universität wien

DIPLOMARBEIT / DIPLOMA THESIS

Titel der Diplomarbeit / Title of the Diploma Thesis

„The Virtual Campfire: Elements of Spoken Language in Written Online Storytelling“

verfasst von / submitted by

Stefan Ostner

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

Magister der Philosophie (Mag.phil.)

Wien, 2017 / Vienna, 2017

Studienkennzahl lt. Studienblatt / degree programme code as it appears on the student record sheet:	A 190 344 884
Studienrichtung lt. Studienblatt / degree programme as it appears on the student record sheet:	Lehramtsstudium UF Englisch, UF Informatik & Informationsmanagement
Betreut von / Supervisor:	Univ.-Prof. PD Mag. Dr. Gunther Kaltenböck, M.A.

Acknowledgments

First and foremost, I would like to thank my thesis supervisor, Univ.-Prof. PD Mag. Dr. Gunther Kaltenböck, M.A., for his support and guidance during the writing of this thesis and for his much-appreciated help on those areas of the thesis that proved to be more difficult than expected.

I would also like to thank all friends and family for their continued support during the writing of this thesis. Their encouragement and support has proven invaluable over these last few months.

Table of Contents

List of figures

List of tables

1 Introduction

David Crystal's book "Language and the Internet" opens with a number of examples taken from contemporary media showcasing how in a comparatively very short time, the Internet has taken over everyday life in Western society – all of them formulated in a way that seems rather dramatic, and culminating in the then French president's gloom-and-doom description of the Internet and its influence on language as a "major risk for humanity" (Crystal 2006: 1). The book, whose first edition was released in 2001, seems almost ancient, considering the rapid spread of and development within the Internet and so-called online culture, but the attitudes shown in these examples continue to live on. The Internet still seems a mysterious Other for many, for digital natives summed up somewhat painfully in German chancellor Angela Merkel's 2013 description of it as "Neuland" – virgin territory ("Die Kanzlerin entdeckt #Neuland" 2013). Yet there is a constant increase of users, with the Internet reaching 85% of EU households in 2016 (eurostat 2017). It is one of the major aims of this thesis to contribute to what may be called a demystification of the Internet and of Internet culture by contributing to academic knowledge on one of its aspects.

Of course, research has been done on the Internet, a vast amount of it at that; and aside from the technical side, researchers have also investigated the Internet's culture from the beginning and, more pertinently for this thesis, its language. These investigations have evolved alongside the technological developments that have allowed the Internet to become more and more fast-paced, and early research is focused on "slower", asynchronous ways of online communication such as e-mail and Internet fora (Murray 1988; Yates & Orlikowski 1993); with the rise of the World Wide Web and thus easy Internet accessibility for the general public and, later, quickly increasing Internet access speeds, research subjects have changed. The quick pace of development is recognizable in almost all of the texts; when Crystal (2006) includes MUDs in his book, but does not yet talk about social media and smartphone messaging simply because these do not exist yet, it highlights very well the problem of any kind of research that pertains to matters of the Internet – it becomes outdated very quickly, in extremes even before its publishing date. In this sense, doing such research is both rewarding, because it is easy to find something nobody has yet investigated, but may also be found to be frustrating for the reasons mentioned above.

The following thesis's aim is the analysis of one aspect of a specialized sub-genre of online forum communication, and this description shows the aforementioned dilemma quite well: while there is a plethora of research on disparate aspects of online forum use –

for example on discourse patters (Morrow 2006), on use by EFL students (Deris, Koon & Salam 2015), or on formality (Montero-Fleta *et al.* 2009) – there is no definite description of the genre of online forum use as a whole, and there are no comparable investigations into other sub-genres of forum use. This may seem like a lack of systematicity, but also showcases the sheer amount of new possible research topics that have exploded onto the scene within the last twenty years.

But why attempt such an analysis? What seems interesting about the genre of online storytelling (as a sub-genre of online forum use) in particular is that it is a written form of something that has traditionally almost exclusively been done orally. There is the well-worn trope of campfire story-telling, whereas the idea of telling these sometimes rather trivial stories in a written fashion, perhaps with a letter, invokes images of time-consuming tedium and the writing of reams of paper. Now, in a time where typing takes much less time than long-hand writing and releasing a text to the public is as simple as the click of a button, we find the genre transplanted into writing. And indeed, the forum under investigation (described in much more detail later in the thesis) may be seen as a "virtual campfire", as the title suggests – or rather an array of virtual campfires, where listeners may stroll from one to the next, each occupied with and telling stories pertaining to a certain topic. The question remains: are these authors more influenced by the medium of writing, since almost everybody has, at some point in their life, had at least fleeting contact with fictional writing in the form of novels or short stories, these also being codified in a certain way and with certain genre conventions? Or are they influenced by the kind of situation they find themselves in – one that has, as mentioned above, long been perceived as one of spoken language? This is what the thesis wants to investigate, utilizing the following research questions:

In the genre of online storytelling, do authors include features associated with spoken language? If so, which?

However, since the view that language has a single axis between speech and writing alongside which all and any genres are located will turn out to be rather simplistic and, more to the point, simply wrong over the course of this investigation, another question follows: what does the genre of online storytelling do that is uniquely characteristic for itself? How does it differentiate itself not only from more conventional, offline methods of communication, but also even from other online genres? Or, in other words:

Which language features, if any, are unique to online storytelling?

This question is one that can be taken quite generally: it applies to all levels of language and includes single graphemes just as much as the entire organization of a text or discourse. It, too, strives to define the genre more closely in order to contribute to a general idea of making online language more tangible and concrete.

To answer these questions, the thesis requires a definition of the features that are common to speech and those that are common to writing, which is harder to find or attain than it may seem to be at first glance. While a complete and comprehensive investigation of such features lies far beyond the scope of this thesis, chapter 2 and its sub-chapters will enumerate and expand upon the features that have been chosen for investigation. Much of this chapter will be based on descriptions by Biber (1988) and Leech (2000) as well as grammars by Quirk *et al.* (1985) and Biber *et al.* (1999). The former is a comprehensive and helpful descriptive grammar, whereas the latter's advantage is that the information contained within has largely been obtained by the use of corpus linguistics, applying its methods to the genres of conversation, fictional writing, news and academic writing much as this thesis will do to the genre of online storytelling. The chapter closes with a brief treatise on corpora; after giving a definition, closely aligned to McEnery & Wilson (2001), it will go more in-depth on corpora use with data collected from the Internet.

Chapter 3.4 will give a brief description of the research methodology used with the data collected from the Internet, while also describing how the data itself was collected and which criteria were used for its selection. To be quite complete, it will also offer a look into the IT side of things; the main focus, however, will be on the linguistic methodology and on the reasoning behind its utilization.

The jump from theory to practice is accomplished in chapter 4, which can be considered the culmination of the previous chapters: the methodology introduced in chapter 3 is utilized to analyze the features which were previously theoretically discussed in chapter 2. Again, the whole section will be closely linked to the findings of Biber *et al.* (1999), since their methodology aligns well with that utilized in this thesis, and the aspects of language discussed herein are a small subset of their far more comprehensive analysis of different genres. Since their analysis contains both speech and writing, the numbers they give provide a useful base from which it is possible to discern whether a particular aspect of language fits in better with one medium or the other.

Finally, chapter 5 offers a short summary of the findings, together with a conclusion which will give an answer to the research question formulated above. A further section will be devoted to a discussion of the limitations of this thesis and the possibilities for further research.

2 Literature review

The following sub-chapters will cover existing research done on a variety of topics that are pertinent to the thesis. I will start with an account of research done on the use of language on the Internet, focusing on genre and register, but also including findings on sociolinguistics and diachronic change, as they may help situate the practical research. This research will serve as a starting point for my own research as well as a tentative basis for the formulation of expectations regarding into the research.

A second sub-chapter will then offer some criteria for the distinction between spoken and written language, which will also be used as basis for the corpus analyses. This section will be further sub-divided into the various separate aspects that are under consideration in this thesis. Each of these aspects will be discussed based on existing corpus findings and other research, based on which a preliminary first expectation for the genre of online storytelling will be formulated and justified. Research will be conducted based on these expectations; if results do not mirror them, the necessary conclusions will be drawn in the practical part of the thesis.

Finally, there will be a short section on previous uses of corpus analysis in conjunction with online language data. This section will examine the discipline of corpus linguistics critically, enumerating some of its advantages and disadvantages and justifying its use in the thesis, before moving on to the specific challenges encountered in trying to analyze Internet language data with the tools of corpus analysis.

2.1 Characteristics of Internet language use

This chapter will offer a general overview of findings gained from the analysis of texts of all kinds gained from computer-mediated communication[1] environments. It will start by defining CMC as well as the two major sub-types of CMC (i.e. synchronous and asynchronous CMC), after which it will go more in-depth on some aspects connected to CMC's position between – or beyond – writing and speech.

CMC is a relatively new phenomenon, only made possible by the rise of the personal computer and, maybe even more importantly, the Internet. The term came into common usage in the 1990s, aided by the decision to include it in the title of the *Journal of Computer-Mediated Communication* (Crystal 2011: 1). Crystal himself criticizes the term

[1] Hereafter referred to as CMC.

both as too broad, including all kinds of non-textual communication (2011: 1), and as too strongly focused on the medium, proposing "Netspeak" as an alternative term (2006: 19). Leaving aside the question whether these two arguments are contradictory, CMC is still a term that is now widely used for communication aided by computer technology; however, the question of how wide or narrow a definition should be used does still arise, as with the advent of ever improving technological possibilities, audio and video communication also falls under the heading of CMC. Perhaps it is best to simply restrict oneself to the most pertinent forms of CMC, as Herring does (1996: 10), which in the case of this thesis is text-based CMC.

CMC has been under linguistic investigation ever since its inception, as it allowed researchers a look at the emergence of a completely new form of communication and, hence, language production. It has been found that CMC users utilize language features both associated with writing and with speaking (although more strongly those associated with writing), as well as creatively inventing new linguistic strategies that have no counterpart in the analogue world (Crystal 2006: 50–51, 2011: 19–21; Murray 2000). The language used also varies strongly within the confines of digital communication: not only do different social groups have different linguistic conventions (Bernstein *et al.* 2011; Kushin & Kitchener 2009; Pérez-Sabater 2013; Tran & Ostendorf 2016), but there are also a variety of forms of CMC, from fast-moving chatrooms to the much slower forms of e-mails or online fora (Crystal 2006).

Generally, it is possible to differentiate between two major groups of CMC. Synchronous CMC, such as chatrooms or instant messaging, allows people to communicate in real-time, emulating face-to-face communication. Asynchronous CMC, such as e-mail, newsgroups or online fora, on the other hand, does not require an instant response by participants, allowing them to instead answer whenever they please (Herring 1996: 1). It might bear mentioning that most technologies can be used for both synchronous and asynchronous CMC, however, common usage divides them relatively neatly into those two groups.

It might be tempting to compare synchronous CMC to spoken interaction and asynchronous CMC to the exchange of written letters, but sadly, the comparison is not that easy. For example, while spoken language is ephemeral, messages exchanged within a synchronous CMC system persist, increasing its usefulness (Herring 1999: Conversational persistence). Nevertheless, the use of synchronous CMC puts users under stronger pressures than the use ofasynchronous CMC; these pressures include time pressure, owing

to the fact that they need to spontaneously generate their responses as well as to the simple factor of typing speed, as well as social pressure (Tu 2002: 18–20). Furthermore, synchronous CMC is more disorderly (Crystal 2006: 61), increasing the demands put on its users further; however, this is mainly an issue of turn-taking, an issue that this thesis, with its focus on singular turns, will not consider.

Studies of asynchronous CMC have largely been confined to smaller-scale investigations, often of communities with a narrow focus. Montero-Fleta et al. (2009) investigate comments on online newspaper articles on politics and on football and come to the conclusion that there is a trend over time of (especially English language) online communication becoming more informal and oral, regardless the seriousness of the topic. There are many studies concerned with the use, advantages and disadvantages of the use of asynchronous CMC in a EFL teaching setting (Deris *et al.* 2015; Stapa & Shaari 2012), whose focus on the social aspects of CMC leaves little space for linguistic investigations. However, some studies of larger-scale asynchronous CMC communities do exist. Bernstein *et al.* (2011), for example, investigate the imageboard *4chan*, whose users are infamous for their use of offensive humor and their occasional forays into organized hacking. The anonymous nature of the imageboard environment – users do not choose a user or nick name, as is the case in most other communities, but are instead assigned a number by the board software – necessitates other forms of identity building. Users who belong to the in-group identify each other by their use of specialized lingo (Bernstein *et al.* 2011: 56):

> To communicate high status in the community, most users tend to turn to textual, linguistic, and visual cues. In many communities, including /b/, slang plays a role in delineating group membership (Eble 1996). […] Lack of fluency is dismissed with the phrase "LURK MOAR", asking the poster to spend more time learning about the culture of the board.

While this focus on insider language may be especially strong on 4chan, it is also common in other communities, including that which forms the focus of this thesis's investigation, Reddit: Tran & Ostendorf (2016) find that the close a user matches a sub-reddit community's language style, the more they will be accepted by the community as a whole. Interestingly, after learning a community's special language quirks and terminology and being fully accepted as a member, users then often do not want the community's language to change any further. While a community's language may continue to change over time, users will often stop changing with it – the sooner a user stops adapting to the changing conventions, the shorter the total timespan they will engage with the community (Danescu-Niculescu-Mizil *et al.* 2013). This thesis will also investigate special stylistic features of

the AskReddit community such as the use of the acronym *TL;DR*, which carries an interesting discourse function.

Asynchronous CMC's position on the written-spoken language continuum has been investigated ever since the inception of CMC. Murray (1988: 370) finds fragmentation and personal involvement in e-mails, two elements she refers to as often identified with oral language use, and Yates (1993: 19) finds oral elements especially in the level of (in)formality of the writers, as well as in the rhythm of the e-mail conversations. The idea that there is a certain amount of orality within the written communication taking place on the Internet has remained a constant in the field. In *Language and the Internet*, Crystal (2006: 31) finds that the language of the Internet "relies on characteristics belonging to both sides of the speech/writing divide", and indeed, more studies than those mentioned above have shown that there are several aspects of language use in CMC that are associated more closely with spoken language (Morrow 2006) – however, the idea that there is a general, uncrossable divide between spoken and written language is questionable in itself, which will be discussed further in section 2.2.

Finally, it is necessary to note that positing CMC as a form of language that must necessarily exist somewhere between some vaguely held notion of a standardized definitions of written and spoken language might be doing it a disservice. As described further in chapter 2.2.8, language is more multi-faceted than the view of a two-dimensional axis between two poles might imply, and there are many new language items that can be encountered exclusively (or almost exclusively) in CMC situations (Bernstein *et al.* 2011: 56; Crystal 2011: 23–24). While this thesis deals with the influence of spoken language on CMC, it may be important to cast a spotlight on some of these language items, as they can migrate the opposite direction (i.e. from CMC to spoken language).

An example given by Crystal is that of instant messaging short forms arriving in spoken language (Crystal 2011: 61; Ulaby 2006). These semi-standardized abbreviations originate mainly in the early days of short messaging where brevity was imperative – especially on cell phones, due to messages being constricted to 160 characters there. They have origins in the short messages of early cellphones, and also in online CMC, more specifically chatrooms such as IRC and instant messaging services such as Skype (Crystal 2011: 5). In recent years, they have finally entered spoken, informal use. While the terms that have managed to jump the divide are scant in number, it is still a remarkable phenomenon that shows that the interactions between different forms of communication – spoken and written, offline and online – are more interwoven than one might, at first glance,

think. Another aspect that Crystal (2011) continuously touches upon, and one that might be, on the whole, the most strongly perceived difference between offline and online communication, is that of emoticon usage. These short, multi-grapheme signs that are used to denote tone or mood, such as :-) and :-(, could and can be found everywhere on the Internet – their successors, emojis, which possess higher graphical fidelity, have entered pop culture to such an extent that Hollywood felt compelled to produce a movie about them (Leondis 2017). Both these discerning features of online communication shall also be analyzed in order to answer the thesis's second research question – before delving into other notable features of the language of online storytelling that may, indeed, be almost unique to the genre.

2.2 Discerning features of spoken language

It might at first glance seem to be obvious that there is a divide between written and spoken language – after all, the media of the two types are virtually incompatible, and writing and speaking are separately taught subskills in language learning. Where speaking is more context-dependent, flexible, informal and uses simpler structures, writing allows for more planning and revision. However, talking of one single "spoken" and "written English", respectively, is far too reductive and simplistic. As an illustrative example, professional speeches are delivered orally, but written and rehearsed in advance, while private diary entries, although produced in the medium of writing, are often spontaneous and informal (see also Murray 1988). This chapter will elaborate on those features which do distinguish written and spoken language, as well as report on those which – perhaps unintuitively – do not.

An obvious distinguishing feature of spoken language that cannot easily be replicated in a written medium is that of voice modulation. While it is possible to add emphasis to one's writing, e.g. by underlining it or, on the computer, changing the font, it comes less naturally than changing one's pitch, speed or volume when speaking, and it is very difficult to achieve the kind of nuance that comes naturally in oral production – as mentioned by Biber *et al.* (1999: 1042), this is also a problem encountered when studying corpora of oral language, as transcribing these phenomena is difficult and there is no universal guideline.

When it comes to the online writing analyzed in this thesis, the starting hypothesis is that very little, if any, of the emphasis innate to spoken language will be replicated in the data. Some ways of representing emphasis, such as putting asterisks around the emphasized

words or writing them in capital letters, do exist, but they are strongly marked, and the use of capital letters is often considered especially impolite and akin to yelling (Crystal 2011: 64) .

Perhaps the only other such clear-cut element that may be present in all forms of spoken language, but does not occur in written language, is that of "normal dysfluency" (Biber *et al.* 1999: 1048), which owes to the real-time nature of spoken language. Written language, which rarely, if ever, occurs in real-time (even synchronous CMC applications allow users to review their messages before sending them), meaning that producers of written language can and do make use of the possibilities of correction and revision before making their product available to the public. Whenever these elements *do* occur in written language, they are strongly marked and used mainly in fictional dialogue to showcase certain character aspects – mostly negatively-connotated traits such as nervousness or a diseased mind (Johnson 2008). It is therefore not expected that users of language working in the medium of writing will use depictions of stuttering, apart from instances where they wish to make a certain point by the inclusion of this strongly marked form.

Apart from these two factors, however, and as alluded to already, it is difficult to argue that there are properties of language that only apply to spoken language and are completely absent from written language, or vice versa, as will be further expounded on in the last section of this chapter, which will once more explicitly deal with the problems of a bipolar view of speech and writing. However, there are of course some features that occur more frequently in one medium than in the other. The rest of this chapter will enumerate some of the more commonly discussed of these factors, giving evidence of prior research and discussing expectations and applications concerning the present dataset. Many of these factors correlate with each other in practice and fall into a common dimension of language that Biber (1988: 107) describes as "Informational versus Involved Production"; indeed, in his analysis (Biber 1988: 128) it is obvious that all kinds of oral text genres, with the exception of prepared speeches, are located decidedly on the involved end of the scale, and later studies confirm this observation (Louwerse *et al.* 2004: 847).

2.2.1 Use of pronouns

One very apparent difference between the two modes of language is the much more frequent use of personal pronouns in spoken language. (Biber *et al.* 1999: 235) find that in spoken conversation, pronouns occur slightly more often than nouns or noun phrases; in the other (written) genres they analyzed, it is nouns that occur more often, sometimes

dramatically so (e.g. in academic writing). Of these pronouns, personal pronouns are especially strongly associated with the spoken language (Biber *et al.* 1999: 333).

One reason for this is the shared context within which most spoken conversation takes place. Conversational language strives towards simplicity and if sufficient identifying information can be taken from context, replacing a noun with a pronoun is more economic both in terms of time – since pronouns are generally very short – as well as complexity (Biber 1988: 104; Leech 2000: 694–695). Leech goes on to relate this phenomenon to the low mean phrase length of conversations in general, an aspect that will be touched upon again in chapter 2.2.7.

But even lacking a strong shared context, there are reasons for language users to choose pronouns over nouns in an informal setting (as most conversations are) – selecting correct, specific nouns require a great deal of presicion and specificity, which is necessary in genres such as academic or news writing where the precise transmission of specific pieces of information is the goal, but less so in most instances of speech (Biber 1988: 104–105).

While the use of personal pronouns – especially that for the first and second person (Biber 1988: 102) – is one of the strongest indicators for spoken language use, other types of pronouns can also be found to replace nouns or noun phrases. For the informational/involved scale mentioned above, Biber especially mentions demonstrative and indefinite pronouns as well as the pronoun *it*. On the other hand, relative pronouns, as parts of relative phrases, signify a more complex style that is generally found in writing (Biber *et al.* 2002: 32). In general, pronouns carry more generalized meaning and are, on the whole, vaguer than the noun phrases that would occur in their place in more formalized, information-focused genres (Biber 1988: 104; Biber *et al.* 2002: 28).

