 To place an order online, kindly visit <https://t.co/UnUKuI8r9G> #libracin

It is worth noting that tweets with a *communication* content and of a scientific article type are often similar to *title+comment* tweets. The title is however excluded in the first case, see example (7), which changes the type of vocabulary used, and might by that also change the understandability for a general audience.

- (7) So excited to see our new paper out, showing architectural stripes and their impact on gene expression and development! Big thanks to @XXXXXX, @XXXXXX and to all other co-authors!  
<https://t.co/Ipm4Mm7axe>

Figure 11 shows the different kind of content detected in the topic *C. elegans/Neuromyelitis*. It is important to remember that this is the numerically smallest topic investigated, leading to few tweets included in this content analysis. This together with the fact that this topic in reality is two separate topics in this dataset, as shown in the descriptive statistic analysis (section 5.1), diminish the usability of these results. The conclusion possible to draw from this topic(s) is nevertheless that it is dominated by titles from scientific articles.

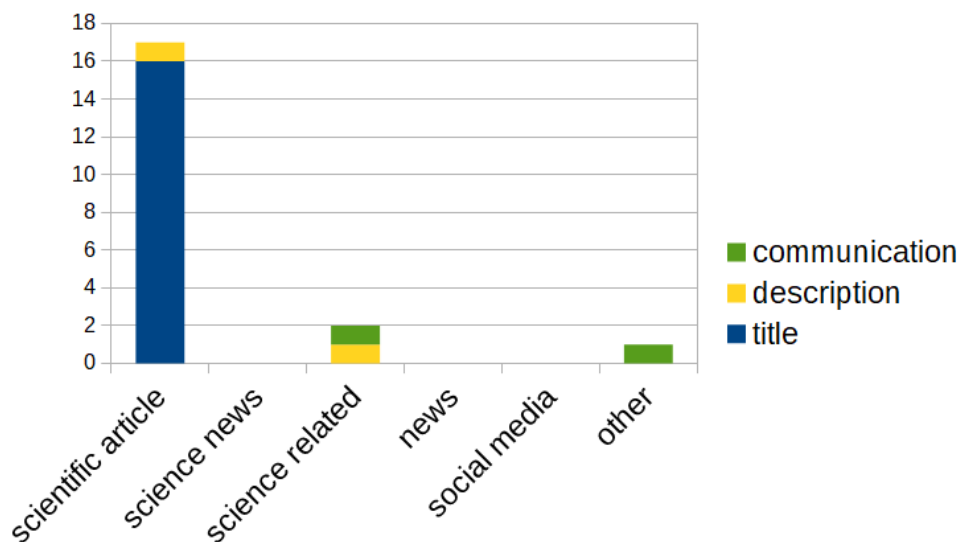


Figure 11: Content of tweets in the topic *C. elegans/Neuromyelitis*,  $n=20$ .

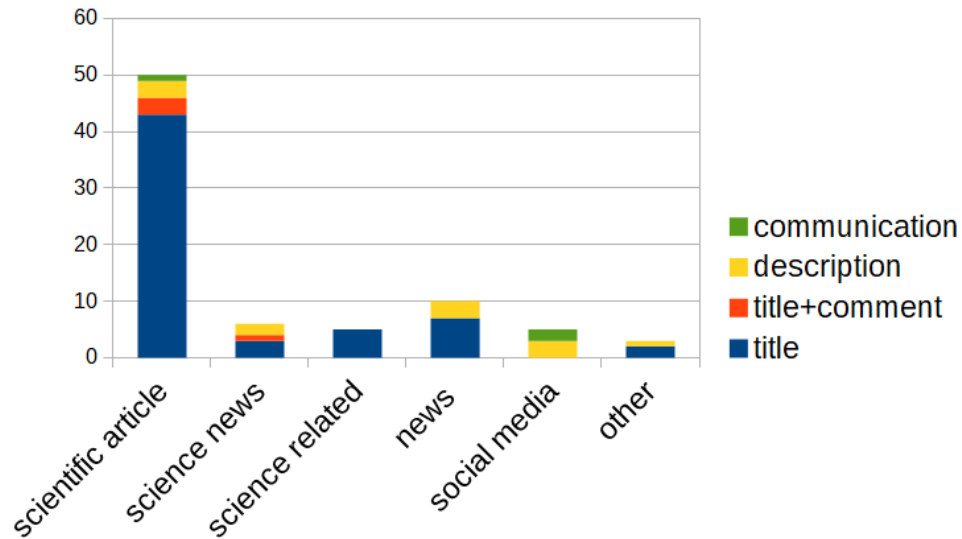


Figure 12: Content of tweets in the topic *Staphylococcus*,  $n=79$ .

The topic *Staphylococcus* displays a similar pattern, as shown in figure 12, most of the tweets are just a notation of the title of the scientific article. The links point to several types of sources, but the title is the predominant content independent of the type, even if it possible to identify some tweets with a more descriptive content.

The dataset for the topic *mRNA expression* consists of more tweets, and a larger number of tweets were therefore analysed. The variety in content is also larger compared to the previously described topics. Almost a third of the tweets with links to scientific articles have a descriptive or communicative content, and the same is true for tweets linking to science news.

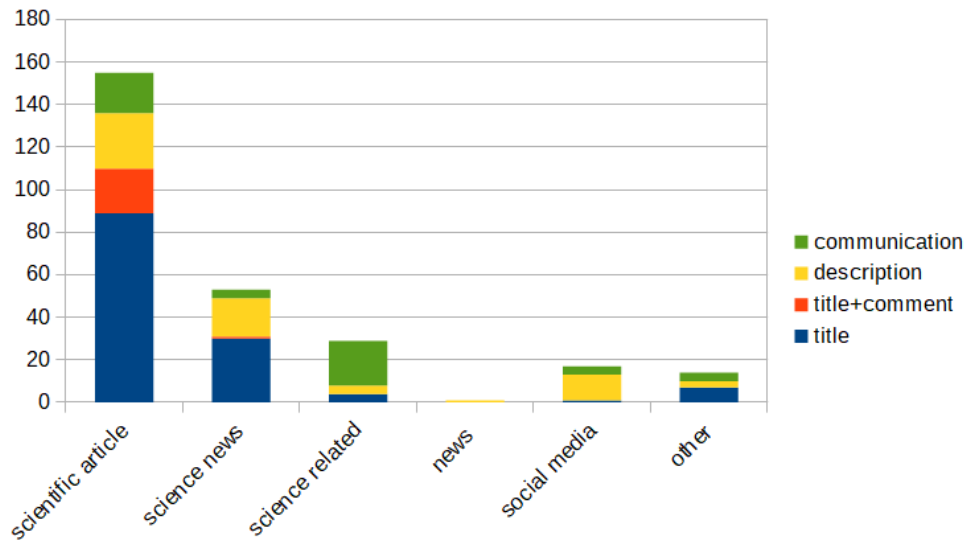


Figure 13: Content of tweets in the topic *mRNA expression*,  $n=269$ .