In an analysis of postings in an online discussion forum, Morrow (2006: 538) also finds a high frequency of personal pronouns, especially those of the first and the second person, which according to him "contribute[s] to a conversational tone"; he relates his findings to previous studies on e-mail communication, which generally assume an informal, conversational tone. Hence, even if it is not to be expected that the stories posted on AskReddit can draw from a strong shared context, since most users are strangers to each other, the general ease of use of and informality derived from the increased use of personal pronouns instead of nouns may very well be reflected in the platform's users' word choice.

2.2.2 Hedging

The term "hedging" was introduced in 1972 by Lakoff, who defines hedges as "words whose job is to make things fuzzier or less fuzzy" (Lakoff 1975: 471). This definition is a useful starting point, however there are many language forms whose status as hedges can be argued. For example, some definitions include approximators such as *often*, which are frequently used to describe data where exact figures are not available or necessary, rather than to lessen the speaker's or writer's individual commitment (Crompton 1997: 279–280). To avoid confusion, Crompton suggests the following definition: "A hedge is an item of language which a speaker uses to explicitly qualify his/her lack of commitment to the truth of a proposition he/she utters" (1997: 281) – this lack of commitment may be because the speaker might know that the utterance is imprecise or because they are uncertain about it (Biber 1988: 106). From Crompton's definition, it is also visible that the definition of a hedge has narrowed since Lakoff's first papers on the topic – nowadays, hedges are only those language items that serve to increase, but not to decrease fuzziness.

Hedges are more common in spoken language, as, according to Leech (2000: 695), hedging "allows a speaker to take refuge in strategic imprecision". Along with coordination tags, hedges belong to a group of spoken language features that allow speakers to be imprecise and to avoid specificity, similar to how more general pronouns are used instead of specific nouns or noun phrases. Biber (1988: 102) also associates hedges with involved rather than informational text production.

However, this is not to say that hedges are only encountered with any remarkable frequency in speech – there are some written genres, such as academic writing, that are also replete with hedges (Biber *et al.* 2002). There are some differences in the hedges used, however. According to Biber *et al.* (1999: 542), the hedges *sort of* and *kind of*, for example, are among the most frequent fixed phrases in conversation[2]; other hedges that only occur in speech are those that are found in coordination tags (Biber *et al.* 1999: 115) – an example given by the authors is found in the last two words of the sentence *They're all sitting down and stuff*. On the other hand, spoken language will rarely include hedges that are associated with more complicated language, such as the foreign language prefixes *pseudo-* and *crypto-*, as a logical consequence of speech's tendency towards simplicity.

[2] However, the same phrases are used in a different context in academia, as parts of noun phrases (Biber et al. 1999: 36) – one example of why the use of blind counts can be problematic.

On a broader basis, the grammatical classes of the words used as hedges are also different between writing and speech. According to Holmes (1988: 27), modal verbs and adverbials are significantly more commonly used as hedges in speech than in writing, whereas nouns and adjectives are more common in writing. However, comparing specifically *academic* writing to speech, the image is a different one (Hyland 1996: 270). Academic writing utilizes adjectives much more commonly than other genres, whereas particularly forgoing the use of modal verbs even more decisively than Holmes's general investigation. Academic use of lexical verbs and adverbials runs counter to Holmes's findings; it seems that within these categories, it is difficult to establish a precise tendency of the difference between spoken and written hedging, if one exists at all. In this thesis, the investigation into the differences between speech and writing, as concerns hedging, will therefore focus mainly on modal verbs, adjectives and nouns.

To analyze the use of hedges, the thesis will use a table of hedges based on one proposed by Lakoff (1975: 472). Although Lakoff's work has, of course, since been amended, his language items are largely used in the way Crompton's definition demands; as he states, it is incomplete, but it can be argued that it is impossible to compile a complete list of hedges. The list has been extended with some more items encountered in the data and will be presented in Appendix A. Special care has been taken to only include those items which are most frequently used in a spoken context.

2.2.3 Lexical density

The type-token ratio is a measure that describes the relative number of word repetition in a text. By putting into relation the number of unique words, or types, and the number of words at all, or tokens, one can calculate a number that reflects the variedness of a text's word use. However, the ratio decreases with a text's length, since an author will introduce fewer types over time (Köhler 2003: 93–95).

In comparison to pre-planned language, spontaneous oral interaction has a low type-token ratio. There are many word groups where speakers commonly use a very restricted set of words very frequently. In addition to pronouns, these groups also include verbs that control *that*-groups (Biber *et al.* 1999: 668) – in spoken language, mainly consisting of the verbs *think, say, know* and in American English *guess* – as well as modals, which in spoken language are mainly represented by *can, will, would* and several semi-modals such as *have to, going to* and *used to*, which are far more common in spoken than in written registers (Biber *et al.* 1999: 487–490). This phenomenon of a small amount of "favorite" words,

however, is not limited to the groups mentioned above, but can be found across the board (Leech 2000: 697–698). The semi-modal examples given before also demonstrate a problem with utilizing the type-token approach: *have to*, for example, is a different token than the two single words *have* and *to*, demonstrating that simply mapping each word to a token is imprecise at best.

Other aspects that reduce the type-token ratio include the replacement of more specific verbs by the general verb *do* as well as the frequent uses of *and* to connect clauses and of *be* to modify nouns – here, the ratio is not lowered by reducing the number of types, but by increasing the number of tokens for a specific type (Biber 1988: 106).

2.2.4 Extra-clausal constituents

A common feature of spoken language is the use of extra-clausal constituents – "expressions which can be analyzed neither as clauses nor as fragments of causes [that] may stand on their own, or precede, follow, and even interrupt a clause […]" (Dik 1997: 379), another indicator for simplified grammar in conversation (Leech 2000: 695). Extra-clausal constituents have been a recent focus of linguistic investigation. While it has been suggested to replace the term with "thetical" so as to not insinuate that these constituents are, in any way, underprivileged in relation to elements within the clause (Kaltenböck, Heine & Kuteva 2011: 856), this thesis will continue to refer to these language items as extra-clausal constituents, since their position outside the clause will be a crucial point of the investigation. They can further be divided into single-word inserts and syntactic non-clausal units, consisting of more than one word (Biber *et al.* 1999: 1082).

Interjections, for example, are a common type of insert in spoken language (Leech 2000: 705); Biber *et al.* (1999: 1083) define them as "inserts which have an exclamatory function, expressive of the speaker's emotion". Interjections can also consist of expletives, another type which is defined by their tabooness and their use in reaction to a negative experience (Biber *et al.* 1999: 1094); while they are also used for other functions, the overlap is considerable (cf. Ljung 2009) . Other items which often fit this category are those which draw attention to the interactive nature of discourse, which is obviously reduced in a CMC situation, compared to a "real" spoken interaction – so-called discourse markers (McCarthy 1993: 172). These are grouped separately here, since interjections and expletives are mainly dependent on their utterer's – sometimes involuntary – expression of emotions, whereas these other items are focused on the utterer's – also sometimes

involuntary – efforts to keep the conversation going. These items include greetings and farewells, discourse markers, attention signals, response elicitors, response forms, hesitators and forms that primarily serve to maintain face and be polite (Biber *et al.* 1999: 1085–1093).

All these classes have in common that they are both flexible and restricted; while they are not interchangeable (*Yuk!* and *Wow!* have quite different meanings), they can be often used in a variety of ways and for a variety of purposes. For an exploration of this flexibility, consider Aijmer (2016). This makes it difficult to investigate the purposes of their use in a largely quantitative study such as this; however, the simple frequency of their occurrence can, of course, be investigated.

The category of extra-clausal constituents is a rather large one, and it cannot be generalized whether their use is typical for spoken or written language; however, as implied by the above section, most of its sub-categories do tend to occur more frequently in spoken language. Discourse markers such as *well* and *now*, to give just one concrete example, are very frequent in spoken language, but are marked in all written genres with few exceptions, namely fictional writing and certain kinds of advertising (McCarthy 1993).

2.2.5 Use of tenses

When dealing with narratives that describe past events, there is a difference between the tenses utilized by the narrator depending on the medium: written narratives tend to continuously remain in the past tense, whereas in spoken narratives, narrators tend to switch to the present tense in the course of the narrative (Tannen 1982: 7). Sometimes believed to be used as an intensifying narrative device (Biber *et al.* 1999: 454), the so-called historical present tense is also seen as a measure to mark the most salient units of a narrative (Fludernik 1991); narrators may choose to use it for what could be called plot points or, as Fludernik (1991: 387) calls it, narrative aorists. The use of the historical present is not unknown to also occur in written (fictional) narratives, where it constitutes "an extension and application of [this oral pattern] to written narrative" (*ibid.*).

The use of historic present tense is only one of a variety of reasons why conversational, oral communication strongly favors the use of the present tense, whereas fictional writing is more likely to use past tense forms. Another has to do with conversation's stronger reliance on the immediate context, which all participants are aware of, i.e. the conversation is strongly tied in with and refers to the present situation (Biber *et al.* 1999: 456–457). While the first reason may point to present tense also being used in the

genre of online storytelling, as the historical present is also encountered in written narrative, the second reason may suggest that it will be used less frequently than in conversation, as the immediate context is not given.

2.2.6 Use of taboo language

One more salient feature that distinguishes spoken from written language is the use of swear words or, more generally, taboo language. Like many of the features touched upon in this chapter, swear words, once again, are considered a mark of informal text production; together with other aspects, such as for example the aforementioned hedges, they are considered to represent attitudes in a way which is not characteristic for formal language (Collins & Hollo 2010: 209). The vast majority of swear words is meant to "disturb the comfort level of the mainstream" (Battistella 2005: 99); still, while some are vulgar in a more general sense, others – epithets – are discriminatory and hurtful towards certain groups of people (Battistella 2006: 74). A finer analysis of the corpus will investigate the frequency of both more general profanity (such as *fuck* or *shit*) and of profanity that is meant to be derogatory against a certain group, whether this is based on race (*nigger*), sexual or gender identity (*fag*) or other properties – i.e. epithets (Battistella 2005: 72)[3]; the frequency of the latter group will be especially interesting in light of its stronger perceived offensiveness, the Internet's reputation for being lenient towards its users and Reddit's self-declared ambition to be an inclusive space, asking its users to not be rude, insult, or "troll" ("Reddiquette" 2017)

This is also borne out by studies such as McEnery & Xiao (2004) on the use of *fuck* in British English: said expletive is twelve times as common in speech as it is in writing. Possible reasons given include, once again, the lower level of formality in speech and possible censorship in writing (McEnery & Xiao 2004: 236). Furthermore, of all occurrences of *fuck* in spoken language, the vast majority occurs in dialogic situations (McEnery & Xiao 2004: 239). Given these findings, it seems that although informal, the written, monologic nature of the genre analyzed in this thesis might preclude swear words from occurring at a high frequency.

[3] Battistella categorizes taboo words into four different categories: epithets, profanity, vulgarity and obscenity; however, this thesis will mainly investigate the difference between epithets and the remaining three groups, i.e. non-epithets.

On the other hand, the Internet is often considered a place with highly profane, vulgar tendencies, where immature adolescents act out their desires to be politically incorrect and offensive, supported by news articles such as Wakefield (2016). While news media tend to sensationalize these tendencies, on the whole, these observations tend to be true – Internet users tend to swear frequently, no matter whether they use MySpace (Hinduja & Patchin 2008), Facebook (Kushin & Kitchener 2009) or 4chan (Bernstein *et al.* 2011). This observation may make it plausible to expect higher levels of profanity in online storytelling than one would in a similar offline genre. One other aspect worthy of mention that relates to the level of profanity is that of anonymity; studies show that if users are able to freely choose a nickname rather than be associated with their real name and identity, the general tone of the site is more vulgar and aggressive (Omernick & Sood 2013; Santana 2014). While Reddit does offer anonymity, however, the subreddit rules of AskReddit, enforced by its moderators, include moderators' "right to remove content or restrict users' posting privileges […] if it is deemed detrimental to the subreddit or to the experience of others […]", qualifying this content as "personal attacks, slurs, or comments that insult or demean a specific user or group of users", together with some more technical ways of abusing Reddit's comment system ("Ask Reddit..." n.d.). Whether this moderation is effective is a question that can only be answered subjectively; however, the data collected in this thesis might help give some valuable factual basis on whether the site's users routinely use profanity or tend to refrain from it.

2.2.7 Sentence Level Aspects

This sub-chapter will discuss the differences between spoken and written language on a sentence level. First, it will deal with the difficulty of speaking about a "sentence" when such a thing is not clearly delineated in the spoken mode, where one cannot simply orientate oneself along full stops, and where, furthermore, fragments and irregular sentences are commonplace. After considering these problems, the chapter will consider differences in the construction of regular sentences.

A sentence, as such, is actually not a very well-defined unit of *speech*. It is one such of *writing*, where it is clearly represented by a number of graphological features – the use of certain punctuation marks, such as full stops, and the capitalization of the first letter of the word following said punctuation mark – but in speech, it simply does not exist (Biber *et al.* 1999: 1039; Halliday & Matthiessen 2004: 6; Miller & Weinert 1998: 71). Sentences can be interpreted in speech, sometimes quite easily, by orientating oneself alongside clues

such as intonation and pitch, but rarely can one ever unequivocally state something to be a spoken sentence, and indeed, as both Halliday & Mathiessen and Biber *et al*. also describe, there are also no definite marks of sentence delineation. Very often, much is left up to interpretation, as shall be demonstrated by the following, made-up examples:

(1) a. I went to the left, John went to the right, and Mary went straight ahead.

b. I went to the left. John went to the right. And Mary went straight ahead.

One may argue about the stylistic merits of the second example, but it is not ungrammatical, and without the graphological features mentioned above, it would be impossible to state whether the whole utterance consisted of one or three sentences.

For this reason, talking about sentences in spoken language is impossible. Many different units of analysis have been proposed as replacements, including, amongst others, the so-called T-unit (Akinnaso 1982; O'Donnell 1974). A T-unit "contains one independent clause and the dependent clauses (if any) syntactically related to it […], it can be the equivalent of a simple sentence or a complex sentence; a compound sentence would contain more than one T-unit" (O'Donnell 1974: 103). In the same study, O'Donnell find out that there is a significant difference in the length of the average spoken and written T-unit, which to him suggests greater complexity or "syntactic density" in writing (O'Donnell 1974: 108–109). This thesis will compare the average length of the T-unit in some of the responses in the dataset to O'Donnell's findings to give an indication of the genre's syntactic density.

A phenomenon that may, perhaps, be more prototypical for dialogic speaking, but can very well also occur in more monologic speech situations, is the appearance of fragmented speech. This can occur for various reasons, including speakers reevaluating and reconstituting their utterances while already in the process of uttering them, and is rarely seen as marked, e.g. as unprofessional (Miller & Weinert 1998: 58–61). Miller & Weinert (1998: 60) give the following example, which I shall also use for demonstration purposes: *[…]if we can get Louise/I mean her mother and father/Louise's parents would give us/they've got a big car and keep the mini for the week […]* – in speech, the content that the speaker desires to communicate seems quite obvious, but in writing, the utterance seems convoluted and grammatically wrong. For Biber (1988: 106), fragmentation is, once again, an indication for a text being located on the involved side of the spectrum (i.e. of the first of his dimensions); he considers the time constraints of spoken language to be a further factor for the higher level of fragmentation in spoken utterances (1988: 43).

Fragmentation is one aspect of a larger aspect of language that occurs more frequently in informal or spoken language, that of the irregular sentence; and here, indeed, the word sentence *is* used. Quirk et al. (1985: 838–849) for example, give an exhaustive overview of irregular sentences, dividing them into three categories – the aforementioned fragmentary sentences, sentences with irregular forms and sentences that are marked as subordinate without being subordinate to a non-subordinate sentence.[4] In short, these sentences are such that they seem grammatically incorrect or incomplete, but which are still recognizable or reconstitutable as sentences. While still recognizing the problematic nature of talking about sentences in speech, figures similar to these irregular sentences do appear in speech also. Many of these are elliptical; while ellipses[5] occur both in writing and speech, the common type of ellipsis is different between the modes. In speech, situational ellipsis is more frequent, whereas in writing, structural ellipsis can be encountered more often (Leech 2000; Miller & Weinert 1998: 211; Quirk *et al.* 1985: 900). Due to the amount of existing research on ellipses and the comparatively clear delineation between a spoken and a different written usage, the analysis of the topic of irregular sentences will focus on those that are elliptical.

The differences between situational and structural ellipsis, briefly explained, is that the words missing from a structural ellipsis can be reconstructed solely from the grammatical structure of the remaining sentence, whereas in the case of a situational ellipsis, extra-linguistic knowledge is needed to be able to complete the sentence (Biber *et al.* 1999: 156–158; Quirk *et al.* 1985: 895–900). In the following examples, both taken from Quirk et al., sentence (2) contains a structural ellipsis, whereas (3) contains a situational ellipsis:

(2) I believe you are mistaken.

(3) Did you get it?

It is sufficient to have a good grasp of the English language to recognize what is missing in (2): The word *that*, which has been elided before *you*. In (3), however, theoretical language knowledge is not enough: we need to know the situation before we can say with any certainty what the speaker means by *it*.

Leaving aside issues of fragmentation and irregularity, Greenbaum & Nelson (1995: 12) present some findings that indicate that spoken language genres contain significantly

[4] They follow this category with that of the so-called nonsentences, which will not be dealt with here in greater detail, but which have some overlap with some of the extra-clausal constituents mentioned in chapter 2.2.4. The distinction between irregular and nonsentences made by Quirk *et al*. sometimes seems rather arbitrary.

[5] The definition of ellipsis utilized in this thesis is that given by Quirk et al. (1985: 884-887)

more simple clauses (or simplexes, i.e. simple clusters) and significantly fewer complex clauses than written language. The amount of compound clauses is similar, with spoken language just containing 0.3% less. However, there are large differences between different types of spoken language – looking at monologues, these contain the second-lowest number of simplexes, less than most genres of writing! The seeming predominance of simple structures therefore has more to do with their predominance in conversations, which may have any number of reasons, not the least of which is likely the time pressure under which the speakers act. Considering the type of text analyzed in this thesis, it would be wrong to compare it with conversation – the storytelling in these posts is uninterrupted and therefore much more comparable with monologues. There is a significant difference between the types of clusters – monologues, compared to any written genre, have significantly fewer complex clusters and significantly more compound clusters (Greenbaum & Nelson 1995: 12). Aside from the analysis of T-units as described above, an analysis of the frequencies of simple, complex and compound sentences would also seem feasible, as the genre's mode of writing might seem to make the task easily accomplishable; however, the site's users frequently utilize non-standard typesetting, for example using commas instead of full stops, muddying the waters on issues such as deciding whether a sentence is a simple or a compound sentence.

The topic of coordination and subordination is closely related to the topic of sentence construction, as complex sentences are the result of clause coordination, whereas compound sentences are the result of clause subordination. A study by Beaman (1984) gives an in-depth analysis of coordination and subordination in spoken and written narratives and serves as inspiration for this part of this thesis. While her initial numbers do not seem to bear this out, it still holds true, as the compound sentences created in speech often contain many clauses joined together; while the proportion of compound sentences may be higher in writing, her spoken language sentences contain compound sentences adjoining as many as 13 clauses, which does not occur in writing (Beaman 1984: 58); she makes the interesting point that many of these overly long compound sentences consist of many clauses, almost exclusively joined together by the word *and*. Maybe, she argues, this word has lost much of its original meaning and serves as a filler word with many functions, including as an indicator of order and as a method to signal the desire to keep talking, i.e. not ending the turn (1984: 61). Her findings also support the notion that subordinate clauses occur more frequently in writing than in speech (1984: 78).

2.2.8 Limitations of the view of a spoken-written continuum

At the chapter's closing, it must once-again be stressed that, while many assertions in the sections above have claimed a language feature to be more predominant in speaking or in writing, there is no simple axis with speaking at one end and writing at the other on which any given language features is positioned (Biber 1988: 24). Some genres may be produced in one medium, but bear properties of texts typically produced in the other. The difference between spoken and written language may not even be one of the strongest differentiating factors between different discourse types; there are indications that the most salient factor is the level of personal involvement by the speaker or writer (Biber 1988: 104; Murray 1988: 370; Tannen 1982: 18). Nevertheless, as Murray points out in the same place, message modality is not a random choice or one that is solely made based on circumstance, as evidenced by an instance she quotes where a conversation was moved from writing to speaking because the topic was considered sensitive; and as Biber (1988: 128) does point out, the level of involvement in a text and its modality are connected.

Many mentions have been made in the previous sub-chapters of the seeming fact that spoken language, due to how it is rooted in the present and its speakers having no way of revising or editing their utterances, is less complex than written language. Yet this might be too simplistic a view – as Beaman (1984: 79) states, "differences in syntactic complexity between the spoken and written modalities […] often turn out to result from differences in the formality and purpose or register of the discourse rather than true differences between spoken and written language". This does not completely preclude the possibility, or rather the fact, that there are fundamental differences between the two modalities; rather, they do not point to a difference in the level of complexity, but only to a difference in purposes. Therefore, it might do well to keep in mind that all of the differences described above are not qualitative judgments, but only quite dispassionate observations, and that a lack of complexity in some areas, such as word use or sentence construction, is weighed up in other areas, such as the necessity to include the present context and the inability to edit one's utterances.

3 Methodology: Corpus linguistics

This chapter will give a brief overview of the methods and history of corpus linguistics, which is relevant to the thesis as its methodology is heavily based on this linguistic subdiscipline. After a short introduction into the field, there will be a special focus on its possibilities and limitations. The chapter will close with a review of existing work done especially in the context of online language use.

3.1 The corpus

Corpora, in the definition used in corpus linguistics, are a relatively young phenomenon. The definition, which was once used to refer to any (somewhat comprehensive) collection of texts, has undergone a change, as described by McEnery & Wilson (2001: 29) explain:

> "the term 'corpus is simply Latin for 'body', hence a corpus may be defined as any body of text. […] But the term 'corpus' when used in the context of modern linguistic tends most frequently to have more specific connotations than this simple definition provides for. These may be considered under four main headings: *sampling and representativeness[,] finite size[,] machine-readable form[, and] a standard reference*." (emphasis my own)

These four aspects of a modern corpus will now be elaborated on in further detail.

Corpora are sampled for certain goals; they are not simply random collections of texts, but are rather intended to aid in the analysis of a certain variety of language. To this end, a corpus should be representative of the variety under investigation. However, complete representativeness is not possible, as it would demand each and every text of that variety to be included in the corpus. However, corpora do strive to include a large number of texts, getting more representative the larger they get; another factor that increases their representativeness is the inclusion of texts from as many different sources as possible (Hunston 2009: 160–161; T. McEnery & Wilson 2001: 29-30; 77-81). Even then, however, it seems difficult to create a perfectly representative corpus, since a large number of variables need to be taken into account: not just topic and speaker diversity, but also questions of gender and of power relations among others (Hunston 2009: 161–162). It is necessary to weigh the idea of equal representation against the idea of realistic representation, i.e. if a certain field or genre is, for example, dominated by members of one gender, it might be more useful for certain tasks to sample more texts produced by the dominant gender, while it might be better suited for other tasks to collect an equal amount of data from both genders. Since the main goal of this thesis is to analyze the language of storytelling on AskReddit "as it is", rather than divided by social categories, taking

countermeasures against biased sampling for certain groups was not deemed helpful. In any case, due to the anonymity afforded to Reddit users by the conventions of using nicknames and of offering as little identifying data as possible, it is practically impossible to account for any variables that the user did not intend to divulge: if a person's gender or age, for example, is not relevant for a story, it will simply not be stated.

Apart from the subcategory of monitor corpora, corpora have a finite size, meaning a certain quantity of data is represented; the BNC, for example, consists of 100 million words, to which no new material is added. Because of their static nature, conventional corpora are in danger of becoming outdated; monitor corpora, on the other hand, while more flexible, are constantly changing and therefore less suited for the generation of authoritative quantitative data (T. McEnery & Wilson 2001: 30–31). Due to the relatively restricted size of this project and the additional criteria added to weed out unwanted types of threads (further discussed in chapter 3.4), the given corpus comprises a relatively small selection of 124 threads collected over a time period of 31 days. Collection was mainly restrained by time constraints, leading to the corpus's finite size defined by the duration of data collection (a month) instead of word count (a relatively arbitrary seeming 711,529 words).

The large size of a corpus means that analyzing it by hand is impossible, or at the least not very sensible. The aid given by computer programs in searching, comparing and compiling certain types of data is invaluable, in comparison to the expenses and fallibility associated with early corpus linguists' analysis by large amounts of human helpers (McEnery & Wilson 2001: 12–13). A unique problem, however, occurs if analyzing a corpus of linguistic data acquired from a digital source, since the corpus needs to be both machine-readable and human-readable – something that cannot be trivially assumed: web pages and CMC programs are optimized for human readability, but may not be machine-readable; the raw data behind a comfortable, human-optimized interface may, on the other hand, quite possibly not be very accessible to human eyes. The challenges that occur at the intersection of corpus linguistics and CMC are elaborated upon more closely in chapter 3.3 of this thesis.

The final point made by McEnery & Wilson (2001: 32) is rather similar to the first one: that it should be possible to consider a corpus a "standard reference for the language variety which it represents". The connection to the aspect of sampling and representativeness lies within the selection criteria – if the sampling is biased, the corpus is not only flawed for its immediate purposes, but can also not be considered a standard

reference. However, McEnery & Wilson also argue that a corpus's state of being a standard reference means that it should be widely available, so that different studies on the same variety of language are able to use the same underlying set of data. This is an attribute that is not taken into consideration during the corpus compilation phase, but instead in later stages, when data collection is complete and matters of distribution and availability need to be considered. Due to the small nature of this project relative to other corpus projects and its rather specialized targets, it is unrealistic to strive to be a standard reference – something that is less prevalent in the intersection of corpus linguistics and the Internet in any case (see also chapter 3.3 of this thesis)

3.2 Possibilities and limitations

Corpus linguistics fulfills an irreplaceable role in linguistics as it allows for the quantitative study of language in a way no other sub-discipline can. Statistical analysis of corpus data can help researchers generate information on aspects such as word frequency, type-token ratio and concordancing.

To analyze word frequency, the simplest way is to generate frequency lists for each lemma. Oftentimes, however, this is not enough, for example when analyzing lexemes, different variants of one lexeme caused by different bound morphemes must be categorized in the same category (Evison 2010: 123–126; T. McEnery & Wilson 2001: 92). Whatever the criteria for sorting words into categories, the final result is a list which gives the number of occurrences for each word. This list can then be used in a variety of ways: among other uses, it can be compared to other corpora of the same languages to find similarities and differences from their frequency lists, in effect mapping the difference between the language varieties covered by the two corpora (Clancy 2010: 88–89; Hunston 2009: 160; Rayson & Garside 2000), it can be compared to corpora of different languages to study differences in language use (M. McCarthy & O'Keeffe 2010: 11), or it can be used for language teaching, such as in the so-called lexical approach, which is based on research into the most commonly and frequently used languages and collocations of a language (cf. Richards & Rodgers 2001: 132–140). Finding out the type-token ratio of a certain corpus is then a relatively easy task and can be accomplished by dividing the number of unique items on the frequency list by the number of words as a whole.

Concordancing refers to a process which is also called "key word in context analysis" which displays all instances of an item, which can be a word, part of a word or a whole

phrase, in context (Evison 2010: 128; Tribble 2010) – displaying the word in the center and showing some context to the left and right of it (see figure 1). It is especially useful for collocation analysis and for researching the frequency of different uses of a polysemous word (Evison 2010: 130). The resulting table is called a concordance; these tables have also already existed before modern corpus analysis was made possible by the emergence of affordable, powerful computers. Like the term *corpus* itself, the term *concordance* has undergone a shift in meaning: initially meaning an index of all words in a literary work, mainly useful for scholars of literature (and, in centuries past, theology), it still means an index of words, although the subject of the index has changed dramatically (M. McCarthy & O'Keeffe 2010: 3).

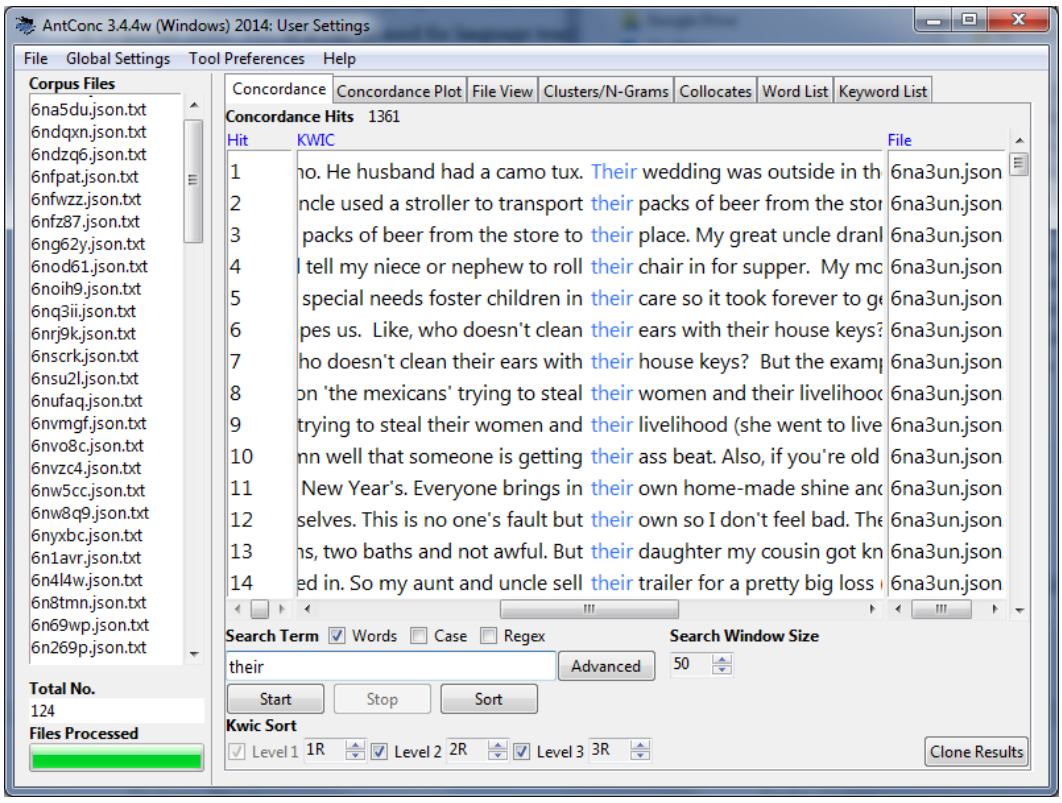


Figure 1: Concordance

However, while corpora may be very useful for quantitative analysis and modern corpus analysis software reduces both the tedium for the researcher and the margin of error by a large margin, there are also some difficulties and limitations regarding corpus use, which shall be elaborated upon in the following paragraphs.

As alluded to before, one of the most important aspects of a corpus is its representativeness; a corpus that cannot accurately represent the type of language it purposes to reflect has limited usefulness. While the difficulties of creating a representative corpus have already been mentioned, even if corpus creators have taken all measures to

maximize representativeness, random chance can lead to rare language items being over- or common language items being underrepresented. The larger the corpus, the more it is possible to avoid these problems occurring, however, especially small, specialized corpora such as the one constructed for this thesis can easily lead to wrong conclusions simply based on the sample being skewed in this way. While this specific problem may not be very pressing, it already indicates a general problem with corpus linguistics: with the computer being such a useful tool, it is tempting to rely on its results without much further investigation. Human control of even seemingly trivial results is still necessary because corpus analysis programs do not understand language semantically – it is not possible for a program, for example, to distinguish between homographs without additional meta-information, and indeed, this has been a repeatedly occurring difficulty throughout the whole project. The mere possibility of making blind counts (i.e., utilizing word frequency tables generated by corpus analysis programs without further scrutiny by the researcher) possibly enables laziness and lack of exactitude in some contexts; on the other hand, checking each and every occurrence of a word by hand is an impossible proposition due to lack of time and resources, and indeed, blind counts have been judiciously utilized in some parts of this thesis. While there may be some general guidelines for the level of scrutiny applied by the researcher (for example, blind counts of very common words, such as *to* or *be*, will very probably include many instances that the researcher didn't intend to find, whereas searches for very specific, technical vocabulary will rarely yield false positives), it is ultimately left up to the judgment of the researcher what to do with a frequency list. For this and other reasons, argue, it can therefore not be said that corpus linguistics is a way of analyzing language objectively and unambiguously.

3.3 Corpus analysis and the Internet

Using corpus linguistics to analyze texts collected from the Internet carries its own difficulties, related mostly to the relative non-prescriptiveness of online language on the one hand and to the environment from which the texts were collected on the other hand.

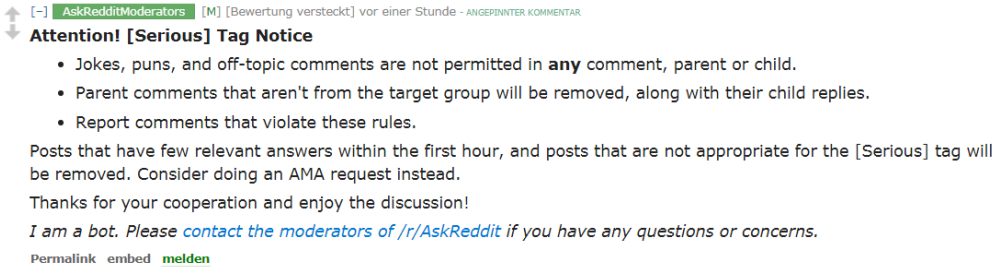
The looseness of the English language in web use has been touched upon before, e.g. in chapter 2.1. To give a concrete, practical example, in standard English each word generally has one form of written representation, whereas on the Internet, the word *please* can also be spelled *pls* or *plz*, or be modified in many different forms, for example vowel repetition in order to empathize (e.g. *pleeeeeaaase*, with varying vowel count). Therefore,

in searching for certain words, one must take into account all possible spellings; still, this does not guarantee that one finds all occurrences of this word, owing to misspellings and especially creative ways of writing. The same applies also for other Netspeak exclusive language items, which often can be expressed in a wide variety of ways (cf. Beißwenger & Storrer 2009: 302–303; the topic has also been touched upon in chapter 2.1 of this thesis).

While many of the environmental issues are especially strong in the analysis of synchronous CMC – issues such as timestamping and transmission lag carry less weight in an environment where real-time communication is less intrinsic, i.e. asynchronous CMC – there are several issues that still need to be considered. These include the format of the environment, which can offer metatextual information that may be reflected, but not explicitly stated within the text proper, such as status messages in instant messaging clients or additional information on a user, such as his status of a community moderator, which is often stated in the profile shown next to the message (Beißwenger & Storrer 2009: 299–300). Other aspects of the text may not, or only with great difficulty, be able to be stored for technical reasons, such as additional formatting for emphasis: the text, shown in figure 2 in the usual human-optimized form Reddit posts take, is represented in the corpus as follows:

> **Attention! [Serious] Tag Notice** * Jokes, puns, and off-topic comments are not permitted in **any** comment, parent or child. * Parent comments that aren't from the target group will be removed, along with their child replies. […]

In this case, the JSON data format chosen by Reddit *does* store formatting, but it does so in a way that is difficult to parse for the human reader.


[–] AskRedditModerators [M] [Bewertung versteckt] vor einer Stunde - ANGEPINNTER KOMMENTAR

Attention! [Serious] Tag Notice

- Jokes, puns, and off-topic comments are not permitted in **any** comment, parent or child.
- Parent comments that aren't from the target group will be removed, along with their child replies.
- Report comments that violate these rules.

Posts that have few relevant answers within the first hour, and posts that are not appropriate for the [Serious] tag will be removed. Consider doing an AMA request instead.

Thanks for your cooperation and enjoy the discussion!

I am a bot. Please contact the moderators of /r/AskReddit if you have any questions or concerns.

Permalink embed melden

Figure 2: Example of a Reddit post in its usual, human-readable form[6]

This example also shows another difficulty that researchers compiling corpora from the Internet need to be aware of: the software used by CMC applications often sends messages of its own. These can be either preprogrammed, such as chat clients displaying users who

[6] Taken from https://www.reddit.com/r/AskReddit/comments/6szuc5/serious_teachers_who_had_a_student_who_commit/dlhdpjd/ (accessed Aug 11, 2017); the same message can be found in any thread tagged as "serious".

have just joined or left a chatroom (Crystal 2006: 169), or configured by administrators or moderators of the CMC platform, such as the post shown above, which is automatically displayed above responses whenever an AskReddit thread is tagged as serious (see also chapter 3.4). These messages are certainly useful for day-to-day users of a CMC platform, alerting them to pertinent information that is not immediately visible, but they may distort the corpus. Software companies including these messages may choose to use a neutral or slightly formal register for these messages in order to be unobtrusive and to not offend, which may stick out in environments which cultivate their own language. Even if this is not the case, these messages are not written with the intent to contribute to the conversation, or, in that moment, consciously chosen to contribute at all; they are strictly generated automatically. Finally, their repeated nature leads to overrepresentation in the corpus, wherefore corpus compilers must manually exclude them from the aggregated data.

A problem that is less based in Internet language use and more in the relative newness of the Internet itself is that of representation in corpora. CMC is not well represented in many large corpora, many of which, such as the BNC, are completed and do not admit the addition of new types of language any more. Researchers of CMC therefore often need to gather corpus data on their own. This can have advantages, for example allowing them to focus strictly on their own needs without taking into account possible future uses for the resulting corpus, but obviously, single researchers or small teams do not have the same resources at their disposal as large, dedicated teams, resulting in small, specialized, secluded corpora which often badly documented (Beißwenger & Storrer 2009: 294–295). In this vein, it has also been necessary to compile a small, dedicated corpus for the purposes of this thesis.

3.4 Corpus acquisition

While it may not be possible to find a consensus on what constitutes "typical" Internet language, owing to the large differences in sociolects between any two websites, this thesis does intend to investigate language that, if not prototypical, is not especially marked in any particular way. This excludes any platform or online community which uses specific in-group language to differentiate itself from outsiders, such as the somewhat infamous imageboard 4chan (Bernstein *et al.* 2011: 56). It also excludes any online community which targets only a very special, restricted set of users, such as e.g. fans of one particular sports team or boards specifically targeted at users of any specific gender, age, race or location. Finally, to be representative, the user base needs to be of or above a certain size.

The AskReddit community fulfils all of the thesis' criteria, i.e. it is large, relatively diverse and its users use comparatively little site-specific language, as evidenced by the findings of chapter 4.8. It forms part of the larger Reddit community, which as of writing is the ninth most accessed website on the Internet (Alexa 2017). While originally conceived as a link-sharing platform, Reddit now hosts a large amount of so-called "subreddits", platforms which serve as discussion boards for certain topics. Within a subreddit, users can post content consisting of text, images or links; each of these content items can then be responded to by other users, who can choose to either create a new comment on an item or respond to another user's comment.

AskReddit is one of the largest of these subreddits, as of writing having over 17 million subscribers ("Ask Reddit..." n.d.). Even if one assumes that the majority of these accounts are inactive, this still leaves a user base of several million users. In the community, users are restricted to post text only, in the form of questions; other users can then answer these questions. Most of these questions are rather open, such as "What makes you incredibly uncomfortable?"[7], leading to longer answers which very often contain stories the users have (allegedly) experienced themselves. Since users can give other users points for good answers, users are incentivized to post answers that are either particularly insightful or with which (they think) the majority of users can agree or empathize.

There are certainly valid criticisms regarding the choice of platform – US citizens as well as males are overrepresented on the platform, according to Alexa (2017). However, it is important to note that the community's open nature does not specifically discourage anybody from posting. This openness, as well as the community's size, make it an appropriate choice for the investigation of relatively unmarked Internet language.

The platform's software does allow thread creators to categorize their own threads as "serious", causing a stricter enforcement of rules and disallowing jokes and off-topic comments within the thread, and as "NSFW" (not safe for work), which means that the thread creator expects the answers to contain content that is sexual, violent or otherwise not suitable for the workplace, during creation. Apart from the deletion of joke answers in "serious" threads, these categories are not enforced very strictly; of course, it is also possible to give serious answers in other threads, and it is also not prohibited to tell off-color anecdotes in threads that are not tagged as NSFW. There is no further categorization

[7] https://www.reddit.com/r/AskReddit/comments/6pnjy8/what_makes_you_incredibly_uncomfortable/, accessed 2017-07-26.

of threads, especially none that is content-related; users must decide on their own whether a particular question strikes their interest and merits viewing or not.

The Reddit platform allows users to fetch any web site stored on the platform not only in the commonly-used form, which is optimized for human readability; rather, it is also possible to acquire these sites in the JSON data interchange format, which its creators describe as both "easy for humans to read and write [and] easy for machines to parse and generate" ("JSON" n.d.), trading some readability and comfort for ease of processing with a computer. To acquire a sizable corpus of AskReddit responses, I wrote several Linux shell scripts which downloaded the JSON files corresponding to the current top "hot" 25 threads on AskReddit and consequently parsed and filtered them to exclude unnecessary meta information, automated top level posts such as the aforementioned reminders in case a post was declared as "serious", as well as any post not on the top level. The scripts were executed periodically in the time period from July 14th to August 15th, 2017. The commented scripts can be found in Appendix B.

There are several arguments for the restriction of data collection to the top "hot" 25 threads:

- Subreddits are presented to users 25 threads at a time. Unless users change their personal settings, these threads are selected by an algorithm that determines a certain "hotness" value for each thread, mainly based on the thread's age and popularity, intending to offer new and popular content to viewers. Therefore, a selection of the "hottest" 25 threads is arguably most closely representative of what an average user will see at any given time.
- Sampling *all* threads created after a certain date might seem to give a more well-rounded impression of the community's language. However, the less "hot" a thread is, the more likely it is to be unpopular, and since popularity is voted on by the community (via an "upvote" system that allows users to give or subtract a single point to a total score attached to each thread or post), it allows insight into what a community deems acceptable or even adhering to a certain standard.
- A further argument against the collection of less popular threads is that due to their status of low visibility, they mostly generate very few answers. Due to this project's limited size and the necessity of further (automatized) manipulation of the data to achieve an easily readable corpus format, the effort invested into the

collection of these less popular threads would have exceeded the possible yield of usable data.