Communication is the most prevalent content for tweets pointing at scientific resources. It is mostly promotion for different courses and conferences relevant for researchers in the discipline.

The most diverse topic when it comes to variation in content, as well the type of source linked to, is *Species diversity/Phylogenetic tree*. The many tweets linking to sources of the category *Other* are mostly due to retweets and have already been mentioned above. It is interesting to see that half of them consisted of communication: a message was sent out, see example (2) in section 5.3.1, and then spread by retweets. Just as for the *mRNA expression* topic, a lot of the content in the science related type can be coded as communication. The most common content for tweets with links to scientific articles is still just the title, but it is also possible to detect more descriptive content. A similar pattern can be seen for the science news type.

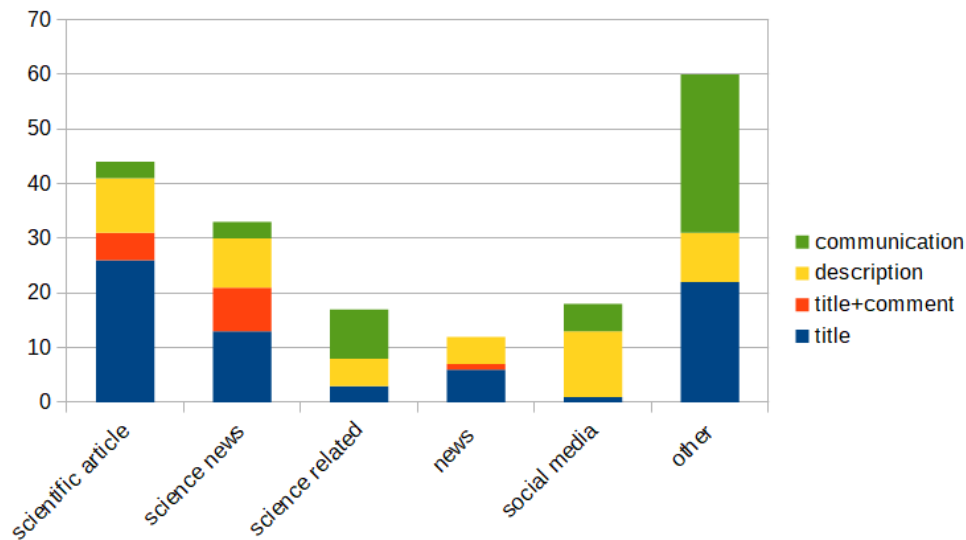


Figure 14: Content of tweets in the topic *Species diversity/Phylogenetic tree*,  $n=184$ .

## 5.5 Vocabulary analysis

The frequencies of terms commonly used as markers for stance, engagement and novelty were calculated for the topics *Species diversity/Phylogenetic tree* and *mRNA expression*. The frequencies were also compared between the two topics and the statistical significance in these comparisons was calculated using the log-likelihood.

The concept of stance can be divided into four parts, each representing different aspects. Some markers within the hedge group frequently used in the investigated topics were *suggest* and *likely*. Example (8) shows how it is used within the tweet; it is interesting to see that the text also contains the word *putative*, a term that also can be grouped within the hedges category.

- (8) RT @PNASNews: A study *suggests* a *putative* gene-expression hallmark common to monogamous male vertebrates of some species, namely dendrobati...

As shown in table 5 the use of markers for hedges are significantly more common in the *mRNA expression* topic than in the *Species diversity/Phylogenetic tree* topic, since the critical value for the 1% level is 6.63. This in contrast to the next two analysed groups, the boosters and the attitude markers, where the word frequencies are similar in both topics.

Table 5: Topic comparison, markers of stance

Type	<i>Species diversity/ Phylogenetic tree</i> (per 1000 words)	<i>mRNA expression</i> (per 1000 words)	<i>Log-likelihood</i>
Hedges	0.84	1.26	9.29
Boosters	0.87	1.07	2.29
Attitude	0.92	0.76	1.52
Self-mention	13.4	24.6	374.18

*Certain*, *demonstrate* and *sure* are some examples of boosters used in both datasets, while *important* and *interesting* are commonly used attitude markers. Example (9) shows the use of *certain* as a booster, however it is important to always look at the tweets themselves. This since the term *certain* is in many cases instead used in order to distinguish one group of specimens from another, as exemplified in (10).

(9) These aren't really optimal conditions, and *certainly* are well beyond the manufacturer's recommendations for storage. Not all mRNA transcripts will degrade in the same fashion. 2) There is a pretty substantial difference in RNA quality from the extractions. 3/n

(10) RT @NationalForests: Did you know that *certain* tree species depend on fire for their survival, while others have evolved thick bark to with...

The attitude markers can be used at a more detailed and technical level (11) as well as in a more communicative manner (12).

(11) *Interesting* to see the YY1 (transcription factor) and RMB25 show similar binding profiles, but experimental data suggest that the RMB25 is the key to start the interaction, followed by YY1. In other words, binding of RNA binding proteins to RNA drives next DNA-TF binding.

(12) RT @XXXXXX: *Interesting* findings regarding how exposure to rat poison has affected the gene expression of bobcats and perhaps altere...

The last type of markers for stance are the self-mentions, including *I*, *we*, *our*, and *us*, and as seen in table 5 this is the most frequently type of marker of stance used, and much more common in the topic *mRNA expression* than in *Species diversity/Phylogenetic tree*. It should be noted that *we*, *our*, and *us*, together with *you* and *your* are also markers of engagement, and table 6 shows that it is again more common in the *mRNA expression* topic.

Table 6: Topic comparison, markers of engagement

Type	<i>Species diversity/ Phylogenetic tree</i> (per 1000 words)	<i>mRNA expression</i> (per 1000 words)	<i>Log-likelihood</i>
Reader mentions	8.39	22.6	776.64
Directives (to instruct the reader)	1.87	1.54	3.67
Questions	0.09*	0.05*	-
Shared knowledge	0.84	1.25	9.25

\*the number of questions is calculated as number of question-marks per number of tweets.

The other markers of engagement investigated in this study are directives, questions, and shared knowledge. Questions are not common in the dataset, however for the markers of shared knowledge the difference between the topics is significant, while this is not the case for the directives. Examples of directives are *should* and *consider*, and (13) gives a good example of such directive.