One further modification to the corpus was necessary, as while storytelling posts are frequent on AskReddit, the open nature of the community which allows any kind of question to be posted also allows for questions whose replies are rather short, often predicated on puns and wordplay – a common kind of question asks users to modify film titles to describe some aspect of their lives. To account for this, all threads with an average comment length of less than thirty words were discarded from the corpus. The remaining data was then analyzed with the freeware corpus analysis software AntConc (Anthony 2014).

The choice to use methods of corpus linguistics in research implicitly means that quantitative research is used. For many of the features discussed in chapter 2.2, building keyword lists and concordances seems appropriate. This specifically includes pronoun use – where it is a simple matter of counting the number of pronoun occurrences and comparing them to similar data generated from other corpora, dealing with other textual genres – as well as hedging, the type-token ratio and, after a fashion, the analysis of extra-clausal constituents. There are a number of problems associated with the latter aspect, however, as the topic covers a rather large number of language features, some of which, those which are unique language forms utilized in no other occasions, (e.g. interjections) are better suited for quantitative analysis as others (e.g. tag questions). Finally, ellipses are rather difficult to locate using frequency analyses and concordances and therefore bear investigating in another, qualitative fashion. All these investigations will be supplemented by comparison to data from other corpora and linguistic genres, both indirectly via Biber *et al.* (1999) and directly. To preclude lemma problems as mentioned in chapter 3.2, a lemma list (Anthony n.d.) was used to search the corpus.

4 Analysis

This chapter will be devoted to the findings gained from the data collected from AskReddit. Each sub-chapter will correspond to a sub-chapter of chapter 2 and present findings regarding one of the features of spoken language listed there. Every item will be discussed in matters of quantity (e.g. the relative frequency of the item), of quality (by giving examples from the corpus) and of implications for the genre.

4.1 Use of pronouns

This chapter will discuss the use of pronouns by users of the AskReddit community. As discussed in chapter 2.2.1, the favoring of pronouns over "full" nouns would indicate a tendency towards the spoken and the informal. On the other hand, in the anonymous, worldwide online environment of the Reddit platform, users cannot presume a shared context to exist for their stories, which would indicate pronouns use to be less frequent than in actual, spoken face-to-face conversation. The interplay of these factors would point towards a pronoun-to-noun ratio somewhere in between written and spoken language; as this chapter will reveal, this is indeed the case.

The corpus includes 75,026 personal pronouns, which alone makes up for more than ten percent of the material. Adjusted for a million tokens, this means 104181 personal pronouns per million words, which falls between the measured amounts for conversation and fictional writing (Biber *et al.* 1999: 333), however leaning more strongly towards fictional writing. The somewhat larger amount compared to "regular" fictional writing may be explained by a simpler, less elegant style utilized in online storytelling posts, which unlike published fictional writing does not undergo a professional editing process. One major reason for the much larger number of personal pronouns in conversation that online storytelling does not include is repetition, which forms part of normal dysfluency – writers of online storytelling may not care for their story to seem dysfluent and as such, unprofessional or possessed of a linguistic ineptitude.

Of these 75026 pronouns, 59966 – or about 81 percent – are nominative personal pronouns; the rest are accusative. Among the nominative pronouns, the first person singular pronoun *I* is by far the most common, making up approximately two fifths. Another frequent pronoun is the third person singular *it*. Other pronouns are rarer, fluctuating around five thousand occurrences. It might be interesting to note that there are significantly fewer female than male third person singular pronouns – a common occurrence, as Biber *et al.* (1999: 333) note – and that the second person pronoun *you* occurs comparatively

infrequently, even though it does not differentiate between singular and plural. Contractions such as *I'm* and *you're* are included in the count, since they are partly made up of pronouns.

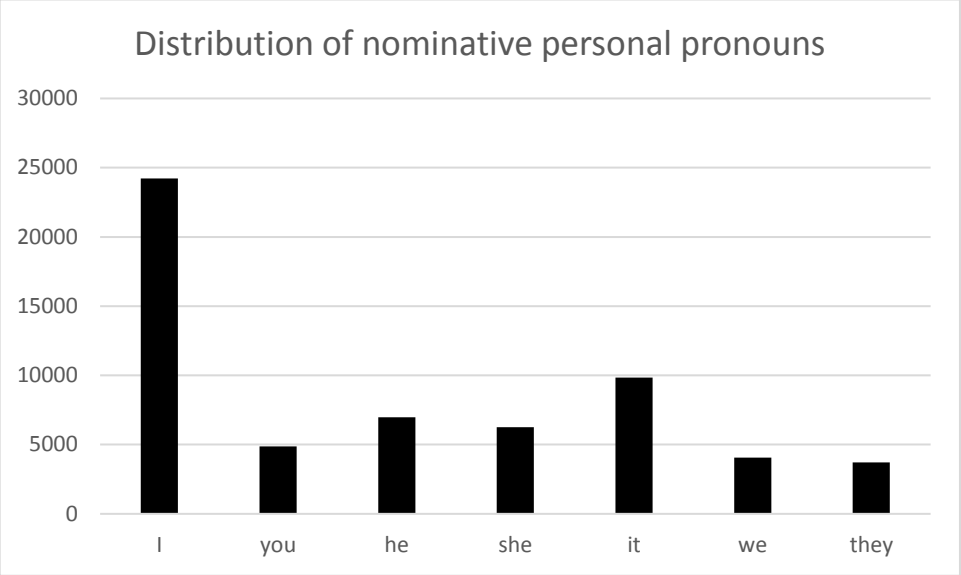


Figure 3: Distribution of nominative personal pronouns

Among the accusative pronouns, distribution patterns are similar. It is notable that *us* is even less frequent among accusative pronouns, relatively, than *we* among nominative pronouns. There is some difficulty in obtaining the exact frequency of the accuastive pronoun *her*, since the third person singular female possessive pronoun takes the same form; from a random sample of 300 occurrences, it was determined that approximately 55% of all instances of the word *her* are used as a personal pronoun in the dataset, the rest being possessives. In the following graph, only the personal pronouns for *her* are shown.

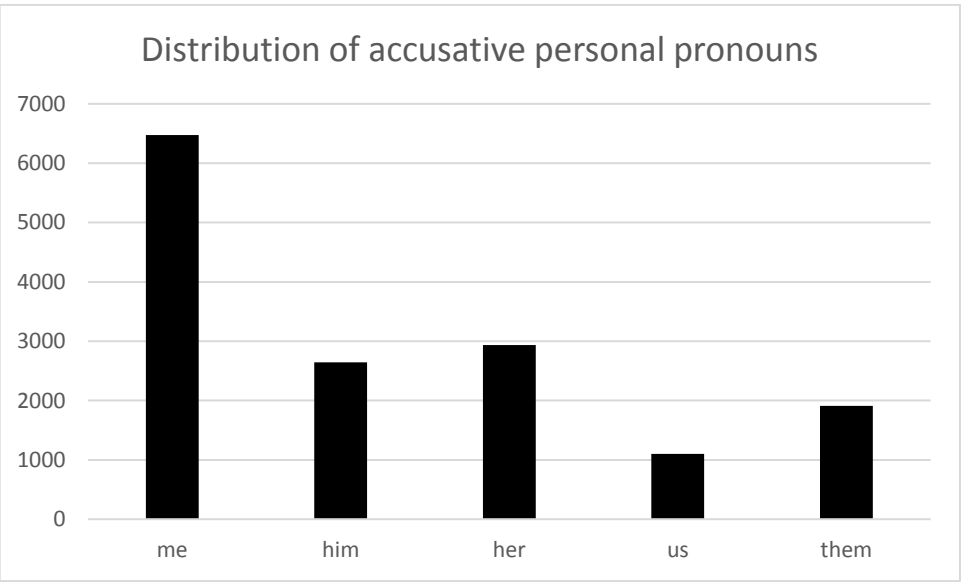


Figure 4: Distribution of accusative personal pronouns

Interestingly enough, there are more instances of *her* than of *him*, the opposite of their corresponding nominal pronouns. There may be several reasons for this phenomenon – do users of AskReddit cast men as active protagonists and women as passive subjects to whom things happen? This is a thought-provoking question this thesis cannot answer by itself, but which will be touched upon again in chapter 4.6.

In comparison with the usage patterns analyzed by Biber *et al.* (2009: 334), it is interesting to note that the strong predominance of *I* corresponds with findings related to spoken conversation. However, in conversation, the pronoun *it* can also be found to occur twice as often as in online storytelling, and the third most frequent personal pronoun is *you*, which is not borne out by the data collected in the corpus – in this case, with the relatively clear reason that in conversation, speakers have a clear opposite which they address, whereas in online storytelling, writers have a more passive, generalized audience. In comparison with fictional writing, on the other hand, *I* occurs more frequently in online storytelling. The following table compares Biber *et al.*'s findings in conversation and fiction with the findings of this thesis in online storytelling. The quantities are adjusted for occurrences per million words; Biber *et al.*'s data is only available in approximations of thousands.

	online storytelling	conversation	fiction
I	34,052	38,000	17,000
me	9,096	4,000	4,000
we	5,714	7,000	3,000
us	1,547	1,000	1,000
you	6,846	30,000	11,000
he	9,811	11,000	17,000
him	3,715	2,000	5,000
she	8,794	8,000	10,000
her	4,122	1,000	3,000
it	1,3834	28,000	13,000
they	5,227	10,000	5,000
them	2,686	4,000	3,000

Table 1: Personal pronoun distribution per genre

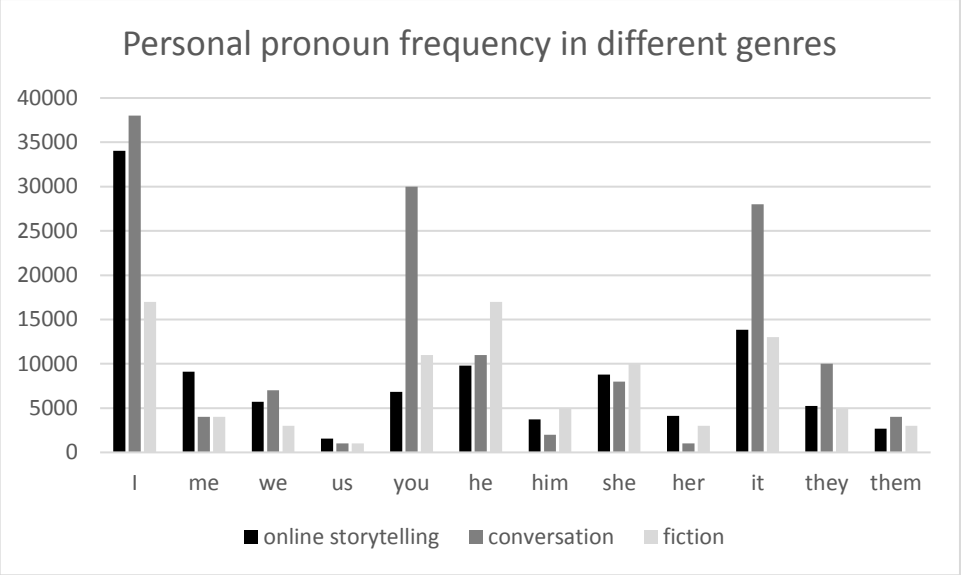


Figure 5: Personal pronoun frequency in different genres

A case-sensitive search for the misspelled first person singular pronoun *i* yields 836 hits; interestingly, many of these hits are not embedded within stories that, as one might expect in an online setting, ignore case-sensitive spelling completely. Sometimes, the right and the wrong case occur side by side in the same sentence:

(1) Sometimes **I** go on those chat rooms online for kids and say you're I'm a 12 year old girl and **i** like older guys. I've done this and I've almost gotten Amazon gift cards before. (6t7z2o)[8]

(2) As **I** left the store, **i** see a girl leaving with the exact shirt I asked for in the exact size I asked for. (6t1gpg)

A search for *u*, which is commonly used in some online communities as a substitution for *you*, however, yields very few results. When one subtracts all mentions of user names (which in Reddit take the form of "/u/[nick name]", one is left with only eight instances. That some of these seem to be parodic or sarcastic, such as in (3), suggests that Reddit users try to distance themselves from more extreme forms of Internet jargon.

(3) I looked him up on Bebo and sent him a message saying `**u** wer so gud in ur skwl play lol im 12 btw lol`. Shockingly, he never replied. (6ndzq6)

The predominance of *I* and *me* is easily explained by the fact that users mostly recount events that they have experienced themselves; questions also do often prompt them to write about their own experiences. It also matches Biber's (1988: 102) assertion of frequent first person pronoun use in informal speech; however, the relatively rare use of *we* does not fit

[8] Quotes from the corpus are referenced by the name of the file in which they appear. These file names have been automatically designated by the Reddit system and are therefore not very easily comprehensible for human eyes. All files have been included on the CD-ROM accompanying this thesis, as their inclusion as a printed appendix would have been prohibitive due to their dimension.

that observation. Concordances do not immediately reveal why that is the case; it does not appear to be supplanted by the phrase *[person] and I* especially often. Perhaps, Reddit users rarely tell anecdotes that happened within a group of friends since that might increase the overhead of context necessary (e.g. the explanation of inside jokes or group dynamics), perhaps the users of the site are more individualistic than the average person or perhaps even, the common cliché of frequent Internet users as loners holds some truth.

On the contrary, as mentioned above, it is rather obvious why the second person pronoun is used so much less frequently. Since communication is not synchronous or dialogical, users rarely have an immediate counterpart with which to interact. Threads that specifically ask for help in a certain matter are an exception to that rule and make up the majority of *you* instances, as they address either the thread creator or other users of Reddit directly, such as in (4) and (5). The difference between advice and pure story threads is quite apparent – in figure 6, the middle bar shows the distribution of *you* in advice threads, whereas the other bars show the distribution in story threads (each occurrence of the word *you* is displayed as a black column):

(4) Check what vaccinations **you** need to get as soon as **you** decide on a date/place, because some take a month long process. (6q4bja)

(5) I strongly recommend that **you** get active in club life. If **you** want a leadership position in a good club (and **you** do, trust me), **you** have to join as a Freshman. (6q5ylt)

Figure 6: Different distributions of you *in advice and story threads*

Other occurrences of *you* include quotations and stock phrases such as *you know* or *thank you*:

(6) In her embarrassment, she says “Fuck **you**, I'm getting a flight home.” (6stqk7)

(7) Because **you** know, sleep talking... UGH. (6ndzq6)

Taking all other types of pronouns into consideration also[9], the number of pronouns increases to 126,901 tokens, or 178,350 per million words. Of these, 20,954, or 16.5%, are possessive pronouns and 1,082, or 0.9%, are reflexive pronouns. These three large groups of pronouns which in the wider sense refer to persons are thus distributed as follows:

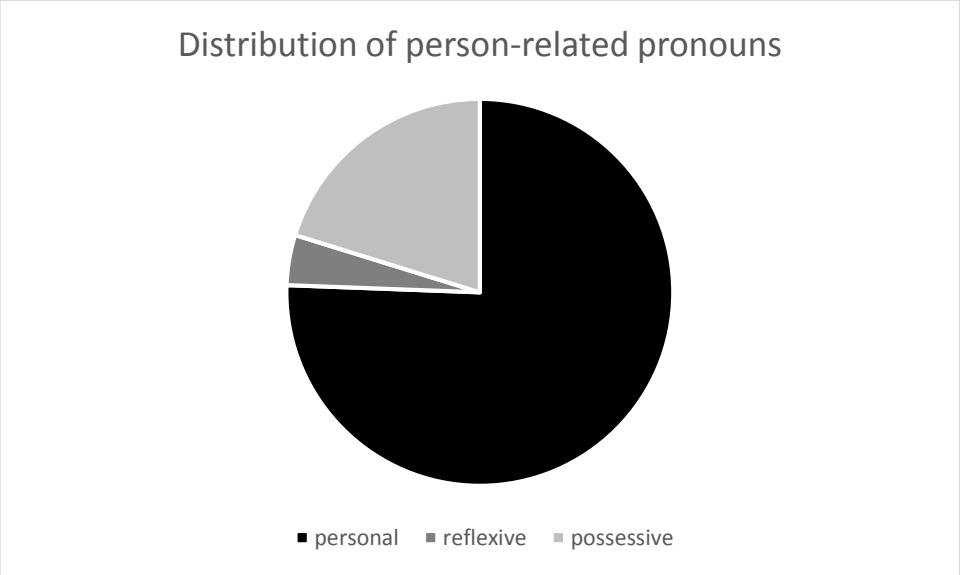


Figure 7: Distribution of person-related pronouns

An analysis of demonstrative pronouns is complicated by the fact that all words falling into these groups can also be used for other purposes – at least, it is possible to use them as demonstrative adverbs, but especially *that* is a multi-purpose word that can be used in many different cases. To find out the true distribution of these words only as demonstrative pronouns, averages were again used. In the case of *these* and *those*, which all in all occur less than 1,000 times, this average was calculated from 100 occurrences; for *this* and *that*, from 300 occurrences, as with *hers* above. The results are as follows:

word	**occurrences as demonstrative pronoun**
this	1,488
these	65
that	1,477.2
those	1,03.2

Table 2: Demonstrative pronoun distribution

All in all, according to these values, the corpus contains approximately 3133 demonstrative pronouns; the distribution of personal, reflexive and demonstrative pronouns

[9] The full table of pronouns used for this analysis can be found in Appendix E.

in online storytelling, compared to the results from conversation and fictional writing researched by Biber *et al.* (1999: 333), is as follows:

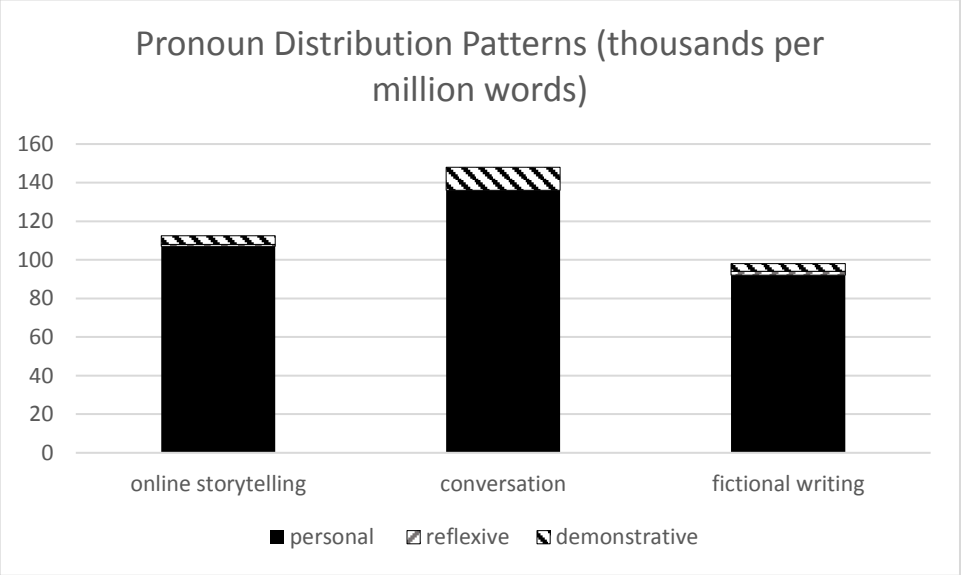


Figure 8: Pronoun distribution patterns

All in all, these distribution patterns seem to be much more similar to fictional writing than to conversation, which is understandable owing to the asynchronous nature of the genre and its written medium. While the number of personal pronouns is situated very much in the middle of the numbers observed for the other genres, the number of demonstrative pronouns – apart from *that* – in online storytelling is even lower than that in fictional writing, again according to the data from Biber *et al.* (1999: 333):

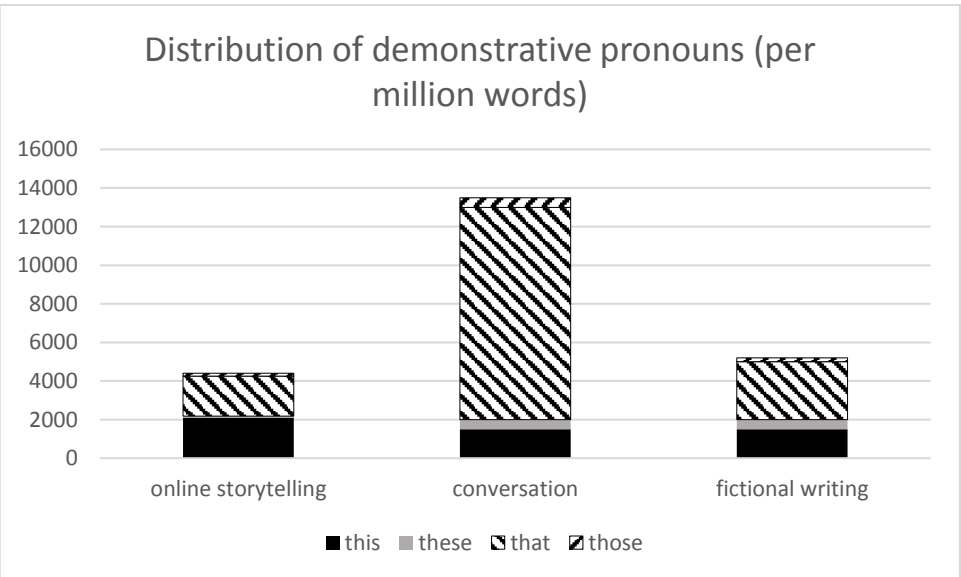


Figure 9: Distribution of demonstrative pronouns

4.2 Hedging

This section will discuss the use of hedges in online storytelling. They are generally associated with conversation (although not solely so; academic writing also utilizes hedges frequently, for example – see chapter 2.2.2.) and so might be expected to occur frequently. Before starting with the analysis proper, it must be noted that there is no objective list of language forms that can be considered hedges, maybe even less than with the other aspects of language dealt with in this thesis. While Lakoff (1975) has offered such a list in his initial paper on the concept, it is of limited usefulness due to the significant semantic change that the concept of the hedge has underwent since then. Nevertheless, this chapter looks at some commonly used language items that are relatively exclusively used for hedging, before venturing into an analysis of hedging by word type.

The list of items analyzed individually is rather short, since there are almost no language items used for hedging that do not also serve a second purpose. Language items such as *sort of* can be used for other purposes than hedging. This kind of use was however deemed infrequent enough to allow for these items' inclusion in this investigation. The following items were chosen based on criteria given by Hyland 1994 and Crompton 1997; they are subjectively ranked from most to least formal. Instances where items were used for purposes other than hedging were excluded.

language items	**total instances in the corpus**
it appears (that)	2
it seems (that)	25
likely	79
somewhat	30
sort of	126
a bit	204

Table 3: Hedges

However, while the number of instances increases the less formal the hedges used are, these results might not be very significant. These are just a small selection of possible hedges, and it might be that the more formal hedges in this list are also comparatively rarely used in more formal settings. Still, it would appear that there is arguably a predominance of the types of hedges used in conversation, rather than those used in academia.

An investigation into the word types discussed by Holmes (1988) is difficult due to the fact that most of these types can also be used to mean other things. Nevertheless, the

following paragraphs shall try to give a quick overview of what can be found out about the distribution of different types of hedges, and whether the distribution pattern can be said to correspond to spoken or written language.

Holmes claims that modal verbs are a word type that is used as a hedge especially often in speech. According to her investigations (1988: 28-29), the most frequently used of these are *will, would* and *might*. *Will* and *would*, as future markers, can be said to always have some kind of predictive function and hence, their higher frequency as predictive hedge is rather logical. However, not all of these are predictions – *I/we will […]*, for example, is less of a hedge and more of a way to state a plan and are therefore excluded from the count. Without these, *will* occurs 138 times in the corpus. *Would* is used almost exclusively for constructions such as *I* ***would*** *take the bus into the city to see her* (6nod61), although there are some rare predictive instances such as *[…] I need to chase thoughts out of my head the* ***would*** *destroy me* (6oz1ci). *Might*, again, whose use as a hedge is to express the possibility of an event (Holmes 1988: 29), and idiomatic uses such as *[as] one might think* are rather rare. Excluding these, one arrives at 200 instances. Again, it is not trivial to compare these instances with those referenced in Holmes, as it is possible that the researchers quoted there might have had other criteria for applicability, but if one does, one comes to the following conclusion (frequency given per 1,000 words):

word	Holmes spoken (25,000)	Holmes written (25,000)	online storytelling (711,529)
will	1.92	2.44	0.17
would	4.8	1.88	(negligible)
might	1.64	0.44	0.28

Table 4: Modal hedge distributions per genre

As a conclusion, modal verb usage can be said to be rather rare in online storytelling.

While modal verbs are considered to be more frequent in spoken language, nouns are especially frequent in academic language. In Holmes's figures (1988: 36-37), most of the nouns expressing epistemic modality are much more frequent in academic writing than in speech (with *chance* and *idea* being exceptions to the rule). The following table will examine the frequency of use of a selection of the most common of these nouns, comparing speech, academic writing and online storytelling (frequency given per million words and rounded):

word	speech (Lund)	academic writing (Brown & LOB)	online storytelling
evidence	31.3	168.8	52
possibility	40.6	156.3	14.1
estimate	3.1	59.4	2.8
assumption	0	50	11.2
tendency	9.4	46.9	2.8
idea	122.9	25	372.4
chance	59.4	18.8	134.9

Table 5: Noun hedge distribution per genre

The last two rows contain the outliers *idea* and *chance*, and interestingly, here it seems that the usage of these words in online storytelling corresponds more to speech than to academic writing. Whereas there does not seem to be any clear pattern regarding individual words, it is noticeable that for the two words more commonly used in speech, the same holds true for online storytelling.

In conclusion, it has to be said that hedging patterns of different genres seem unpredictable. Modal verb use for hedging seems to be much rarer than in both spoken and academic use; nouns use to show epistemic modality seem to be used somewhat frequently, and in a pattern more strongly corresponding to how they are used in speech. Adverbial hedge use is extremely common, completely different to Holmes's findings (1988: 27), where they are used more frequently in speech than in writing, but not to such an extent. Therefore, even if there are definite tendencies present in the genre, I hesitate to use them to position online storytelling alongside the spoken-written continuum, making this sub-investigation a prime example of why the reservations treated in chapter 2.2.8 are relevant.

4.3 Lexical density

With 24,905 types and 711,529 tokens, the corpus as a whole has a type-token ratio of 0.035, which of course means that only 3.5% of all words in the corpus are in fact distinct, the rest being made up from repetitions of these words. However, a total type-token ratio bears little information, as the type-token ratio typically decreases over time, as fewer new types are introduced into a text with increasing length, the most common words having already been introduced. Furthermore, some of these distinct types actually are part of the same lemma.

To make the lexical density of text genres comparable, still, Biber *et al.* (1999: 53–54) analyze the type-token ratios of the different text genres in stretches of fixed lengths. In order to rule out outliers, they do this with several stretches of text, averaging the results. Their results make obvious that length is an extremely important factor; the differences between the type-token ratios of texts of 100 words and of 10,000 words are immense.

Again, their results will be used as point of reference for the results generated on the basis of the AskReddit corpus; hence this study will also investigate the type-token ratio of stretches of 100, of 1,000 and of 10,000 words. Based on the means of 25 random 100- and 1,000-word stretches as well as 24 10,000-word stretches taken from twenty random threads, the following picture emerges:

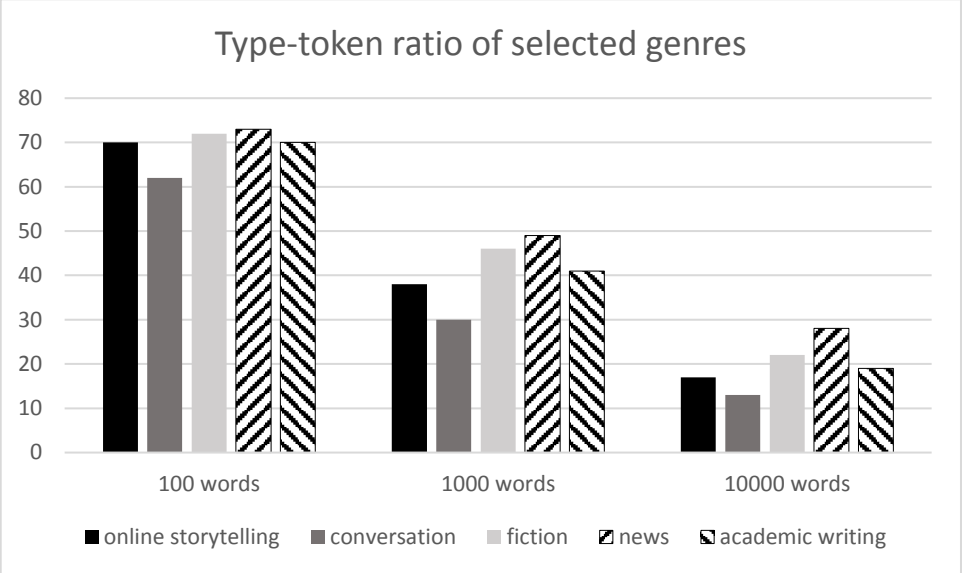


Figure 10: Type-token ratio of selected genres

As seen in figure 10, online storytelling consistently has the second-lowest type-token ratio (especially in the longer stretches of text). Thus, the genre is situated between conversation, which is restricted for a variety of reason including spontaneity and shared context, and academic writing, which is also restricted because it uses specific technical vocabulary for which it is rarely possible to use substitutions (Biber *et al.* 1999: 53–54). In other words, online storytelling is posited between spoken and written genres. However, the distinction between conversation on the one hand and written genres on the other hand is quite clear, and online storytelling seems much more closely aligned with the numbers found for other genres of writing than for conversation. Some of this may be due to the spoken genre being conversation, rather than a more monological genre, but it might also be connected to writing's higher complexity, as mentioned, for example, by Leech (2000). All in all, as with

pronoun usage, online storytelling here, too, is positioned between spoken and other written genres, but with a tendency towards the type-token ratio of writing.

4.4 Extra-clausal constituents

This chapter examines the use of extra-clausal constituents by AskReddit users – first holistically, using a table of common constituents from Biber *et al.* (1999: 1096), and then more in-depth regarding several of the subtypes of extra-clausal constituents mentioned in chapter 2.2.4.

The list of common extra-clausal constituents mentioned above is reproduced in the following table. The results of the analysis of the AskReddit corpus is compared with the results from Biber *et al.*'s analysis of American English conversation, since AskReddit is a mainly US-centric website; numbers are per million words and rounded.

word	**online storytelling**	**AE conversation**
oh	339	800
well	992	600
you know	180	450
I mean	80	200
yes	(252)	150
yeah	(190)	1,150
no	(2,517)[10]	550
mm	4	1,000
uh huh	--	1,500
okay/ok	406	550
uh/er	83	650
um/erm	20	300

Table 6: Common extra-clausal constituents per genre

As expected, with extra-clausal constituents being very much a hallmark of spoken language, the number of occurrences is significantly lower in online storytelling due to its written medium. The numbers for *well* are approximate due to the adverb *well*'s status as a homophone; *yes*, *yeah* and *no* obviously do occur throughout the corpus, but very rarely as extra-clausal constituents such as response forms. The values in the table are inside

[10] Words in parentheses are displayed for the sake of completeness, but have little informational value due to their frequent usage outside of extra-clausal constituents.

parentheses as they represent the total number of these forms, hence they cannot really be compared with the results of Biber *et al.*'s study: blind counts are always problematic, but in this case especially so, as the use of the forms in other ways than as extra-clausal constituents is frequent. When these forms *are* used as extra-clausal constituents, it is often within representations of direct speech:

(8) "Hmmm - Was it a fulltime job?" "**Yes** I guess so. […]" (6nvmgf)

(9) I say **no**, it's okay, I'll just tell the sound guy to skip the duet. (6noih9)

Even so, *yeah* seems underrepresented in comparison to real spoken interaction. This might be due to its routine, automatic use in speech, which users may not find important to represent in writing. Interestingly, *yeah* is often used as a slightly sarcastic or cynical marker that contrasts a point of view portrayed immediately before, such as in (13) and (14).

(10) […] 'women coming to the New World' to start families with the settlers? **Yeah**, they were all prostitutes. (6n5du)

(11) I was stupid enough to continue seeing her when she promised she ended it etc. **Yeah** she never did. (6nfpat)

Constituents that seem especially "oral" are used the least – utterances that are seen more as vocalizations than words, such as the hesitation markers *uh* and *um*, which are generally consciously avoided, are represented less than 50 times in a million words. When they are used, its often to represent spoken language, or as what could be described as a slightly sarcastic hedge:

(12) […] says, `What are you doing right now?" I say, "**Uh**, poopin?" So then he says, "Look, I'm gonna have […] (6phxp3)

(13) […] Call them and tell them we need seating for 8." "**Uh**, ma'am, that's not something we do." "Then […] (6sjip7)

(14) […] security than I thought. And the surrounding... **um**...village had a four room bed and breakfast. We […] (6nsu2l)

As table 6 shows, the frequency of these extra-clausal constituents in online storytelling does not seem to be related to their frequency in conversation. One may be able to say that they generally occur in fewer instances than in conversation, but even so, *oh*, for example, seems to occur with some regularity in online storytelling, whereas *uh* seems to be almost unused, despites its frequency in conversation. The reason may seem obvious – *uh* is almost never used consciously, and speakers indeed try to minimize the frequency with which they produce these hesitation sounds, whereas *oh* can carry some meaning.

Oh is the most common example of an interjection, along with *ah*, *wow*, *(wh)oops*, *ugh*, *ow/ouch*, *(a)argh*, *urgh* and *hm* as well as some other, more peripheral and less frequently used utterances (Biber *et al.* 1999: 1083–1085; Leech 2000: 697). These are the forms investigated in this thesis by means of a frequency analysis based on a blind count (since none of these forms are ambiguous or homographs of other words); further qualitative investigation will focus on the question whether these interjections are mainly used in representations of speech or whether AskReddit users also use them outside of these.

Interjection	**Total count**	**Count per million words**
oh	241	338.7
ah	10	14.1
wow	30	42.2
(wh)oops	8	11.2
ugh	15	21.1
ow/ouch	7	9.8
(a)argh	0	0
urgh	0	0
hm(mm)	4	5.6

Table 7: Interjection counts

As the counts above demonstrate, the frequency of interjections in the genre of online storytelling are quite low, pointing towards a more "written" style. The findings by Biber *et al.* (1999) show interjections to be more frequent by more than an order of magnitude in speech. For example, in their analysis of conversation, *oh* occurs approximately 8,000 times per million and *ah* approximately 300 times per million[11]. The corresponding counts per million for online storytelling are 339 for *oh* and 14 for *ah*, respectively.

The following examples show some of the usage patterns of *oh* in the genre. The interjection appears especially frequently either at the beginning of a post (as in (18) and (19)) and at the start of a stretch of text that is intended to represent direct speech (as in (20) and (21)):

(15) **Oh** boy. So my freshman year of high school I took AP European History (AP Euro) named Mr. Regar. […] (6nfz87)

(16) **Oh** I've got this one. My fiancé planned a girls trip with her bridesmaids out of town to visit another one of her bridesmaids, V […] (6s9kqq)

[11] in American English, which is assumed to be the more common variety of English used on AskReddit; it is far more frequent still in British English.

(17) […] he tried to blame it on my sister. ‘**OH** it's because your daughter was not following my instructions. […] (6qrrpk)

(18) […] When I asked about the postcards the mechanic said `**oh**, yeah, I get a few of those each summer. […] (6t5e2n)

Interestingly and as shown in the examples above, in the instances where *oh* occurs at the beginning of a post, these post-initial sentences is not used to immediately transport important information, or “plot” – rather, they are often used to set the tone of the story or to prepare the reader for what the author thinks is an especially shocking or sordid tale. It often collocates as *oh boy* or *oh man*, which could both be classified as interjections as well based on the definition given in chapter 2.2.4.

A third way in which *oh* is used is not at the beginning, but towards the end of longer posts or paragraphs when the authors intend to add one more detail that goes along well with the rest of their story. In this case, it is used at the start of the sentence, collocating with *and*:

(19) […] **Oh** and he was a narcissistic babbling logorrheic, who couldn't STFU when walking down the street. […] (6o854p)

(20) **Oh**, and he eats stinky onion sandwiches in the office every day. I hate him. […] (6nfwzz)

All three usages are similar in the sense that the *oh* is not provoked by a sudden reaction towards something occurring in the context of writing, but is rather used as an eye-catcher or to structure the story.

It may be of little use to provide a closer analyses of all the other interjections due to their relative infrequency. *Wow* is used as an example here, with the other cases being rather similar. *Wow*’s relative frequency can be explained by the fact that it is also an abbreviation for the popular online role-playing game *World of Warcraft* (as in (24)), causing four false positives, and its usage in “edits” (parts of the post added by the user after the original post has been published and others have had a chance to read it) which thank readers for their positive reception (as in (25)):

(21) […] I stick with a sad routine of work, **WoW**, Netflix, and doing a little working out a few times a week... sometimes. […] (6q2h9l)

(22) […] Edit: **Wow**, I did not expect this much traction. First off, thank you for all the condolences and support. […] (6sp1wr)

Other than that, the usage patterns are similar to those of *oh* – post-initial and at the beginning of representations of direct speech.

The other type of extra-clausal constituents investigated in-depth in this chapter is that of expletives. There is some significant overlap with the chapters on taboo language as concerns lexical items; the difference is in the focus of the investigation, as this chapter focuses on practical usage whereas chapter 4.6 focuses on semantic content. Furthermore, of course, not all taboo words are used as expletives – of course this chapter solely deals with those instances where such a usage can be confirmed. For the difference between expletive and non-expletive usage, let us recall Biber *et al.* (1999: 1094)'s definition as "taboo expressions […] used as exclamations, *especially in reaction to some strongly negative experience*" (emphasis my own). Using taboo words, such as e.g. *fucking*, for emphasis, does not constitute its use as an expletive under their definition, which this chapter follows; this, for example, excludes all adjectives on their own (*"Fucking!" would seem rather strange to most speakers of English). Since it seems difficult to get an exact count of all expletives while excluding non-expletive taboo words – and also since it may sometimes be a matter of interpretation if an instance of a word falls into one category or the other – the analysis will be mainly qualitative. Generally, taboo words are rather rare, as will be further elucidated in chapter 4.6 – it may therefore be assumed in this chapter that expletives do not occur in any significant number.

Fuck, for example, is an example of a word that can occur in any number of contexts. Apart from the possibility of expletive usage, it can also be a verb with a variety of meanings and a pejorative way of referring to a person of disdain as well as being part of a number of figures of speech. Examples of clear expletive usage (which can include composites such as *holy fuck*) include the following:

(23) […] Holy **fuck**, this is the first time I've had an overwhelming amount of people interested in my story, well the story is long and way more things happen. […] (6tei79)

(24) Gas Blower he starts up every time there is anything on his street or driveway, just pick it up. **Fuck**!! (6nscrk)

(25) […] As my eyes slowly opened they were greeted with a full blast of pepper spray. Holy **Fuck**! What the hell! […] (6sp1wr)

(26) […] **Fuck**, when I helped my parents do some yardwork once and we were cleaning up near the property line, *his entire family came out to watch us.*[…] (6ouc9l)

While (26) and (27) refer to strong reactions the narrator has "right now" (i.e. at the time of telling the story), and (28) and (29) seem to indicate their emotions at the time of the

narration, the salient point seems to be the strength of the (negative) emotion felt by the narrator.

Another frequently used expletive is *shit,* whose frequency and usage in the corpus is similar to that of *fuck*, reinforcing these results: it is used in an expletive fashion mainly as a reaction to strong emotion and in the representation of thought or speech. Both expletives – and, by extension, likely others as well – occur, as Biber *et al.* (1999: 1094) find, mainly in initial position or after verbs of thinking or saying (e.g. *I mean **shit,** I'd love to* (6p103x)).

All in all, these findings indicate that extra-clausal constituents – at least the two categories analyzed in more detail – are used sparingly in online storytelling on AskReddit. Frequent usage of interjections and expletives point towards would indicate a more speech-like use of language, which on the other hand means that the relatively low frequency of their occurrences indicates that Reddit users tend towards a more "written" style. This, in turn, might suggest that these stories are not written in the spur of the moment (maybe in a somewhat rambling and spontaneous manner) and immediately posted, but are instead planned and revised before publishing. Finally, when these language patterns *do* occur, they are used relatively prototypically – one might even say conservatively.

4.5 Use of tenses

This chapter describes the usage of tensed verb forms in the corpus, especially focusing on past and present tense, as explained in chapter 2.2.5. Since a close analysis of all verbs is beyond the scope of this thesis, it will focus on a list of some of the most common verbs of the English language; the number of occurrences of their tensed forms will be counted and juxtaposed.

The investigation will focus on the following points:

- A quantitative analysis of the present and past simple forms of the 30 most frequent English language verbs. Special attention will be paid to the frequency of speech-act verbs used in the historic present, as elaborated upon below.
- An investigation into whether the historic present is only used for certain parts of the text – as is the case in the genre of conversation – or for whole stories.
- A qualitative analysis of the differences between tense use within direct speech and outside of direct speech. Both conversation and fictional writing are genres the frequently incorporate direct reported speech, of course also incorporating the pertinent temporal deictic shift. This step is therefore intended to ascertain

whether present tense occurrences are largely restricted to direct speech or whether there is indeed, as hypothesized, a sizable presence of the historic present.

Some of the forms have been deleted since there has been some difficulty investigating verbs such as, for example, *must* (which has no past form at all) or *put* (where present and past forms are the same). Another danger with especially frequently used verbs is that their usage may very well differ significantly from the average usage of less frequently used verbs, necessitating the inclusion of research on phenomena such as speech-act verbs, especially *say*, being used in the historic present especially frequently, with *say* alone accounting for 35% of verb usage in the historic present (Biber *et al.* 1999: 455; Wolfson 1979: 179).