- (13) People are great at confirmational bias and compartmentalization. *Considering* how much longer in our evolutionary history we've been guided 100% by emotions rather than knowledge, it's amazing we as a species can be rational at all.

*Common* and *routinely* are some words used to signal shared knowledge in the datasets, (14) below shows one example.

- (14) @XXXXXX This is an *all-too-common* vague and meaningless question. Please specify which measure(s) of 'biodiversity', membership/definition of 'community', and which function (i.e. 'the importance of vascular plant species richness for vascular plant GPP in 1 m<sup>2</sup>')

The last kind of markers investigated were markers of novelty. They are more common for the *Species diversity/Phylogenetic tree* topic than for *mRNA expression*, and are besides reader and self-mentions the most frequent types of markers, as seen comparing table 7 with table 5 and 6. *New/newest* are by far the most frequent terms used in this regard, followed by *important/importance*.

Table 7: Topic comparison, markers of novelty

Type	<b>Species diversity/ Phylogenetic tree</b> (per 1000 words)	<b>mRNA expression</b> (per 1000 words)	<b>Log-likelihood</b>
Novelty	11.2	8.04	57.53

## 6. Discussion

If one with research dissemination means spreading scientific result from the academia out to a more general public, this study shows that tweets with scientific content and links are not fulfilling this criteria. A large part of the tweets with embedded links, in all the four topics, contained only the title of a scientific article and no attempts to explain the content of the link for lay-people were made. A few of the tweets included a comment together with the title, however – these comments were mostly still in the tone of a colleague-to-colleague conversation, just making an announcement that the specific article was available online. Some institutions, departments and libraries as well as publishers also used this type of shout-out and title construct in order to promote available resources. This highlights that it is not only individual scientists that use Twitter as a platform, it is also a common communication way for the scientific journals themselves to promote new issues and their featured articles.

It is possible to see a difference in the amount of jargon used between tweets with and without embedded links, but in general the use of rare vocabulary is high in all the tweets discussed in this master thesis. On the other hand reader mentions are common in the tweets, and these are generally considered as markers for engagement, that the author in this way is trying to communicate with the readers. This means that science is not only present on Twitter as a one-way promotion of article, these markers for engagement shows that the intention to discuss different topics are there.

In this final section of the master thesis the research questions will be discussed in relation to earlier research. The concept of investigating Twitter and the methods used in this study, including its limitations, will also be examined. The discussion will end with conclusions and suggestions for future research.

### 6.1 Representativity, generalisation, and limitations

In a research project such as this we can never know who is writing on Twitter, in part due to ethical considerations. In this specific project it can be formulated like this: if we get little jargon in our jargon detector – does this mean that there are scientists writing in non-jargon, or does it mean that non-scientists writes about science? This kind of complication does not mean that

we should refrain from doing this kind of research, but need to be aware of the limitations while drawing conclusions.

As mentioned above it is not only the author, and co-authors, to an article that tweet about it; journals and publishing companies are also conducting this practice. This habit has in previous research sometimes been seen as both self-promotion and noise in the data (Haunschild et al., 2021; Haunschild et al., 2020; Luzón & Albero-Posac, 2020). It is also considered a problem when discussed in relation to altmetrics, how much social media activity, or interaction with society, does the score actually measure when it is just the title of the article that is tweeted – either by the author or the journal (Álvarez-Bornstein & Montesi, 2019). Nevertheless, a scenario where the authors, or journals, are not posting tweets about their own research on Twitter, or other social media, would lead to even less research dissemination and science communication.

In this context of science communication the question of who is tweeting is maybe of less importance than the content of the tweets. This means that the answer to the question in the beginning of this section might not even matter. The content of tweets will be discussed further in the next section, however, who is tweeting can be interesting in other aspects such as representativity. One such phenomenon is the presence of bots, automatic social media accounts, that generate tweets. Previous research investigating tweets linked to scientific articles about opioid research has found that these bots seem to behave in the same way as human users, and they were not considered to disturb the results (Haunschild et al., 2021). It is also interesting to connect the behaviour of the bots to the discussions about regarding the mechanical behaviour of human users, scientists and researchers, who more or less automatically tweet the title and a link to their recent article (Robinson-Garcia et al., 2017).

Connected to representativity is also the possibility of generalisation. It is important to remember that the dataset is a snapshot of what happened during a specific month and a specific year, it is a moment frozen in time. One example is the herbal treatment against *Staphylococcus* infection, a story originating from what it looks like an article in a Nigerian newspaper. This story is present in tweets with several types of links in my analysis, not only retweets of one original tweet but instead something that are mentioned in different ways. Only ten percent of the tweets with links were analysed manually, in numbers 79 tweets for this topic, and a news story like this can still be detected. Even if this was not the type of content analysis this master thesis focuses on it is a good example of the kind of results a snapshot-technique like the one used in this analysis can give. This study is based on tweets from one month and it would be interesting to do the same collection

from several months throughout a year. Then it would be possible to follow how the discussions about a specific topic changes.

Beside the above-mentioned limitations that are present in most kind of Twitter research, there also are some specific details for the methods used in this study that can be mentioned. The first is that “one get what one asks for”, four different scientific topics are investigated in this study, how much and in what ways these topics are presented, represented and discussed on Twitter varies. The datasets were generated using query-terms and these naturally have a great influence on the content of the tweets collected. Two topics were retrieved using the scientific name on organisms and diseases: *C.elegans/Neuromyelites* and *Staphylococcus*. As predicted in the description of the datasets the interest was higher for the more common human pathogen *Staphylococcus*, shown as a variety in types of pages linked to, as well as in the content. The topic *mRNA expression* was collected using specialist terms, and the tweets also had a technical, academic content. The topic *Species diversity/Phylogenetic trees* was collected using a wide range of query-terms including both academic terminology as well a more common vocabulary. This was also the topic showing the least amount of jargon and the most variable content.

While analysing text, regardless if it is the content or the vocabulary, it always comes back to the necessity of going back to the tweets and looking at the words in context. This can be seen in this study in the vocabulary analysis where word frequencies for specific words was used as markers for stance and engagement. It is important to take the context of the words into consideration when interpreting these results. One clear example of this is the word *certain* that have different meaning depending on context.

The lack of context is also an issue that needs to be discussed for the jargon detector. It should be mentioned that the De-Jargonizer tool used here is not developed as an analytical tool for research, but instead is supposed to be an aid for researchers while writing text for the general public (Rakedzon et al., 2017). The words are not analysed in their context and this means that some words that should be considered jargon, such as the *expression* in *mRNA expression* if the context was taken into consideration, will instead be labelled as common.

Nevertheless, since many of the patterns detected in this study are common for all the topics studied it is still possible to draw some conclusions.