The list of the thirty most frequent verbs used for this analysis is taken and adapted from a frequency analysis of the BNC done by Leech, Rayson & Wilson (2001) and can be found in Appendix C.

All in all, for most verbs, the frequency of present forms found in the text outweighs the frequency of past forms, if only slightly. This can have several reasons, such as present forms often coinciding with the verb's noun counterpart, such as with *need* or *use*, or coincidental analogy to an unrelated noun (such as in *mean*), or also the aforementioned possibility of large amounts of present use in direct reported speech. The median number of present forms for the list of verbs utilized lies at 788.5 times in the corpus, while it lies at 564 for past forms. Past forms overtake present forms concerning the mean, however, with 1290.47 to 1062.57 occurrences.

The fact that the mean is larger for past forms whereas the median is larger for present forms points to the presence of a strong statistical outlier favoring the past forms. This discrepancy is due to the one verb which is most frequent by far, which is *to be* in all its forms. Even though derived forms such as *I'm, we're* etc. were, of course, also counted, there are more than double the amount of past forms than there are present forms, which is not representative of the common trend. Other verbs that strongly go against the general grain include *will*, whose past form, *would*, is used for different and more numerous purposes than its present form, and *shall*, whose present form is barely utilized today at all. However, all in all, present forms still outweigh past forms in frequency – as can be ascertained also from their larger median amount.

Curiously, *say* is another verb whose past forms outnumber its present forms. While it might be expected that due to its especially frequent usage with the historic present, *say* might present itself to be a strong outlier in favor of present form usage, the opposite is the

case. The results of a qualitative investigation into the use of the verb can be found further towards the end of this chapter.

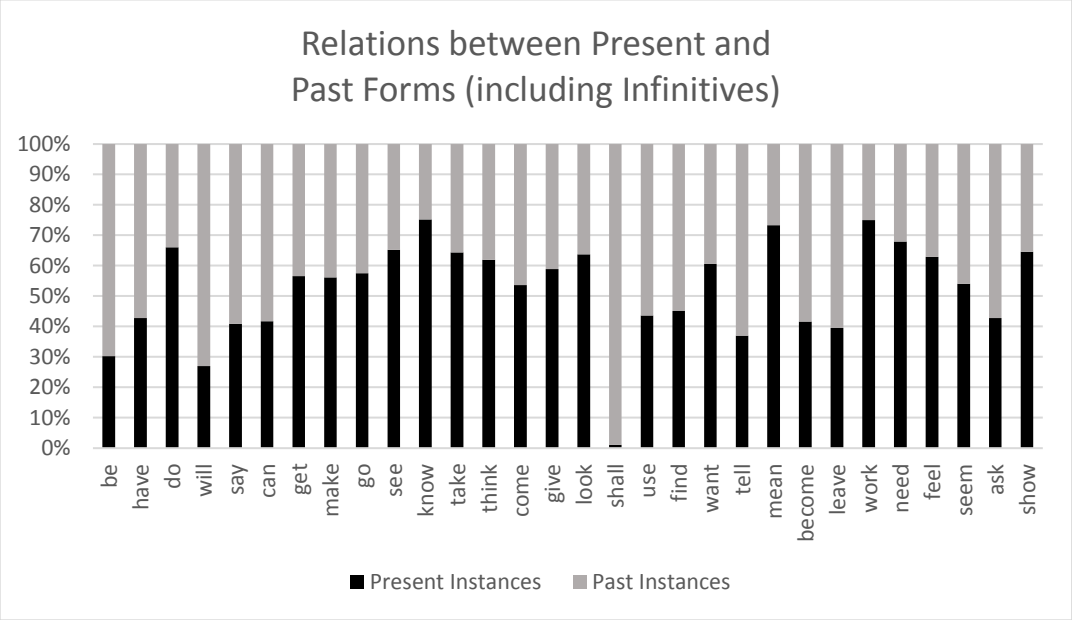


Figure 11: Relations between present and past forms (including infinitives)

As for the types of verbs that are predominantly used in their present or past forms, respectively, it is of course somewhat daring to speak of definite results, considering the relatively small number of verbs analyzed. However, it seems that those verbs that represent a mental process are especially weighted towards their present forms – represented in the sample by the verbs *know, think, want, need* and *seem*. They have a median ratio of 1.58 present forms per past form, as opposed to the general median ratio of 1.29 present forms per past form. Other trends are not immediately visible from the data.

A qualitative analysis of the data reveals that the predominance of present forms is due to a number of reasons, including the fact that present forms also double up as the infinitive, which may be one of the driving forces for their frequent use for some verbs, including *make* and *show*:

(27) The hair on my arms were turning white because there wasn't enough nutrients **to make** them the proper color. (6n8tmn)

(28) This happens a lot and gets annoying as people need **to make** themselves feel involved in this. (6ndqxn)

(29) Turns out that police coming into schools **to show** kids some drugs makes them more likely to actually use drugs. (6sjawn)

(30) It's not good enough **to show** up and do well on tests and essays. (6q5ylt)

Another diagram which shows the same data as above, but excluding infinitives constructed with *to*, looks as follows:

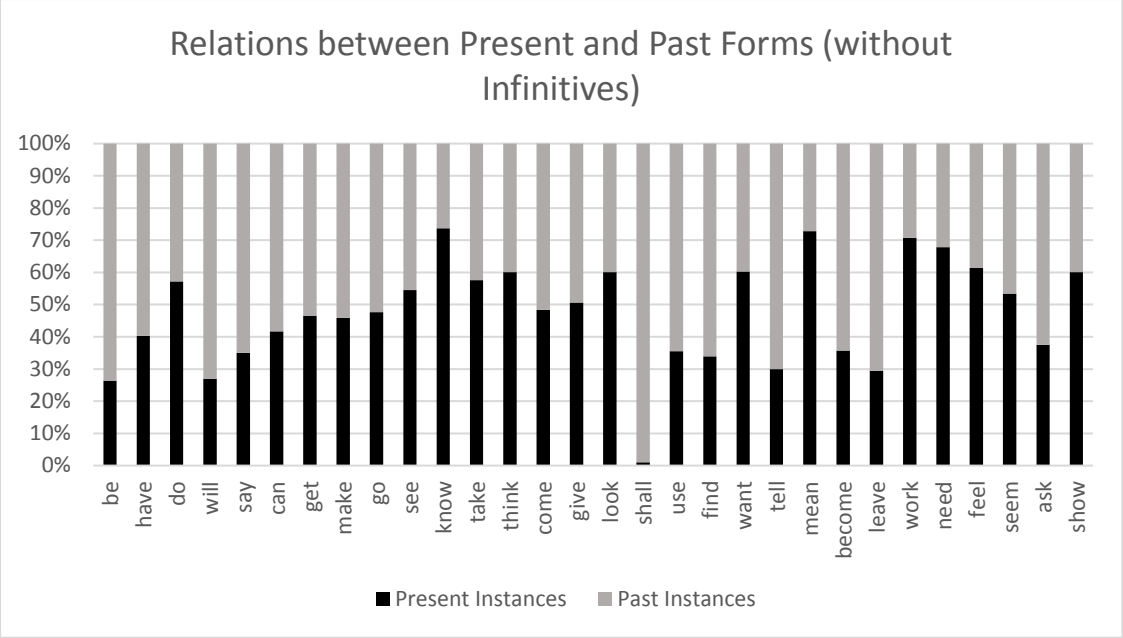


Figure 12: Relations between present and past forms (without infinitives)

The average frequency of present tense verbs moves down to 864.7, the median to 612 and the standard deviation to 1035. It lowers the average considerably still, but the median does not reach that of the past tense forms. By excluding the infinitive form, the divide between the amount of present and past forms grows larger, with past forms becoming, relatively speaking, even more common.

The examples given still demonstrate, however, that present tense use is generally frequent, especially in comparison to fictional writing. This might be due to users having to share general information about the setting, place or time in which their story takes place. Some more examples include:

(31) Apparently stuff like that **happens** pretty often in our family. (6s3h81)

(32) I **have** a camera system that's just inside the side yard entrance with big floor to ceiling windows. (6qo8y4)

(33) […] whenever I **visit** he always **goes** above and beyond all expectations to **make sure** I enjoy myself. (6n269p)

Of course, the historic present is also used in many of the stories, utilized both as intensifying device *and* as narrative aorist (Fludernik 1991). In the first of the following examples, it is used as an intensifying device; the use stretches over almost the entirety of the story. The second example, in turn, represents its use as narrative aorist, where the present form occurs only when there is a new 'plot point', so to say. Noticeably, most stretches of the historic present are longer than just one or two verb occurrences – examples

(37) and (38), already longer than the usual examples given in this thesis, omit parts of the narration for reasons of length.

(34) My friend and I **were** at somebody's birthday party. […] As the night **goes** on and we've had some drinks I **see** my friend sitting next to a girl talking to her. […] He just **looks** at me and says: `she speaks Russian`, while the girl **looks** at me in disbelief and shock in her eyes. I **mumbled** something about how sorry I am and quickly **left** […] (6nyxbc)

(35) But I **did** end up working with one of the guys who interviewed me at C*** and **was** telling this story at the Christmas party at which point he **goes** 'OMG, that was you?! […]' (6nvmgf)

Other reasons for present form usage include a number of stories being indeed told in the present tense throughout, users often giving advice or recounting lessons learned from the happenings in their stories, or other instances where the use of present tense is common, such as rhetorical questions:

(36) I **try** as hard as I can to get the bar to lock, but no avail. Customer **makes** a scene, says I**'m** not doing it right […] (6telk6)

(37) Your main goal **is** to study. You **are** not there to party, to drink or to waste time. (6q5ylt)

(38) What **happens** if the relationship goes wrong; will he react badly? (6q2h9l)

A final note on the use of present is that the dataset also includes threads that can reasonably be called storytelling threads, but which do not tell stories that are set in the past. Rather than try to describe this phenomenon in theoretical terms, I will refer to one of these threads as a practical example. This thread is designated "6ndqxn" in the dataset and originally bore the title "For those who struggle with depression and suicidal ideation - what do you most want to hear from people who want to offer comfort or help? What don't you want to hear?"[12]. It was also tagged as "[s]erious". While many responses are not "stories" in a traditional sense, they do contain and recount personal experiences; together with their consciously chosen location within a forum mainly used for storytelling, they can be argued to represent storytelling as well, in a less conventional form. Some of the shorter responses are reproduced below (all examples, of course, taken from the thread):

(39) I **want** to hear that I am cared for/loved by them. Depression **tends** to make you feel mentally alone or that people secretly hate you. Those little reminders **can** help to alleviate that anxiety.

(40) I **don't** want to hear anything. **Leave** me alone.

[12] Accessible under https://www.reddit.com/r/AskReddit/comments/6ndqxn/ (last accessed Sept 17th, 2017).

(41) An invite to an event or gathering. I might say no but if you **force** me to go I **promise** I won't regret going.

It is apparent that these answers are almost invariably written in simple present tense.

Regarding the use of the verb *say*, it seems likely that the difference between the frequent present form usage in conversation and the less frequent usage in the genre of online storytelling lies in an argument made by Wolfson (1979: 179): she postulates that the phenomenon of frequent present tense *say* usage in spoken conversation is based on "loss of significance through overuse". In writing, where each word choice is deliberate and which is typically less spontaneous, it is likely that users will opt for the form which they perceive as more correct – i.e., the past form is chosen because the present form *say*, surrounded by past form verbs, simply "looks wrong" to them, for whatever conscious or unconscious reason. There are 54 occurrences of *I say* and 87 of *he/she says* in the corpus, which, apart from being quite a small number in comparison with the magnitude of the effect measured by Wolfson, are also rarely used in the way she describes. In the cases where they are, they occur within a story that is wholly told in the present tense, thus also not fulfilling the criteria. It can therefore be concluded that the use of "[personal pronoun] say(s)" as a speech-act verb, or speech tag, within a past tense environment, is something that, while occurring frequently in spoken conversation, is very rare at best in online storytelling; this aspect indicates a strong influence by written style.

4.6 Use of taboo words

This chapter will venture slightly into the field of sociolinguistics, as the use, disuse or condemnation of profanities and swear words can illuminate much about a certain community of speakers. As mentioned in chapter 2.2.6, after a general analysis of the use of taboo words in the AskReddit community, special attention will be paid to the differences between general profanity on the one hand and epithets denigrating certain groups of people on the other hand. Another sub-chapter will investigate the use of profanity in "NSFW" threads in contrast to regular threads, which is intended to give insight into which profanities are "marked" enough to largely not enter into conversations not marked as possibly offensive – even on the Internet – and which have become commonplace.

As with the other chapters, the appendix will include a table of obscenities investigated. It needs to be stressed that some of these categorizations are somewhat

subjective due to changing perceptions: for example, *bastard* is classified here as a non-epithet, while this may have been different some decades ago – the implication on the addressee's mother's sexuality is not in the foreground any longer; in a similar loss of original meaning, a *fucker* is not necessarily a person engaging in sexual intercourse. Where in doubt, it has been decided to classify words that directly attack a part of a person's being as epithet – thus, *faggot*, directly deriding homosexual men, is classified as an epithet, whereas the aforementioned *fucker* (or, more generally, *fuck*), is not directly targeting an aspect of the addressee's sexuality, even if it might at first glance be expected; see also Battistella (2005: 72), on whose classificatory systems this section is based and who elaborates more closely. As swear words and epithets are a rather subjective matter, often interpreted wildly different in strength and degree of tabooness by different groups of people, the caveat must be made that all attempts at classifying these words into discrete categories can only ever be best-effort rather than authoritative. What has deliberately not been included are surrogate words that clearly describe one of the epithets while not using it, so as to not offend, such as the use of the phrase *the N-word*.

Most writing and research on the frequency of taboo words has been done with regards to British English, which in this case differs significantly from the English used by the mainly American audience of AskReddit, for example by the frequent use of *bloody*, which is practically absent in American English (Biber *et al.* 1999: 565); the results may, nevertheless, offer a basis for comparison with the findings of this section. By far the most common swearwords in British English, based on a study of the BNC, are *God*, *fucking* and *bloody* (McEnery 2006: 29). These words are followed by *fuck, pig, hell* and *bugger* before the first epithet – *bitch* – is recorded. A study of London teenage slang (Stenström, Andersen & Hasund 2002: 80) comes to a similar conclusion, with the exception only that *shit* is far more frequent (the exclusion of *God* from these findings may be due to a narrower definition of the term swearword). The salient point – that the first epithet is relatively low in the frequency list – remains. Whether this is also the case in the kind of online discourse that is found on AskReddit is subject of this investigation.

The sum total of all occurrences of taboo words, as enumerated in the appendix, is 2709, or 3807 per million words. Their distribution is as follows:

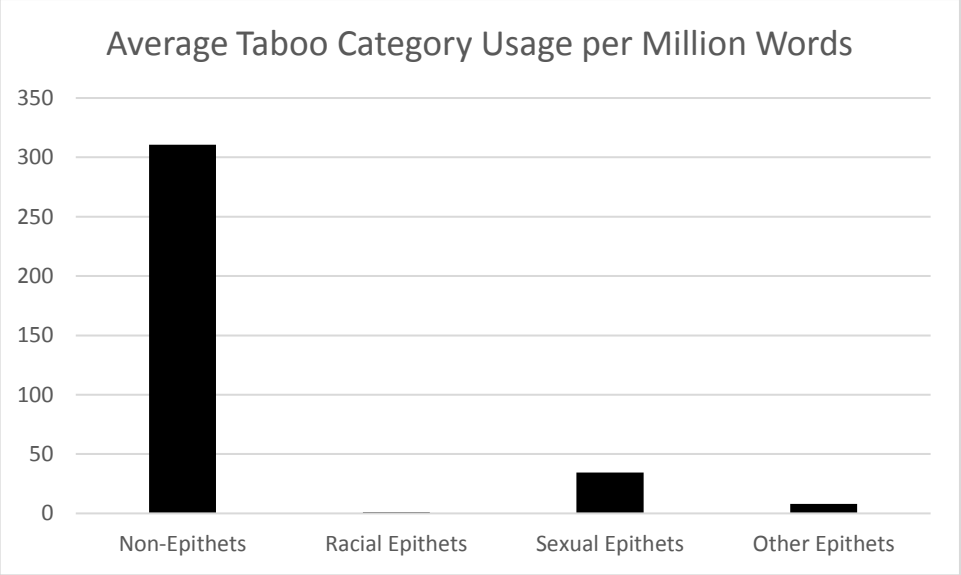


Figure 13: Average taboo category usage

This means that a (fictitious) "average" non-epithet word is used 311 times per million words, whereas a (similarly fictitious) "average" racial epithet is used only 1.05 times per million words. For both the racial and general epithet category, less than fifteen occurrences of the words enumerated in the appendix have been found; sexual epithets, while occurring significantly more often (382 times per million words), are still much less frequent than non-epithets, which occur 3417 times per million words. As expected, *fuck* and *shit* are by more the most frequent of these, occurring 899 and 710 times in the data set, respectively; the next most frequent term, *ass/arse*, occurs 175 times. All in all, these results conform to the expectations set by previous studies. While sexual epithets will be discussed in further detail in the following paragraphs, (45) and (46), given below are two examples of other epithets as used in AskReddit discourse:

(42) I've personally heard people say that they've never met a Trump supporter that wasn't **retarded**. (6obe27)

(43) So the girls dad decides to say, 'get out of my class you stupid **nigger**!' (6rri6y)

It is of note that all racial epithets are used in the way they are used in the above example – not by the narrator themselves, but by characters within their stories. Clearly, racism is a problem concerning which Reddit users are very sensitive, which cannot be said about sexism.

It may also be of note that not only are there more sexual epithets targeted at females, they are also used more frequently. 79.0% of all sexual epithets used in the corpus were those aimed at females, while only 3.7% were aimed at males. The 17.2% remaining were

neutral. This is an alarming number, and while there is no explicit rule against sexist content in Reddit's Content Policy or the so-called Reddiquette, the latter, for example, does encourage users to, amongst other things, "[r]emember the human [and] ask yourself 'Would I say [a comment] to the person's face", to "[not b]e (intentionally) rude at all" and to "[not] insult others" ("Reddiquette" 2017). On the other hand, Reddit's administrators argue that free speech should be allowed even if it may cause discomfort, as long as the content posted is legal (BBC News 2012).

The implications of this discovery definitely merit further investigation. What does it say about the website's users (or, maybe, its moderators) that they almost consistently avoid racial or ableist epithets – quite opposite to, say, 4chan (Bernstein *et al.* 2011: 53) – but do not seem to share the same sensibility regarding epithets based on (mostly female) sexual behavior and morality? As mentioned above, there are very few occurrences each of racial and ableist epithets in total, but 9 of *whore* alone (which, unlike *prostitute*, does have very strong negative connotations), 7 of *slut*, and 139 of *bitch,* although it must be noted that the verb *to bitch*, which is included in the latter count, does not have sexual connotations – examples given below. In contrast, the male-targeted epithet *fag(got)*, which is frequent on other platforms (cf. Bernstein *et al.* 2011), never occurs. These findings are in line with other research, though, that suggests that while both women and men face harassment online, it is mainly women who are harassed in a sexual manner (Duggan 2014).[13]

(44) **Bitch**, you cheated on me with over 20 men, I have to get tested for AID's and shit now because your a stinking **whore** pretending to be the opposite. (6nfpat)

(45) I knew I was right but said nothing. Fuck that saggy **slut** (6n4l4w)

(46) She was that kind of super hot **bitch** that you would typically see with some rich dude at some fancy night clubs (6nfwzz)

While many of these occurrences, such as (50) and (51), are instances where the storyteller merely quotes another person – often one that takes on an unsympathetic role in the story – it still begs the question why there is no similar phenomenon for other epithets.

(47) The boys (who were actually french speaking Belgians) were talking amongst themselves about how 'American girls were easy **sluts**' […] (6nyxbc)

[13] To be clear, sexually charged epithets are still very rare in relation to the text total, but still, the difference to other epithets is remarkable, especially since online sexual harassment towards women is a constant topic of discussion; due to the breadth and controversy of this topic, it will not be expounded upon further in this thesis.

(48) An ex can't yell `She's a **WHORE**!` and expect to stop the ceremony. (6sqyfp)

One more aspect of taboo language use that this thesis investigates is whether there is a big difference in its use in NSFW and non-NSFW threads. As explained before, the abbreviation NSFW stands for "not safe for work"; subreddits often have slightly different definitions of what content, exactly, must be declared NSFW, but most follow the definition given by Reddit itself, which says that "[c]ontent that contains nudity, pornography, or profanity, which a reasonable viewer may not want to be seen accessing in a public or formal setting such as in a workplace should be tagged as NSFW" ("Reddit Content Policy" 2017). In other words, an NSFW-tagged thread can be seen analogous to an R-rated movie – do users have less inhibitions regarding taboo language use in such a space?

Since the number of NSFW-tagged threads in the dataset is very low (only 9 out of 124 threads were tagged as NSFW), these findings must be looked at with an especially critical eye since a sample size of nine cannot be seen as especially representative. Nevertheless, there are some interesting observations: the relative proportion of NSFW-tagged threads in the dataset (7.32%) is much larger than that in the full data dump which includes those threads with shorter answers that were not included in the corpus (4.5%) – or in other words, NSFW threads tend to have longer answers on average. This may just be coincidental, but may also suggest that NSFW-tagged threads induce users to elaborate more.

While the low numbers make any kind of authoritative statement difficult, there is a noticeable difference – in NSFW threads, 0.55% of all items fall under taboo language, whereas in SFW threads, it is only 0.37%. Both numbers are very low in general – what may perhaps be the strongest conclusion that can be drawn from this chapter is that AskReddit users are not especially keen on using taboo language, perhaps choosing to express strong feelings in other fashions. Chapter 2.2.6 offers up arguments and data that suggest that swearing and the use of taboo language is, in general, not only an informal thing, but also one that is only strongly represented in spoken language; the numbers generated in this thesis would support this idea – it should not be forgotten that the total swear words found is just 2709, in a dataset that includes more than 700,000 words. Finally, it should be emphasized, that the findings from this chapter, especially, are only applicable to AskReddit and should not be generalized to other CMC situations or, even, other subreddits on the same website.

4.7 Sentence Level Aspects

Due to the nature of this section, which demands an analysis above the word level, it is largely impossible to use quantitative methods for the analysis of its topics. In order to have a manageable selection of text for qualitative analysis, the three top posts were taken from the fifteen longest threads to form a small sub-corpus for analysis.

There are few instances of sentence fragments in the corpus, which does indeed make sense considering that writers have the possibility to reread and revise their writing previous to posting it on the website. As touched upon in chapter 2.1, time pressure and turn-taking are issues that are more relevant for synchronous CMC; in asynchronous CMC, users do have time to revise their utterances, and they also make use of it. Due to the informal nature of the discourse on AskReddit, its users do not feel constrained to always use grammatically complete sentences, however, they do not venture into the realm of fragmentation as much as the example from speech in chapter 2.2.7.

The most common type of incomplete sentences in the corpus involves ellipses. The number of ellipses used in a story varies significantly from author to author. There are long stories without a single instance of ellipses, and there are others where the author chooses to use initial ellipsis (i.e. the omission of the subject) in every sentence, possibly as a conscious, stylistic choice.

Only a few types of ellipses are commonly used. By far the most common one is the sentence-initial omission of the subject, sometimes accompanied by omission of the verb. Sentence-initial ellipsis of *there is/are/was* is also rather common; other forms of situational ellipsis are rare, if they occur at all. As for structural ellipsis, the most common instance is *that*-deletion, followed by the omission of pronouns (however, it could be just as possible that pronoun omission owes to careless typing – it seems that this kind of ellipsis is especially common in stories with typographical errors). In general, structural ellipsis occurs much less frequently than situational ellipsis, which would correspond with a more casual, "spoken" style. Some examples follow below – the bold words in parentheses indicate the ellipses and are not actually present in the original text:

(49) **(She is)** Always late. **(She has)** a Loud, fake cackle laugh that would embarrass a hyena. (**She)** Takes conference calls at her desk on speakerphone

(50) **(She was)** Charming. But every now and then there would just be this onslaught of abuse. (**It was)** Nasty and hurtful

(51) I had a dream that my roommate's friend was over our place and (**that)** we were sitting on the couch […]

Other types of ellipses do occur as well, such as ellipsis in question-answer sequences (see Biber *et al.* 1999: 157), especially at the beginning of posts, when users answer the question posed in the thread's title. For example, in a thread asking users who suffer from depression how they are feeling, one user responded with *Not so great, honestly*, leaving out the obvious words "I'm feeling […]". Nevertheless, omission of subject and that-deletion are the most common kinds.

As regards coordination and subordination, quite impressive examples of both can be found in the texts. Example (55), for example, directly echoes Beaman's (1984) findings of long coordinated sentences using several instances of *and*:

(52) […] It was a nice dog **and** not really scary or aggressive at all **and** went up to the dude **and** he got our dog to get in the van **and** then they kept driving towards the house **so** my sister came in **and** locked all the doors **and** started freaking out. […] (6nq3ii)

The special place of the conjunction *and* deserves special investigation. Beaman's investigation shows it occurring approximately twice as often in speech as in writing – 72.9 times per 1000 words in speech, and 35.9 times per 1000 words in writing[14]. A simple blind counting of the total instances of *and* in the dataset reveals a count of 22482 instances, translating to 31.6 instances per 1000 words. This is even less than Beaman counted for writing, and it is very far removed from her results for speech – quite as opposed to to the examples presented above. One possible cause of this is a generally rather terse style of writing that can be found frequently in the corpus, tending towards short and simple sentences. Just looking at *and* in isolation, this finding strongly suggests that in this case, the written nature of the genre dominates over whatever influences by speech there may be.

The common subordinating conjunction *when* occurs 2,984 times in the dataset, translating into 4.19 times per 1,000 words (the amount of *when*s used as interrogative pronouns is negligible). This is more frequent even than Beaman's number for writing, which gives the impression that the writing style of online storytelling is, in this case, almost "hyper-written". Again, this might have to do with the relative terseness of the overall style, it might have to do with Beaman's relative small sample size, or it might just really be the case that most writers of online storytelling strongly prefer a style of coordination and subordination that is congruent with existing written styles.

The wide range of different styles also makes T-unit analysis difficult. In fact, the most pertinent result of the analysis resulted from the preparation for the process, when it

[14] It must be noted that Beaman's sample size is rather small.

became apparent that while much of the stories' texts conform to standard written grammar, there are many quirks on the level of sentence delineation and, more generally, construction. A relative terseness can again be found, with several T-units consisting only of one word (due to ellipses). Even in postings with long sentences, some have short T-units due to their frequent usage of coordinating conjunctions (as seen in the example above). On the other hand, there are also users whose T-units stretch over more than one sentence, as in the example "If you're looking up your gf/bf's behavior on the internet. You need to get out." (6rre0r). This construction can rather clearly be considered a single T-unit (i.e. a common conditional sentence), however, it is partitioned into two sentences by the use of a full stop. Still, however, in a random selection of postings, the results are quite extreme, with "short" T-units (of up to 18 words of length, as by O'Donnell's definition (1974:106)) taking up 91% to 100% of the text – much more extreme, in fact, than even O'Donnell's results pertaining to speech, where short T-units made up approximately 62% of the texts. Clearly, and perhaps paradoxically given the length of some of their posts, being short and to the point matters greatly to these online storytellers.

These results do not seem very congruent - how do these numbers fit with the extreme run-on sentence above? There are a number of these in the corpus, enough so they cannot be called extreme outliers. The answer simply lies within the difference of styles between individual users. Indeed, in conclusion it seems that the way these sentence- and clause-level features are used are on the whole more dependent on users' individual styles rather than on genre conventions of any kind. The numbers indicate that writers tend to keep more towards what they perceive as the written standard, but there are obvious exceptions, as indicated above.[15]

4.8 Unique features of online storytelling

Chapter 2.1 lists some aspects of online communication that are not only unique to this genre but also, in a way, emblematic. Even those who rarely use computers will, at one point, have seen the emoticon *:-)* or the abbreviation *LOL*. The first part of this chapter will analyze these two major features of online communication, whereas the second will investigate any features that are unique to online storytelling, if there are any.

[15] There might however be an argument that the seeming terseness of the language actually indicates the attempt by writers to translate the rambling, *and*-heavy, spoken way of communicating into more "writerly" prose. Connecting the sentences in example (52) with *and*s, for example, results in a text that would not feel exceedingly strange if uttered in conversation.

The first aspect to look at are emoticons, because the analysis of the corpus regarding these special pictographs gives us an immediately impressive result: they are almost never used. For something that has almost become a symbol of Internet jargon, the results are quite underwhelming – an investigation into those emoticons which are most common on Twitter (Schnoebelen 2012: 117) yields the following results for the corpus data (which, again, is comprised of more than 700,000 tokens):

emoticon	**absolute quantity in corpus**
:)	29
;)	5
:(	25
:D	4
:-)	1
:P	1
(:	1
:/	13
XD	1
=)	0

Table 8: Emoticon instances

Also, while it might not strictly fall under the heading of "emoticon", the <3 heart symbol is also used sparingly – it appears only twice in the corpus[16]. There may very well be slight differences in the emotions expressed by Twitter users in comparison to those expressed by users of AskReddit, but the overwhelming difference between these counts and those found by Schnoebelen (2012) speak for themselves. AskReddit users seem to prefer not using emoticons or emoji and thus, these items are rare in this specific genre of online storytelling. Interestingly, this is only one of several aspects of very informal online communication that is rejected by the community – most writers also pay attention to spelling and grammar and, as seen in chapter 4.1, reject the use of short forms such as *u* for *you*. On the spectrum of online language use as a whole, they might well be strongly on the conservative side.

The issue of abbreviations may be a bit more complicated, as AskReddit users do use standardized abbreviations, but usage is different from other online environments. The very

[16]To be very exact, this count applies for the character sequence *<3* – owing to the technical specifications of HTML and JSON files, the < character is saved as *<* in the corpus.

common and well-known *lol/LOL* abbreviation appears 49 times, whereas the less widely spread abbreviation *TL;DR* and its cousin *TLDR* appear 117 times. A later part of this chapter will explain the use of this abbreviation in more detail due to its interesting use (or rather, its interesting dual uses – two practices of usage have developed); however, before that section, table 8 offers a sober, quantitative look at some common Internet and short messaging abbreviations and their distribution in the corpus.

There is no common consensus on which Internet language acronyms are the most common or maybe even "canonized" in whatever way; however, studies (e.g. Crystal 2011; Kinsella 2010; Varnhagen *et al.* 2010) consistently include the forms *lol*, *brb* and *omg*. Other frequent forms include *rofl, btw* and *ttyl*. These forms can be said to either be spontaneous expressions of emotion, especially laughter, or strongly embedded in a context – *brb*, standing for "be right back", for example only really makes sense within a certain situation. It seems therefore sensible to expect these abbreviations to occur less frequently in asynchronous CMC, where language users act less spontaneously and less embedded in any kind of common context and have more time to reflect on their language use – quite apart from whatever stylistic qualms one may have about the use of these acronyms or short forms. The data bears this out: the rarity of *lol* has already been noted, and the following table contains the other aforementioned forms, as well:

Acronym	**Meaning**	**Frequency**
lol	laughing out loud	49
brb	be right back	0
omg	oh my god	13
rofl, rotfl	rolling on the floor, laughing	0
btw	by the way	11
ttyl	talk to you later	0

Table 9: Acronym instances

It bears especially mentioning that the two context-dependent abbreviations *brb* and *ttyl* never occur in the corpus.

As mentioned above, the following paragraphs will go more in-depth concerning the abbreviation *TL;DR* (the semicolon being facultative), which appears disproportionally often in AskReddit discourse – even more so when considering its most commonplace

usage, which takes place on a meta-level regarding the rest of the text.. Where the more common abbreviated forms such as *lol* and *rofl* serve as more spontaneous, context-embedded forms, *TL;DR* – whose letters come from the phrase "too long; didn't read" serves more of a meta-function: it is put in front of story summaries, often located at the very end of a post. These summaries' main function is to give the casual readers scrolling by some hint whether the preceding story might be interesting to them (Adams Sheets 2012). That the term makes little sense in its current use – why would users introduce summaries with "too long; didn't read" – hints at a change in usage. Originally, *TL;DR* was used as a somewhat caustic reply to very long texts in discussion forums – users informed the posters they were replying to that their long, time-consuming and maybe somewhat overly detailed text had fallen on deaf ears.

TL;DR's usage has evolved, and it now is an indicator of a summary, but this change has only taken place on Reddit, where it can even be said that another, further change has taken place, because *TL;DR* can seemingly be used in two different ways – both indicating summaries of the preceding text. One type tries to summarize the post in a matter of fact way, such as in (56):

(53) […] "**TL;DR**: Had 2 run of the mill jock-esque neighbors in college who ended up dropping out due to poor grades, and one of them starting an argument with racist insults in it." (6nscrk)

There is another type, such as shown in (57), where the writer tries to make this summary as vague and enticing as possible, summarizing the preceding story rather amusingly and sometimes misleadingly. In this case, the writer's ultimate goal is not to inform the reader of the story's contents and therefore to allow them to decide whether it is of interest, but to persuade them to read it no matter the content, as for example in the following, very succinct example:

(54) […] "****TL;DR**: Clever canine counters caper.**" (6ripcd)

As apparent from the above example, the latter kind of *TL;DR* applies stylistic methods such as alliteration and repetition, whereas the "original" kind rarely does so.

TL;DRs are even more frequent on other, less frequented subreddits which are even more devoted to online storytelling (in contrast to AskReddit, which does often also allow threads with shorter answers, as discussed in chapter 3.4) and are, as such, part of a site-wide culture – neither restricted to one subreddit, nor applicable to any kind of online story-telling on other websites. However, even though they pose an interesting part of Reddit's site-wide culture, their importance should also not be overstated – while it *does* prove

interesting that this rather less well-known acronym occurs more than twice as commonly as *lol*, its appearances are still constrained to a relatively meager number of 117 times in the entire corpus. More than any particular fascination or predilection with *TL;DR* on the part of AskReddit users, the takeaway from this chapter seems to be that AskReddit in general frowns on the excessive use of abbreviations, being quite stylistically conservative in comparison to other websites and the general public perception of Internet language use.

This is not only confined to the issue of acronyms and emoticons, however. In general, it seems that the AskReddit community's approach to its position as an Internet community, including all the trappings such a community has, is to quietly accept it without bringing any attention to it. A quick search for other Internet or Reddit specific terms yields few results – these include more general Internet terms such as *link* or *forum* as well as those specific to the Reddit ecosystem, with its *karma* point system which is influenced by *upvotes* and *downvotes*. While it may be expected that focused storytelling might not include these topics very frequently – after all, which storyteller at a real-life campfire would constantly refer to the fire and his audience's reactions? – these results, too, are worthy of note.

5 Conclusion

In this chapter, a concise summary of the main findings is given, supplemented by some notes of the limitations of the study and several ideas for avenues along which further research could be conducted.

5.1 Summary of main findings

This thesis has analyzed a number of features of language that are generally held to distinguish speech from writing and has applied them to a corpus of language gained from the online storytelling platform AskReddit. In many cases, the results for these different features are similar: while the genre of online storytelling may in some cases take on elements of spoken language, it is, ultimately, still rooted in the medium of writing, with which its users choose to communicate. How strong this claim can be made is different from language feature to language feature – the results are summarized briefly in the next paragraphs.

Users' pronoun usage is located between conversation and fictional writing, according to a comparison of the language data with Biber *et al.* (1999). Some features, such as very frequent usage of the first person singular pronoun *I*, correlate more strongly with conversation; others, such as pronoun distribution patterns in general, correlate more strongly with fictional writing. Both the type-token ratio and the usage of extra-clausal constituents are used in a way that correlates more closely to written usage; the same applies to tense usage: although present forms are used more frequently than in other written genres, they are less frequent than in conversation, especially concerning the use of the present form for speech-act verbs. Taboo words are rare, and on the sentence level, findings diverge widely: the authors' individual styles are much more apparent here. While there may be some writers who use long, coordinated sentences, most write rather tersely and use subordinations, as is more indicative for written language.

These findings, combined, also gives an answer to the first research question: AskReddit users indeed do use aspects of language more commonly associated with speech than with writing. They do, however, not do so to an overly large extent – for those features whose usage seem to be indeed between the two modes of communication, such as lexical density, it can generally be said that they are closer to (informal) writing than to speech.

Moving on to the second focal point of this thesis, it has started with an overview of language use in CMC and touched upon a variety of features that are said to be unique features of CMC. Of course, not every feature connected to CMC will be found in every

variety of this rich and vast repository of genres and language use, but there are several aspects that are strongly associated with Internet language as a whole. Wherever people communicate online, at least informally, one might expect short forms of text or the extremely abbreviated language items that have emerged from short messaging, and one might also expect the usage of smilies, emoticons or emoji – three generations of face-based combinations of graphemes that transport emotion. Both features – indeed, almost clichés – are rare, if at all present, in the genre of online storytelling as it can be found on AskReddit. It seems as if the common consensus of the platform is to present as little of these common features as possible. This thesis does not presume to find the meaning of this phenomenon: Do the platform's users wish to present themselves as more sophisticated or articulate than the average user of the Internet? Do they find that the interruption of text by these short forms, maybe even by pictograms, lends the text a jagged quality that makes it harder to read? (Notably, the only Internet-exclusive, and maybe even platform-exclusive, feature that users *do* utilize – the summarization of their stories under the heading of the abbreviation *TL;DR* – occurs outside of the body of the story proper.) Whatever the cause may be, the platform's users eschew these common qualities of CMC.

Another remarkable aspect of the genre – one that is less rooted in the peculiarities of CMC – is the extreme terseness of many of the texts. Ellipses, other omissions and short T-units can be found to an amount that exceeds findings even in other written genres. There are various possibilities to why this might be the case – for example, users might wish to communicate as efficiently as possible. Another possible reason is that it represents an attempt by users to "translate" the long, *and*-heavy coordinated clause structures of spoken language into writing by replacing the *and*s with periods.

Apart from this aspect, there is a general red thread found throughout the thesis: AskReddit users are not staunchly conservative and formal – their narrations are more often informal than formal – but they tell their stories in a much more conventional form than one might expect from an online platform such as they are using. The general style is terse, and the genre's lexical density is lower than in other written genres, but it always stays on this side of grammaticality. Ellipses are used, but sentence fragmentation is very rare. The historical present – found both in speech and in writing – is utilized, but the much more speech-exclusive use of *say* in the present tense within past tense narratives is not.

A final point that must be made, however, is that users have a rather large amount of leeway concerning their writing styles. The community does hold writers to a certain standard, but still, there are writers whose styles differ little from published fictional writing,

while others include more features of speech, such as long coordinated clause structures. As such, all findings in this thesis should be seen more as community trends, not as absolute statements on each and every story.

5.2 Limitations

Of course, the criteria chosen for analysis in this thesis have been carefully selected, and their results paint a certain picture. As such, it cannot replace a far more detailed and in-depth genre analysis – such a thing, however, was not feasible within the scope of the thesis, both concerning time and resources. It still holds some important points and discoveries about the genre, however.

While the selection criteria both for the dataset (i.e. those governing which threads were to be included in the corpus) and the criteria undergoing analysis were chosen with care, it cannot be ruled out that due to their incompleteness, they may paint a skewed picture; just as has been discussed in chapter 3 that in order to be absolutely representative, one would need to include absolutely every possible applicable language item into the corpus, the problem here has also been the necessity to exclude some features that could possibly discern speech and writing. As it is, the items chosen are some which possess both high visibility, often being immediately apparent even to a layperson, and high impact – it is hoped that by a combination of these factors and those further stated in their respective sub-chapters, the justification for their inclusion is obvious.

A final note regarding the limitations of this study concerns the methodology used – while much care has been taken to approach the analysis of the data given by the corpus sensibly, it would have exceeded the scope of this thesis by far to properly and qualitatively analyze each and every language item contained within. With those items that were more ambiguous or prone to false positives or negatives, averages were calculated from a smaller random sampling, but random samples can be skewed by random chance; with those items considered less at risk for these problems, blind counting has sometimes been used. Of course, these results have also all been double-checked, but it is not inconceivable that something may have slipped past these checks. Even so, however, considering a corpus of over 700,000 words, some false positives or negatives may be expected – and as mentioned in chapter 3.2, these are among the risks and limitations of corpus linguistics in general.

5.3 Avenues for further research

The analysis done within this thesis has, of course, only touched upon a few aspects of one particular facet of the genre of online storytelling. While the criteria selected have been chosen in order to hopefully gain well-rounded insight that confirms the exact position of the genre between spoken and written – or, in many aspects, rather formal or informal – discourse, there is a plethora of other matters that can be investigated, from word creativity to conjunction use. The thesis's analytical part has often referred to Biber et al. (1999), and within good reason: their analysis of four different genres concerning every conceivable aspect has made them open books that other genres can be compared with. There are many purposes for which such a detailed analysis seems desirable and applicable, ranging from teaching to marketing to sociology. Such an analysis might not only serve scientific interests, but also create a better understanding for online interactions, as the Internet and its emerging culture is still something that many find opaque and difficult to understand.

And yet, this thesis has focused only on a small part of what is possible in an online discussion forum. One possible avenue of research might, for example, involve the use of inter- and metatextuality that users of online storytelling websites often use. While perusing the corpus, one may find many hyperlinks to website, images or videos, presented in a variety of different ways – ranging from their use as argumentative support to seething sarcasm. Many times, these links form part of a sort of in-group experience, akin to an inside joke; often, they are standardized, called "memes" in the Internet jargon. This kind of communication seems quite new and fascinating, and although AskReddit may be on the conservative side of things regarding the adoption of such elements, it is quite clear that they cannot be completely escaped from on the Internet, no matter with which sites one engages. It might be a very rewarding avenue of investigation to focus on this non-textual or hypertextual meta level.

While the analysis of the dataset has proven interesting enough, a more thorough investigation on a larger scale might even include interviews with prolific writers engaging in online storytelling. Motivations for how and why certain language items or features are used could shed more light on the background for particularities of the genre and on the motivations for its relative conservativeness.