## 6.2 Communication on Twitter

Even if neither the vocabulary nor the content in the tweets live up to requirements considered suitable for communication and outreach, it is still

possible to detect the intent to communicate in the high numbers of retweets and mentions in the datasets. The same pattern is seen in other studies where it also was concluded that scientist seems to retweet slightly more than the general user, and were including links in their tweets much more often (Büchi, 2017). No division of the tweets into link-containing and link-lacking was made in that study, so it is not possible to see if the pattern detected in this study regarding the relationship between retweets and links was the same. The results presented in this study are interesting since they are consistent over the four investigated topics, despite the fact that the datasets are of a different size. The lack of links in retweets might also be explained by the fact that the retweets often are used in combination with a mention. While creating the mention the @-character is used, and in most available apps for Twitter the @ followed by an username is a link in itself. It will however take the user to the account and not to specific tweets. Nevertheless it can still be considered communication and a direction to the link in the original tweet, if the original tweet contained one. When discussing retweet patterns it is important to remember that tweets generally are read by someone already following the author, or is a retweet by someone the reader is following. Although it is possible to search for keywords and hashtags on Twitter, this requires taking an active interest in the topic. In effect this means that the readers and authors are part of the same network, and it is not clear how often non-academics are part of the academic networks (Mohammadi et al., 2018). Some studies have shown that the larger the (scientific) account is, the more followers from outside academia it has (Côté & Darling, 2018). It is important to remember that journals also are part of those academic networks and are promoting their articles, as described above (Klar et al., 2020).

The difference in retweet patterns between tweets with and without links are also visible in the comparisons of word frequencies between the two subsets. Most of the words in these comparisons are however clustered in the middle, with similar frequencies in both kinds of tweets. For some topics it is possible to see a few more “scientific” words for tweets with links. The topic *Species diversity/Phylogenetic trees* shows a more spread picture, probably due to the wide range of subjects covered within the topic, due to the selection of the query-terms used to generate the dataset. It is also possible to detect the temporal nature of the dataset, and how social media in general is something that visualise the here and now. It has for example been shown that most links to scientific articles that actually get clicked on, are clicked on during the first two days (Fang et al., 2021). For the datasets analysed in this study one can conclude that it is unlikely that the word *bavaria* would occur at high frequency in other studies of the topic *Species diversity/Phylogenetic trees*, but it gives a picture of how scientific topics can be, and are, discussed in different contexts on Twitter.

A more scientific vocabulary used in tweets with links, compared to tweets without, are evident using the De-Jargonizer tool. This is true for all topics studied, even if the degree differed. It is not surprising that the topic *Species diversity/Phylogenetic trees* consists of tweets using a lower number of rare words compared to the other topics investigated. Already in the analysis of query terms in the preprocessing step it is possible to see that a more common language is used. If the authors' intent is that the tweets should be suitable for, and aimed at, the general public a lower percentage of words from the rare- and mid-frequency classes is recommended, numbers as low as 2% are mentioned (Rakedzon et al., 2017 and references therein). That it is hard to achieve such low level even in texts that are explicitly intended as popular science have been concluded in a study investigating so called lay summaries. These summaries were introduced as a complement to the regular abstract in order to facilitate research dissemination (Kuehne & Olden, 2015). The amount of jargon in those were later analysed using the same jargon detector as in this study (Rakedzon et al., 2017), with a result of around 8-12% jargon for the lay summaries. These numbers are similar to the values for the topics *mRNA expression* and *Species diversity/Phylogenetic trees* analysed in this study, while the two topics with the most scientific query-terms, *C.elegans/Neuromyelitis* and *Staphylococcus* are much higher with up to 20% and over. These numbers might not be as alarming as it seems, this since it is again important to remember that Twitter works as a network. Most users that reads tweets with scientific content are those that already are interested in the topics (Côté & Darling, 2018).

An explanation for this high amount of jargon can be found in the content analysis. Most of the tweets only contain the title of the scientific article and naturally the amount of jargon in these cases will be high. The detection of the practice of only tweeting the title of the scientific article was not unexpected since it has been seen in several earlier studies, as summarised in Sugimoto et al. (2017). What might be a little surprising is the fact that the content of these tweets have not changed more since the older studies were conducted. Instead it looks like the practice of just tweeting the title is still the dominant way of spreading the link, and maybe this practice is performed more or less routinely as suggested by Robinson-Garcia et al. (2017) and already mentioned in this discussion.

Some tweets with links to articles do however have the construction *title+comment*, and can be seen as signs of the wide use of Twitter as tool to maintain professional networks and collaborations, something often mentioned as one of the main reasons for using Twitter among scientists (Britton et al., 2019; Mohammadi et al., 2018). That Twitter is used in this way can also be seen in the topics *mRNA expression* and *Species diversity/Phylogenetic trees* where it is possible to detect tweets with a

*communicative* content. This kind of content is mostly present in tweets with links to different kinds of scientific resources, which can be online courses and tools, workshops and conferences. The practice can be seen as an example of communication between researchers, a variant of the preaching to the choir as it is described by Côté & Darling (2018). It also strengthens the opinion that Twitter is used as a communication tool as described in earlier research (Büchi, 2017).

The contents of the tweets are similar for all topics, dominated by titles of articles as well as the titles of science news and regular news. For the two numerically larger topics, *mRNA expression* and *Species diversity/Phylogenetic trees*, it is, beside the communicative content described above, also possible to identify more descriptive content, both in tweets linking to scientific articles and science news. It should be noted that the border between communication and description is sometimes hard to draw. For a majority of the tweets that were differently classified in the first and second round of content analysis were the discrepancies in the distinction between these two categories. These potential missclassifications however do not matter much in the overall discussions about the results. The interesting thing is that it is possible to detect content in the tweets that are descriptive with an intention to communicate. Both *mRNA expression* and *Species diversity/Phylogenetic trees* are also the topics with a larger proportion of the tweets linking to science news and science blogs, something that might indicate a larger tendency to successful outreach within those topics.

It was only tweets from the topics *Staphylococcus* and *Species diversity/Phylogenetic trees* that had links pointing to regular news, something that can be seen as an indicator of the interest in these topics for a wider audience. It also fits with the suggestions that scientific articles covering topics considered relevant for society also generates more tweets (Haunschild et al., 2021). The lack of links to news media in the topic *mRNA expression* implies that this was a topic not discussed in media in early 2019. One can only speculate that a collection of tweets during the same month in 2021 would give a different result, due to the fact that several of the vaccines against COVID-19 are based on mRNA. A comparison between these two years would probably illustrate the sometimes rapidly changing interest in a good way. Instead it is a tweet highlighting the consequences a decreasing species diversity can have for human food production that is in focus and spread through retweeting. Earlier studies have shown that tweets where an emotional aspect of the scientific content is highlighted, or tweets that in some other way are connected to current events or everyday life, generate more response from the public (Denia, 2020). This fact may also be connected to the beginning of this section, tweets with links are not retweeted to the same degree as tweets without. Since we have seen that it is still

common to only include the title of the scientific article in the tweet, and this is not generally considered interesting enough for a larger audience, the tweets are therefore not spread outside the limited circle. What a tweet or any other mention in social media actually can say about the academic or societal impact of an article are also commonly discussed in relation to altmetrics. In these discussions the distinction between the public interest and the societal impact is highlighted, where tweets are considered more often to reflect the former and not the latter (Tahamtan & Bornmann, 2020).

In the analyses of the vocabulary using the stance and engagement model the communication aspect of the tweets is also in focus. Here the two topics *mRNA expression* and *Species diversity/Phylogenetic tree* were compared to each other. It was possible to detect statistically significant differences between the two, where the frequencies of markers for both stance and engagement were higher for *mRNA expression* in several cases. It is however more interesting to discuss which markers that dominated in both datasets. Self-mentions and reader-mentions both occurred in much higher frequencies than all the other kinds of markers. This follows the patterns seen for blog-post where half the markers for stance and engagement came from reader-mentions (Zou & Hyland, 2019). These kind of mentions were also present in a dataset analysing tweeting in relation to conferences. The terms were there used in order to engage in the conference, by submitting contributions before the deadlines, or in discussions about the talks during the conference (Luzón & Alberó-Posac, 2020). This behaviour is also seen in the content analysis in this study where many tweets in the topics *mRNA expression* and *Species diversity/Phylogenetic tree* had a communicative content.

Considering the use of the term *interesting* and other attitude markers in science blogs (Zou & Hyland, 2019) it is a little surprising that these kind of markers for stance and engagement are not more common than they are in the investigated topics. Markers for novelty occurred on the other hand frequently, something also regularly seen in other types of popular science. Novelty is often combined with the practise of highlighting the relevance of the topic (Hyland, 2010), which also have been seen to increase the interest in tweets with scientific content (Denia, 2020).

### 6.3 Conclusions and further studies

When it comes to research dissemination and science communication I think this quote – part of the title of an article used in this study – summarises the feeling: “Think about how fascinating this is” (Zou & Hyland, 2020). That is what most scientists think about their topic, what they have chosen to study and research. But even the most devoted scientist knows that the public can not always feel the fascination about gene regulation in *C. elegans* or how

revolutionary it is when two species change places in the basal branches of the phylogenetic tree of chordates.

The topic *Species diversity/Phylogenetic tree* was the topic with the least amount of jargon in this analysis. The links included in this topic also points at the most diverse number of sources, showing that it is a topic probably discussed outside the academic circle both on Twitter and in other media. The comparison of vocabulary use, following the stance and engagement model, did however show that when it comes to both self- and reader-mentions they were more frequently used in the topic *mRNA expression*. These kinds of terms are considered as markers for engagement with the reader and could be a sign of an attempt to outreach. It might therefore be that part of the diversity in the content and jargon for *Species diversity/Phylogenetic tree* is a result of the wide range of material this topic covers.

Another conclusion from this study is that Twitter in many ways still is a place for science communication between peers, colleagues and collaborators. It is important that this practice is accounted for in studies addressing the questions discussed in this master thesis. No study investigating Twitter as a platform for research dissemination and science communication will find content and vocabulary fully adjusted to suit the general public. This because the researchers and scientists are also part of the social media, where they interact with each other, both in their professional and private roles. Twitter can and should not only be seen as a place for communication from the academia and out, it needs to be looked at from several perspectives. It would therefore be interesting to do follow-up studies with a more in-depth content, in order to try to identify the intended receiver for tweets with a communicative, or descriptive scientific content. If, and how, the vocabulary changes depending on these different kinds of readers is also an interesting question.

As mentioned in the literature review, one drawback while investigating Twitter using query terms is that it is only possible to capture part of the conversations. Considering the amount of reader mentions available in the data-sets it would also be interesting and useful to extend the study to include conversations. Analyses of interactions in science blogs have shown an interesting difference between the vocabulary used in blog posts and the comments. The posts were both informing about the content as well as took the role as a moderator by selecting topics and stimulate discussion. This while the comments included a high frequency personal pronouns, questions and phrases signalling politeness (Freddi, 2020). Future studies using Twitter conversations might show similar, or different results, and can be connected to discussions about how different readers are addressed.

The current COVID19 pandemic have in several ways highlighted the importance of researching how science is discussed on Twitter. The pandemic is a good example of how topics related to the everyday life of people generate the highest activity. When the general public got more interested in science and scientific articles, the scientists and researchers suddenly had a reach outside their normal academic networks on Twitter. In times like this when everyone can proclaim themselves an expert it is extra important that people with the academic knowledge takes care in their choice of vocabulary. It becomes clear that all tweets, even those that might be intended as communication between peers, can be misinterpreted by someone not familiar with the scientific language.

It also illustrates the changing nature of Twitter. One can only speculate about what will be considered interesting in the future, an uncertainty that is both a challenge and an asset while doing research.



## References

- Ädel, A. (2006). *Metadiscourse in L1 and L2 English*. John Benjamins Publishing Company.
- Ädel, A., & Mauranen, A. (2010). Metadiscourse: Diverse and Divided Perspectives. *Nordic Journal of English Studies*, 9(2), 1. <https://doi.org/10.35360/njes.215>
- Ahmed, W., Bath, P. A., & Demartini, G. (2017). Using Twitter as a Data Source: An Overview of Ethical, Legal, and Methodological Challenges. In K. Woodfield (Ed.), *The Ethics of Online Research*. (pp. 79–107). <https://doi.org/10.1108/s2398-601820180000002004>
- altmetric.com. (n.d.). altmetric.com. Retrieved from <https://www.altmetric.com/>
- Álvarez-Bornstein, B., & Montesi, M. (2019). Who is interacting with researchers on Twitter? A survey in the field of information science. *JLIS.It*, 10(2), 87–106. <https://doi.org/10.4403/jlis.it-12530>
- Bawden, D., & Robinson, L. (2015). *Introduction to Information Science*. London: Facet Publishing.
- Bettini, E., & Locci, M. (2021). SARS-CoV-2 mRNA Vaccines: Immunological mechanism and beyond. *Vaccines*, 9(2), 1–20. <https://doi.org/10.3390/vaccines9020147>
- Bornmann, L., Haunschild, R., & Adams, J. (2018). Convergent validity of altmetrics and case studies for assessing societal impact: an analysis based on UK Research Excellence Framework (REF) data. *Proceedings of Science and Technology Indicators (STI)*, (September), 12–14. Retrieved from [www.altmetric.com](http://www.altmetric.com)
- Britton, B., Jackson, C., & Wade, J. (2019). The reward and risk of social media for academics. *Nature Reviews Chemistry*, 3(8), 459–461. <https://doi.org/10.1038/s41570-019-0121-3>
- Büchi, M. (2017). Microblogging as an extension of science reporting. *Public Understanding of Science*, 26(8), 953–968. <https://doi.org/10.1177/0963662516657794>
- Bullock, O. M., Colón Amill, D., Shulman, H. C., & Dixon, G. N. (2019). Jargon as a barrier to effective science communication: Evidence from metacognition. *Public Understanding of Science*, 28(7), 845–853. <https://doi.org/10.1177/0963662519865687>

- Burns, T. W., O'Connor, D. J., & Stocklmayer, S. M. (2003). Science communication: A contemporary definition. *Public Understanding of Science*, 12(2), 183–202. <https://doi.org/10.1177/09636625030122004>
- Compagnucci, L., & Spigarelli, F. (2020). The Third Mission of the university: A systematic literature review on potentials and constraints. *Technological Forecasting and Social Change*, 161(August), 120284. <https://doi.org/10.1016/j.techfore.2020.120284>
- Considine, M. J., Siddique, K. H. M., & Foyer, C. H. (2017). Nature's pulse power: Legumes, food security and climate change. *Journal of Experimental Botany*, 68(8), 1815–1818. <https://doi.org/10.1093/jxb/erx099>
- Costas, R., Zahedi, Z., & Wouters, P. (2015). Do “Altmetrics” Correlate With Citations? Extensive Comparison of Altmetric Indicators With Citations From a Multidisciplinary Perspective. *Journal of the Association for Information Science and Technology*, 66(10), 2003–2019. <https://doi.org/10.1002/asi>
- Côté, I. M., & Darling, E. S. (2018). Scientists on Twitter: Preaching to the choir or singing from the rooftops? *Facets*, 3(1), 682–694. <https://doi.org/10.1139/facets-2018-0002>
- Data4Impact. (n.d.). Big Data approaches for improved assessment of the societal impact in the Health, Demographic Change and Wellbeing Societal Challenge. Retrieved from <http://www.data4impact.eu/>
- Della Giusta, M., Jaworska, S., & Vukadinović Greetham, D. (2021). Expert communication on Twitter: Comparing economists' and scientists' social networks, topics and communicative styles. *Public Understanding of Science*, 30(1), 75–90. <https://doi.org/10.1177/0963662520957252>
- Denia, E. (2020). The impact of science communication on Twitter: The case of Neil deGrasse Tyson. *Comunicar*, 28(65), 21–30. <https://doi.org/10.3916/C65-2020-02>
- Díaz-Faes, A. A., Bowman, T. D., & Costas, R. (2019). Towards a second generation of ‘social media metrics’: Characterizing Twitter communities of attention around science. *PLoS ONE*, 14(5), 1–18. <https://doi.org/10.1371/journal.pone.0216408>
- Didegah, F., Mejlgaard, N., & Sørensen, M. P. (2018). Investigating the quality of interactions and public engagement around scientific papers on Twitter. *Journal of Informetrics*, 12(3), 960–971. <https://doi.org/10.1016/j.joi.2018.08.002>

- Fang, Z., Costas, R., Tian, W., Wang, X., & Wouters, P. (2021). How is science clicked on Twitter? Click metrics for Bitly short links to scientific publications. *Journal of the Association for Information Science and Technology*, (September 2020), 1–15.  
<https://doi.org/10.1002/asi.24458>
- Feidenheimer, A., Frietsch, R., Schubert, T., & Neuhäusler, P. (2018). *Intermediate report on the conceptual framework*.
- Fiesler, C., & Proferes, N. (2018). “Participant” Perceptions of Twitter Research Ethics. *Social Media and Society*, 4(1).  
<https://doi.org/10.1177/2056305118763366>
- Freddi, M. (2020). Blurring the lines between genres and audiences: Interaction in science blogs. *Discourse and Interaction*, 13(2), 9–35.  
<https://doi.org/10.5817/DI2020-2-9>
- Gaffney, D. F., & Puschmann, C. (2013). Data Collection on Twitter. In K. Weller, A. Bruns, J. Burgess, M. Mahrt, & C. Puschmann (Eds.), *Twitter and society* (pp. 55–67). [https://doi.org/10.11207/soshikikagaku.48.4\\_47](https://doi.org/10.11207/soshikikagaku.48.4_47)
- Gnanamani, A., Hariharan, P., & Paul-Satyaseela, M. (2017). Staphylococcus aureus: Overview of bacteriology, clinical diseases, epidemiology, antibiotic resistance and therapeutic approach. *Frontiers in Staphylococcus Aureus*, 4(28).
- Haahr, M. (2021). RANDOM.ORG: True Random Number Service. Retrieved March 3, 2021, from <https://www.random.org/>
- Hargittai, E., Füchslin, T., & Schäfer, M. S. (2018). How Do Young Adults Engage With Science and Research on Social Media? Some Preliminary Findings and an Agenda for Future Research. *Social Media and Society*, 4(3). <https://doi.org/10.1177/2056305118797720>
- Haunschild, R., Bornmann, L., Potnis, D., & Tahamtan, I. (2021). Investigating Diffusion of Scientific Knowledge on Twitter : A Study of Topic Networks of Opioid. *ArXiv Preprint ArXiv:2101.11483*, 1–45.
- Haunschild, R., Leydesdorff, L., & Bornmann, L. (2020). Library and Information Science Papers Discussed on Twitter: A new Network-based Approach for Measuring Public Attention. *Journal of Data and Information Science*, 5(3), 5–17. <https://doi.org/10.2478/jdis-2020-0017>
- Haunschild, R., Leydesdorff, L., Bornmann, L., Hellsten, I., & Marx, W. (2019). Does the public discuss other topics on climate change than researchers? A comparison of explorative networks based on author

- keywords and hashtags. *Journal of Informetrics*, 13(2), 695–707. <https://doi.org/10.1016/j.joi.2019.03.008>
- Holmberg, K., & Thelwall, M. (2014). Disciplinary differences in Twitter scholarly communication. *Scientometrics*, 101(2), 1027–1042. <https://doi.org/10.1007/s11192-014-1229-3>
- Hyland, K. (2005a). *Metadiscourse: exploring interaction in writing*. London: Continuum.
- Hyland, K. (2005b). Stance and engagement: A model of interaction in academic discourse. *Discourse Studies*, 7(2), 173–192. <https://doi.org/10.1177/1461445605050365>
- Hyland, K. (2007). Is There an “Academic Vocabulary”? *Tesol Quarterly*, 41(2).
- Hyland, K. (2010). Constructing proximity: Relating to readers in popular and professional science. *Journal of English for Academic Purposes*, 9(2), 116–127. <https://doi.org/10.1016/j.jeap.2010.02.003>
- Hyland, K. (2017). Metadiscourse: What is it and where is it going? *Journal of Pragmatics*, 113, 16–29. <https://doi.org/10.1016/j.pragma.2017.03.007>
- Hyland, K., & Jiang, F. (Kevin). (2016). “We must conclude that...”: A diachronic study of academic engagement. *Journal of English for Academic Purposes*, 24, 29–42. <https://doi.org/10.1016/j.jeap.2016.09.003>
- Jackson, N. A. C., Kester, K. E., Casimiro, D., Gurunathan, S., & DeRosa, F. (2020). The promise of mRNA vaccines: a biotech and industrial perspective. *Npj Vaccines*, 5(1), 3–8. <https://doi.org/10.1038/s41541-020-0159-8>
- Joubert, M., & Costas, R. (2019). Getting to Know Science Tweeters: A Pilot Analysis of South African Twitter Users Tweeting about Research Articles. *Journal of Altmetrics*, 2(1), 2. <https://doi.org/10.29024/joa.8>
- Ke, Q., Ahn, Y., & Sugimoto, C. R. (2017). A Systematic Identification of Scientists on Twitter. *PLoS ONE*, 12(4).
- Klar, S., Krupnikov, Y., Ryan, J. B., Searles, K., & Shmargad, Y. (2020). Using social media to promote academic research: Identifying the benefits of twitter for sharing academic work. *PLoS ONE*, 15(4), 1–15. <https://doi.org/10.1371/journal.pone.0229446>

- Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research. *Journal of Chiropractic Medicine*, 15(2), 155–163. <https://doi.org/10.1016/j.jcm.2016.02.012>
- Kuehne, L. M., & Olden, J. D. (2015). Opinion: Lay summaries needed to enhance science communication. *Proceedings of the National Academy of Sciences of the United States of America*, 112(12), 3585–3586. <https://doi.org/10.1073/pnas.1500882112>
- Larivière, V., Sugimoto, C. R., & Cronin, B. (2012). A Bibliometric Chronicling of Library and Information Science’s First Hundred Years. *Journal of the American Society for Information Science and Technology*, 63(5), 997–1016. <https://doi.org/10.1002/asi>
- Letierce, J., Passant, A., Breslin, J. G., & Decker, S. (2010). Using Twitter during an academic conference: The #iswc2009 use-case. *ICWSM 2010 - Proceedings of the 4th International AAAI Conference on Weblogs and Social Media*, May, 279–282.
- Lorentzen, D. G., & Nolin, J. (2017). Approaching Completeness: Capturing a Hashtagged Twitter Conversation and Its Follow-On Conversation. *Social Science Computer Review*, 35(2), 277–286. <https://doi.org/10.1177/0894439315607018>
- Luzón, M. J. (2013). Public Communication of Science in Blogs: Recontextualizing Scientific Discourse for a Diversified Audience. *Written Communication*, 30(4), 428–457. <https://doi.org/10.1177/0741088313493610>
- Luzón, M. J., & Albero-Posac, S. (2020). ‘Had a lovely week at #conference2018’: An Analysis of Interaction through Conference Tweets. *RELC Journal*, 51(1), 33–51. <https://doi.org/10.1177/0033688219896862>
- McGrath, L., & Kuteeva, M. (2012). Stance and engagement in pure mathematics research articles: Linking discourse features to disciplinary practices. *English for Specific Purposes*, 31(3), 161–173. <https://doi.org/10.1016/j.esp.2011.11.002>
- Moernaut, R., Mast, J., Temmerman, M., & Broersma, M. (2020). Hot weather, hot topic. Polarization and sceptical framing in the climate debate on Twitter. *Information Communication and Society*, 0(0), 1–20. <https://doi.org/10.1080/1369118X.2020.1834600>

- Mohammadi, E., Thelwall, M., Kwasny, M., & Holmes, K. L. (2018). Academic information on Twitter: A user survey. *PLoS ONE*, *13*(5), 1–18. <https://doi.org/10.1371/journal.pone.0197265>
- Mullen, L. (2016). *tokenizers: A Consistent Interface to Tokenize Natural Language Text*. Retrieved from <https://cran.r-project.org/package=tokenizers>.
- Nelhans, G., & Lorentzen, D. G. (2016). Twitter conversation patterns related to research papers. *Information Research*, *21*(2), 1–32.
- Nelhans, G., Papageorgiou, H., Pukelis, L., & Demiros, I. (2020). *Analysis of company, EU Projects, policy documents, clinical guideline and social media / media data report*. 1–64.
- NHS. (2020). Neuromyelitis optica. Retrieved March 15, 2021, from <https://www.nhs.uk/conditions/neuromyelitis-optica/>
- NobelPrize.org. (2002). The Nobel Prize in Physiology or Medicine for 2002 [Press release]. Retrieved March 12, 2021, from Nobel Media AB 2021 website: <https://www.nobelprize.org/prizes/medicine/2002/press-release/>
- Pace, L. A., & Livingston, M. M. (2005). Protecting Human Subjects in Internet Research. *EJBO-Electronic Journal of Business Ethics and Organization Studies*. <https://doi.org/10.1093/acprof:oso/9780195375893.003.0007>
- Peters, H. P., Dunwoody, S., Allgaier, J., Lo, Y., & Brossard, D. (2014). Public communication of science. *EMBO Reports*, *15*(7), 749–753. <https://doi.org/10.15252/embr.201438979>
- Priem, J., Taraborelli, D., Groth, P., & Neylon, C. (2010). *Altmetrics: A manifesto, 26 October 2010*. Retrieved from <http://altmetrics.org/manifesto/>
- Pukelis, L., & Stanciauskas, V. (2018). Big Data approaches to estimating the impact of EU research funding. *STI 2018 Conference Proceedings*, 429–435. Centre for Science and Technology Studies (CWTS).
- Radzikowski, J., Stefanidis, A., Jacobsen, K. H., Croitoru, A., Crooks, A., & Delamater, P. L. (2016). The Measles Vaccination Narrative in Twitter: A Quantitative Analysis. *JMIR Public Health and Surveillance*, *2*(1), e1. <https://doi.org/10.2196/publichealth.5059>
- Rakedzon, T., Segev, E., Chapnik, N., Yosef, R., & Baram-Tsabari, A. (2017). Automatic jargon identifier for scientists engaging with the public and science communication educators. *PLoS ONE*, *12*(8), 1–13. <https://doi.org/10.1371/journal.pone.0181742>

- Rayson, P. (n.d.). Log-likelihood and effect size calculator. Retrieved from <http://ucrel.lancs.ac.uk/llwizard.html>
- Rayson, P., & Garside, R. (2000). Comparing corpora using frequency profiling. *The Workshop on Comparing Corpora, October*, 1–6. <https://doi.org/10.3329/akmmcj.v8i1.31665>
- Revelle, W. (2021). *psych: Procedures for Psychological, Psychometric, and Personality Research*. Northwestern University, Evanston, Illinois. R package version 2.1.3. <https://CRAN.R-project.org/package=psych>.
- Robinson-Garcia, N., Costas, R., Isett, K., Melkers, J., & Hicks, D. (2017). The unbearable emptiness of tweeting—About journal articles. *PLoS ONE*, *12*(8), 1–19. <https://doi.org/10.1371/journal.pone.0183551>
- Schmitt, M., & Jäschke, R. (2017). What do computer scientists tweet? Analyzing the link-sharing practice on Twitter. *PLoS ONE*, *12*(6), 1–28. <https://doi.org/10.1371/journal.pone.0179630>
- Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4).
- SFS (1992:1434). (n.d.). Högskolelagen. Retrieved from <https://lagen.nu/1992:1434#K1P2S2>
- Silge, J., & Robinson, D. (2017). *Text mining with R: A tidy approach*. O'Reilly Media, Inc.
- Su, L. Y. F., Scheufele, D. A., Bell, L., Brossard, D., & Xenos, M. A. (2017). Information-Sharing and Community-Building: Exploring the Use of Twitter in Science Public Relations. *Science Communication*, *39*(5), 569–597. <https://doi.org/10.1177/1075547017734226>
- Sugimoto, C. R., Work, S., Larivière, V., & Haustein, S. (2017). Scholarly Use of Social Media and Altmetrics: A Review of the Literature. *Journal of the Association for Information Science and Technology*, *68*(9), 2037–2062. <https://doi.org/10.1002/asi>
- Sundström, G. (2010). *Evolution of the neuropeptide Y and opioid systems and their genomic regions*. (Doctoral dissertation, Acta Universitatis Upsaliensis).
- Tahamtan, I., & Bornmann, L. (2020). Altmetrics and societal impact measurements: Match or mismatch? a literature review. *El Profesional de La Informacion (EPI)*, *29*(1). <https://doi.org/10.3145/epi.2020.ene.02>

- Thelwall, M., Haustein, S., Larivière, V., & Sugimoto, C. R. (2013). Do Altmetrics Work? Twitter and Ten Other Social Web Services. *PLoS ONE*, 8(5), 1–7. <https://doi.org/10.1371/journal.pone.0064841>
- Theodoridis, S., Fordham, D. A., Brown, S. C., Li, S., Rahbek, C., & Nogues-Bravo, D. (2020). Evolutionary history and past climate change shape the distribution of genetic diversity in terrestrial mammals. *Nature Communications*, 11(1), 1–11. <https://doi.org/10.1038/s41467-020-16449-5>
- Vande Kopple, W. J. (1985). Some Exploratory Discourse on Metadiscourse. *College Composition and Communication*, 36(1), 82–93. <https://doi.org/doi:10.2307/357609>
- White, M. D., & Marsh, E. E. (2006). Content analysis: A flexible methodology. *Library Trends*, 55(1), 22–45. <https://doi.org/10.1353/lib.2006.0053>
- Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., ... Yutani, H. (2019). Welcome to the Tidyverse. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>
- Willoughby, S. D., Johnson, K., & Sterman, L. (2020). Quantifying scientific jargon. *Public Understanding of Science*, 29(6), 634–643. <https://doi.org/10.1177/0963662520937436>
- Wilsey, B. (2020). Restoration in the face of changing climate: importance of persistence, priority effects, and species diversity. *Restoration Ecology*, 1–10. <https://doi.org/10.1111/rec.13132>
- Zou, H. J., & Hyland, K. (2019). Reworking research: Interactions in academic articles and blogs. *Discourse Studies*, 21(6), 713–733. <https://doi.org/10.1177/1461445619866983>
- Zou, H. J., & Hyland, K. (2020). “Think about how fascinating this is”: Engagement in academic blogs across disciplines. *Journal of English for Academic Purposes*, 43. <https://doi.org/10.1016/j.jeap.2019.100809>

## Appendix

Table A1. Query terms used while generating the datasets

<i>C. elegans/ Neuromyelitis</i>	<i>Staphylococcus</i>	<i>mRNA expression</i>	<i>Species diversity/ Phylogenetic trees</i>
caenorhabditis elegans	staphylococcus	mrna expression	species genus
nematode elegans		mrna levels	phylogenetic tree
neuromyelitis		mrna decay	phylogenetic trees
optica		splice site	phylogenetic analyses
		mrna splicing	phylogenetic analysis
		alternatively spliced	phylogenetic relationships
		alternative splicing	tree species
		splice variant	species diversity
		splice variants	related species
		pre mrna	species distribution
		mrna stability	species tree
		mrna	species richness
		transcripts	genus species
		messenger rna	species identification
		mrna protein	species specific
		rna binding	species genera
		gene expression	evolutionary history
	protein expression		

*Table A2. Words used as markers for stance*

<b><i>Hedges</i></b>	<b><i>Boosters</i></b>	<b><i>Attitude markers</i></b>	<b><i>Self-mention</i></b>
apparently	certain	agree	I
appear	claim	disagree	we
indicate	clearly	important	our
knowledge	demonstrate	interesting	us
likely	extremely	promising	
perhaps	indeed		
possible	must		
suggest	obviously		
	prove		
	sure		

*Table A3. Words used as markers for engagement*

<b><i>Reader mentions</i></b>	<b><i>Shared knowledge</i></b>	<b><i>Directives</i></b>
you	apparently	demonstrate
your	common	assess
we	obviously	calculate
our	traditional	compare
us	integrate	define
	usually	pick
	routinely	allow
		apply
		follow
		classify
		develop
		imagine
		estimate
		remember
		increase
		determine
		consider
		add
		must
		analyse
		go
		should

*Table A4. Words used as markers for novelty*

***Novelty***

novel/novelty

advances/advancing

important/importance

recent/recently

new/newest

discover/discovery