Staying closer to the linguistic core of this thesis, many mentions have been made in chapter 2 regarding the Internet as its own linguistic realm, situated apart from those of spoken or written (or, as it were, formal or informal) language. While Internet lingo has not

been the focus of this investigation, this might, also, provide an interesting angle, especially considering how quickly it changes. Many observations made by researchers in the late 20th century, although often generally still applicable, seem almost quaint for today's Internet users: smilies have evolved into emojis, the use of "textspeak" has decreased and many fads have appeared and disappeared in what on an academic timescale is less than the blink of an eye. However, just as these analyses have brought not only linguistic, but also cultural and historic insights, so can also research written in the present fulfil both these purposes. The potential for investigation is vast – from lexis (such as the aforementioned vocabulary item "meme") to grammar (for example in so-called "greentexts", a very particular form of short story telling used predominantly on 4chan), many things differ from everyday offline English, and there may be just as many different genres online as there are offline. This even applies to the genre of online storytelling alone – many of the pronunciations that have been made for the genre in the thesis might do well with the qualifier "on Reddit". Online storytelling on other websites, such as Tumblr or 4chan, might have completely different features again.

Even the qualifier "on Reddit" may be overstating the generalizability of these findings. The platform hosts many other subreddits – fora – where people may tell stories with a more specific focus, the most popular of which may be, for example, "Tales from Tech Support"[17], where users who work in technical support share their stories of experiences with (often technically inept) customers. In general, these subreddits share the same qualities – conservative in regards to the use of CMC-exclusive language, usage of *TL;DR* to summarize longer stories – but there will surely be subtle differences which could be an avenue of analysis for further research.

Apart from the aforementioned, more general possible avenues for further research, there have of course also been some more concrete ones that have been unveiled by the results of the analytical part of this very thesis. These include, amongst others, the extreme terseness of the writers' language and the predominance of female-gendered epithets.

[17] Accessible at www.reddit.com/r/talesfromtechsupport.

6 References

Adams Sheets, Connor. 2012. "What Does TL;DR Mean? AMA? TIL? Glossary Of Reddit Terms And Abbreviations". *International Business Times*. http://www.ibtimes.com/what-does-tldr-mean-ama-til-glossary-reddit-terms-abbreviations-431704?i10c (7 Oct 2017).

Aijmer, Karin. 2016. "Pragmatic markers as constructions. The case of anyway". In Kaltenböck, Gunther; Keizer, Evelien; Lohmann, Arne (eds.). *Outside the Clause: Form and Function of Extra-Clausal Constituents.* Amsterdam: John Benjamins Publishing Company, 29-58.

Akinnaso, F.Niyi. 1982. "On the Differences Between Spoken and Written Language". *Language and Speech* 25(2), 97–125.

Alexa. "Reddit.com Traffic, Demographics and Competitors - Alexa". 2017. Retrieved July 26, 2017, from http://www.alexa.com/siteinfo/reddit.com

Anthony, Laurence. 2014. *AntConc*. (Version 3.4.4w). [Computer Program]. http://www.laurenceanthony.net/ (August 12, 2017).

Anthony, Laurence. (n.d.). "AntBNC Lemma List". [Additional resource for a Computer Program]. http://www.laurenceanthony.net/resources/wordlists/antbnc_lemmas_ver_001.zip (August 12, 2017).

Ask Reddit.... (n.d.). https://www.reddit.com/r/AskReddit/ (July 26, 2017).

Battistella, Edwin L. 2005. *Bad Language: Are Some Words Better than Others?* Oxford: Oxford UP.

BBC News. "Reddit will not ban “distasteful” content, chief executive says". 2012, Oct 17. http://www.bbc.com/news/technology-19975375 (7 Oct 2017).

Beaman, Karen. 1984. "Coordination and Subordination Revisited: Syntactic Complexity in Spoken and Written Narrative Discourse". In Tannen, Deborah (ed.). *Coherence in Spoken and Written Discourse*. Norwood: ABLEX Publishing Corporation, 45-80.

Beißwenger, Michael, & Storrer, Angelika. 2009. "Corpora of computer-mediated communication". In Ludeling, Anke; Kytö, Merja (eds.). *Corpus linguistics: an international handbook.* Berlin: Walter de Gruyter, 309-327.

Bernstein, Michael S.; Monroy-Hernández, Andrés; Harry, Drew; André, Paul; Panovich, Katrina; Vargas, Gregory G. 2011. "4chan and/b: An Analysis of Anonymity and Ephemerality in a Large Online Community.". In Nicolov, Nicolas; Shanahan, James G. (chairs). *Proceedings of the Fifth International Conference on Weblogs and Social Media.* Menlo Park, CA: The AAAI Press, 50-57.

Biber, Douglas. 1988. *Variation across speech and writing*. Cambridge: Cambridge University Press.

Biber, Douglas; Conrad, Susan; Reppen, Randi; Byrd, Pat; Helt, Marie. 2002. "Speaking and Writing in the University: A Multidimensional Comparison". *TESOL Quarterly* 36(1), 9–48.

Biber, Douglas; Johansson, Stig; Leech, Geoffrey; Conrad, Susan; Finegan, Edward; Quirk, Randolph. 1999. *Longman grammar of spoken and written English* (Vol. 2). MIT Press Cambridge, MA.

Clancy, Brian. 2010. "Building a corpus to represent a variety of a language". In O'Keeffe, Anne; McCarthy, Michael (eds.). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, 80-92.

Collins, Peter, & Hollo, Carmella. 2010. *English Grammar: an introduction* (2nd edition). Houndmills: Palgrave Macmillan.

Crompton, Peter. 1997. "Hedging in academic writing: Some theoretical problems". *English for Specific Purposes* 16(4), 271–287.

Crystal, David. 2006. *Language and the Internet* (2nd ed.). Cambridge: Cambridge UP.

Crystal, David. 2011. *Internet linguistics: A student guide*. Abingdon: Routledge.

Danescu-Niculescu-Mizil, Cristian; West, Robert; Jurafsky, Dan; Leskovec, Jure; Potts, Christopher. 2013. "No country for old members: User lifecycle and linguistic change in online communities". In Schwabe, Daniel: *Proceedings of the 22nd international conference on World Wide Web*. Geneva: International World Wide Web Conferences Steering Committee, 307-318

Deris, Farhana Diana, Koon, Rachel Tan Hooi, & Salam, Abdul Rahim. 2015. "Virtual Communities in an Online English Language Learning Forum". *International Education Studies* 8(13), 79-87.

Dik, Simon C. 1997. *The theory of functional grammar: The structure of the clause* (2nd ed.). Berlin: Mouton de Gruyter.

Duggan, Maeve. 2014. "Online Harassment", 22 October. http://www.pewinternet.org/2014/10/22/online-harassment/ (6 Sept 2017)

eurostat. 2017. "Internet access and use statistics - households and individuals - Statistics Explained", January. http://ec.europa.eu/eurostat/statistics-explained/index.php/Internet_access_and_use_statistics_-_households_and_individuals (25 September, 2017)

Evison, Jane. 2010. "What are the basics of analysing a corpus?". In O'Keeffe, Anne; McCarthy, Michael (eds.). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge, 122-135.

Fludernik, Monika. 1991. "The historical present tense yet again: Tense switching and narrative dynamics in oral and quasi-oral storytelling". *Text* 11(3), 365–397.

Greenbaum, Sidney; Nelson, Gerald. 1995. "Clause relationships in spoken and written English". *Functions of Language* 2(1), 1–21.

Halliday, M. A. K; Matthiessen, Christian M. I. M. 2004. *An introduction to functional grammar* (Third Edition). New York: Oxford UP.

Herring, Susan. 1996. "Introduction". In S. Herring (Ed.), *Computer-mediated communication: linguistic, social and cross-cultural perspectives*. Amsterdam/Philadelphia: John Benjamins Publishing Company, 1-11.

Herring, Susan. 1999. "Interactional coherence in CMC". *Journal of Computer-Mediated Communication*, 4(4). Online journal, http://onlinelibrary.wiley.com/doi/10.1111/j.1083-6101.1999.tb00106.x/full (1 Nov 2017).

Hinduja, Sameer; Patchin, Justin W. 2008. "Personal information of adolescents on the Internet: A quantitative content analysis of MySpace". *Journal of Adolescence* 31, 125–146.

Holmes, Janet. 1988. "Doubt and Certainty in ESL Textbooks". *Applied Linguistics* 9(1), 21–44.

Hunston, Susan. 2009. "Collection strategies and design decisions". In O'Keeffe, Anne; McCarthy, Michael (eds.). *Corpus linguistics: an international handbook*. Berlin: Walter de Gruyter, 154-167.

Hyland, Ken. 1994. "Hedging in Academic Writing and EAP Textbooks". *English for Specific Purposes* 13(3), 239-256.

Hyland, Ken. 1996. "Talking to the Academy: Forms of Hedging in Science Research Articles". *Written Communication* 13(2), 251–281.

Johnson, Jeffrey K. 2008. "The Visualization of the Twisted Tongue: Portrayals of Stuttering in Film, Television, and Comic Books". *The Journal of Popular Culture* 41(2), 245–261.

"JSON". (n.d.). http://json.org/index.html (26 July 2017)

Kaltenböck, Gunther, Heine, Bernd, & Kuteva, Tania. 2011. "On thetical grammar". *Studies in Language* 35(4), 852–897.

Kämper, Vera. 2013. "Die Kanzlerin entdeckt #Neuland". 2013. *Spiegel Online*. June 13. Retrieved from http://www.spiegel.de/netzwelt/netzpolitik/kanzlerin-merkel-nennt-bei-obama-besuch-das-internet-neuland-a-906673.html

Kinsella, Naomi. 2010. "Btw its just netspeak lol". *Griffith Working Papers in Pragmatics and Intercultural Communication* 3(2), 64–74.

Köhler, Reinhard. 2003. "Zur Type-Token-Ration syntaktischer Einheiten: Eine quantitativ-korpuslinguistische Studie". In Cyrus, Lea; Feddes, Hendrik; Schumacher, Frank (eds.). *Sprache zwischen Theorie und Technologie: Festschrift für Wolf Paprotté zum 60. Geburtstag*. Wiesbaden: Deutscher Universitäts-Verlag, 93-101.

Kushin, Matthew J.; Kitchener, Kelin. 2009. "Getting political on social network sites: Exploring online political discourse on Facebook". *First Monday* 14(11). Online journal, http://journals.uic.edu/ojs/index.php/fm/article/viewArticle/2645/2350 (1 Nov 2017).

Lakoff, George. 1975. "Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts". In Hockney, D.J.; Harper, William L.; Freed, B. (eds.). *Contemporary Research in Philosophical Logic and Linguistic Semantics*. Dordrecht: Springer Netherlands. 221-271.

Leech, Geoffrey. 2000. "Grammars of Spoken English: New Outcomes of Corpus-Oriented Research". *Language Learning* 50(4), 675–724.

Leech, Geoffrey; Rayson, Paul; Wilson, Andrew. 2001. "Companion Website for: Word Frequencies in Written and Spoken English: based on the British National Corpus. List 5.2: Frequency list of verbs (by lemma)". London: Longman. http://ucrel.lancs.ac.uk/bncfreq/lists/5_2_all_rank_verb.txt (7 September 2017)

Leondis, Tony. 2017. *The Emoji Movie*. Columbia Pictures.

Ljung, Magnus. 2009. "The functions of expletive interjections in spoken English". In Renouf, Antoinette; Kehoe, Andrew (eds.). *Language and Computers Studies in Practical Linguistics*. Leiden: Brill, 155-171.

Louwerse, Max M.; McCarthy, Philip M.; McNamara, Danielle S.; Graesser, Arthur C. 2004. "Variation in language and cohesion across written and spoken registers". *Proceedings of the Cognitive Science Society* (26). 843-848.

McCarthy, Liam. 2016. "Literally Speaking". *Lingua Frankly*, 3. Online journal, https://ejournals.bc.edu/ojs/index.php/lingua/article/download/9278/8525 (1 Nov 2017).

McCarthy, Michael. 1993. "Spoken discourse markers in written text". In Sinclair, John M.; Hoey, Michael; Fox, Gwyneth (eds.). *Techniques of Description: Spoken and written discourse* (pp. 170–182). London/New York: Routledge, 170-182.

McCarthy, Michael; O'Keeffe, Anne. 2010. "What are corpora and how have they evolved?". In O'Keeffe, Anne; McCarthy, Michael (eds.). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge (3-13).

McEnery, Anthony; Xiao, Zhonghua. 2004. "Swearing in modern British English: the case of "fuck" in the BNC". *Language and Literature* 13(3), 235–268.

McEnery, Tony. 2006. *Swearing in English: Bad language, purity and power from 1586 to the present*. London/New York: Routledge.

McEnery, Tony; Wilson, Andrew. 2001. *Corpus Linguistics: an introduction* (2nd ed.). Edinburgh: Edinburgh University Press.

Miller, Jim; Weinert, Regina. 1998. *Spontaneous Spoken Language: Syntax and Discourse*. Oxford: Clarendon Press.

Montero-Fleta, Begoña; Montesinos-López, Anna; Pérez-Sabater, Carmen; Turney, Ed. 2009. "Computer mediated communication and informalization of discourse: The influence of culture and subject matter". *Journal of Pragmatics* 41(4), 770–779.

Morrow, Phillip R. 2006. "Telling about problems and giving advice in an Internet discussion forum: some discourse features". *Discourse Studies* 8(4), 531–548.

Murray, Denise E. 1988. "The Context of Oral and Written Language: A Framework for Mode and Medium Switching". *Language in Society* 17(3), 351–373.

Murray, Denise E. 2000. "Protean Communication: The Language of Computer-Mediated Communication". *TESOL Quarterly* 34(3), 397.

O'Donnell, Roy C. 1974. "Syntactic Differences between Speech and Writing". *American Speech* 49(1/2), 102–110.

Omernick, Eli; Sood, Sara Owsley. 2013. "The impact of anonymity in online communities". In Bilof, Randall (ed.). *International Conference on Social*

Computing, SocialCom 2013, SocialCom/PASSAT/BigData/EconCom/BioMedCom 2013. Washington DC: IEEE (526-533).

Pérez-Sabater, Carmen. 2013. "The linguistics of social networking: A study of writing conventions on facebook". *Linguistik Online* 56(6). Online journal, https://bop.unibe.ch/linguistik-online/article/view/257/347 (1 Nov 2017).

Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey; Svartvik, Jan. 1985. *A comprehensive grammar of the English language*. Harlow: Longman Group.

Rayson, Paul; Garside, Roger. 2000. "Comparing corpora using frequency profiling". In *Proceedings of the workshop on Comparing Corpora* (pp. 1–6). Stroudsburg, PA: Association for Computational Linguistics, 1-6.

"Reddiquette". 2017. https://www.reddit.com/wiki/reddiquette (6 September 2017).

"Reddit Content Policy". 2017. https://www.reddit.com/help/contentpolicy (11 September 2017).

Richards, Jack C.; Rodgers, Theodore S. 2001. *Approaches and Methods in Language Teaching* (2nd edition). Cambridge: Cambridge UP.

Santana, Arthur D. 2014. "Virtuous or vitriolic: The effect of anonymity on civility in online newspaper reader comment boards". *Journalism Practice* 8(1), 18–33.

Schnoebelen, Tyler. 2012. "Do You Smile with Your Nose? Stylistic Variation in Twitter Emoticons". *University of Pennsylvania Working Papers in Linguistics* 18(2), 117–125.

Stapa, Siti Hamin; Shaari, Azianura Hani. 2012. "Understanding Online Communicative Language Features In Social Networking Environment". *GEMA Online Journal of Language Studies* 12(3), 817-830.

Stenström, Anna-Brita; Andersen, Gisle; Hasund, Ingrid Kristine. 2002. *Trends in Teenage Talk: Corpus compilation, analysis and findings*. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Tannen, Deborah. 1982. "Oral and Literate Strategies in Spoken and Written Narratives". *Language* 58(1), 1–21.

Tran, Trang; Ostendorf, Mari. 2016. "Characterizing the language of online communities and its relation to community reception". In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Proceedings*. Stroudsburg, PA: Association for Computational Linguistics, 1030-1035.

Tribble, Christopher. 2010. "What are concordances and how are they used?". In O'Keeffe, Anne; McCarthy, Michael (eds.). *The Routledge Handbook of Corpus Linguistics*. Abingdon: Routledge (167-138).

Tu, Chih-Hsiung. 2002. "The impacts of text-based CMC on online social presence". *Journal of Interactive Online Learning* 1(2), 1–24.

Ulaby, Neda. 2006, Summer. "OMG: IM Slang Is Invading Everyday English". http://www.npr.org/templates/story/story.php?storyId=5221618 (5 August, 2017).

Varnhagen, Connie K.; McFall, G. Peggy; Pugh, Nicole; Routledge, Lisa; Sumida-MacDonald, Heather; Kwong, Trudy E. 2010. "lol: new language and spelling in instant messaging". *Read Writ* 23, 719–733.

Wakefield, Jane. 2016,. "Microsoft chatbot is taught to swear on Twitter". *BBC News*. March 24. http://www.bbc.com/news/technology-35890188 (1 Nov 2017).

Wolfson, Nessa. 1979. "The Conversational Historic Present Alternation". *Language* 55(1), 168–182.

Yates, JoAnne, Orlikowski, Wanda J. 1993. "Knee-jerk anti-LOOPism and other e-mail phenomena: Oral, written, and electronic patterns in computer-mediated communication". Retrieved from https://dspace.mit.edu/bitstream/handle/1721.1/2474/SWP-3578-45849142.pdf?sequence=1 (1 Nov 2017).

Appendix A: Scripts

The following script – the "scraper" downloads the front page of AskReddit at any given time in a machine-readable format, extracts the filenames of the 25 top threads, creates a folder for the given day and copies the current AskReddit front page as well as the necessary further scripts into that folder.

```
#!
wget www.reddit.com/r/askreddit.json
jq .data.children[].data.url askreddit.json > urldump
sed -ie 's/\/[a-zA-Z0-9_]*\/\"/.json/g' urldump
sed -ie 's/\"//g' urldump
wget -i urldump
mkdir `date +%m-%d`
mv *.json `date +%m-%d`/
cp *.sh `date +%m-%d`/
cd `date +%m-%d`
./parser.sh
./cleaner.sh
cd ..
```

Script 1: scraper.sh

The next script, the "parser", extracts all usable text data from the child threads, leaving metadata (such as post creation and author name) aside.

```
#!/bin/bash
mkdir parsed
FILES=*.json
for f in $FILES
do
      jq .[1].data.children[].data.body $f > $f.parsed
done
```

Script 2: parser.sh

The third script, the "cleaner", finally replaces single letters and signs that might interfere with the analysis with other, equivalent letters and signs. In particular, it deletes paragraph newline symbols, which would have significantly interfered with human readability of the corpus and which was possible since the paragraph structure of the texts was not under scrutiny, and replaces double quotation marks (") with single ticks (`), which was necessary since the individual story posts are also delineated with double quotation marks, introducing too much ambiguity.

```
#!/bin/bash
FILES=*.parsed
for f in $FILES
do
      sed -i 's/\\n/ /g' $f
      sed -i 's/\\\"/`/g' $f
done
```

Script 3: cleaner.sh

While these three scripts were sufficient to create a workable corpus, the decision was made to limit analysis to threads whose average posts were of a certain minimum length, in order to separate storytelling threads from those where the thread creator intended for shorter responses (see also chapter 3.4). The demarcation point was set as 30 words per post, which proved to leave out short answers while still giving enough material to work with. The so-called "shibboleth" script copies all files that fulfil the criterion into a subfolder and returns the number of selected threads to the user.

```
#!/bin/bash
FILES=*.parsed
mkdir useful
r=0
for f in $FILES
do
      h=$(grep ' ' -o $f | wc -l)
      t=$(grep '"' -o $f | wc -l)
      if (($((h/t)) > 30))
      then
            cp $f useful/$f
            ((r++))
      fi
done
echo $r
```

Script 4: shibboleth.sh

Appendix B: List of Most Frequent Verbs

Taken from Leech, Rayson & Wilson (2001) and edited. Words under investigation are bolded.

be 42277
have 13655
do 5594
will 3357 + would
say 3344
can 2672 + could
get 2210
make 2165
go 2078
see 1920
know 1882
take 1797
could 1683
think 1520
come 1512
give 1284
look 1151
(may omitted because there is no direct past – might has problems)
should 1112 (includes shall)
use 1071
find 990
want 945
tell 775
(must omitted because there is no past form at all)
(put omitted because forms indistinguishable)
mean 677
become 675
leave 647
work 646
need 627
feel 624
seem 624
(might omitted, see above)
ask 610
show 598
[…]

Appendix C: List of Taboo Words Investigated

This list is based on a list presented in Battistella 2005: 72.

Non-epithets:

Hell, damn, goddamn, shit, fuck, ass, bastard, crap, Jesus/Christ, piss

Epithets:

- **Racial**

  Wop, raghead, nigger, wetback,

- **Sexual**

  (son of a) bitch, fag, whore, motherfucker, cunt, pussy, screw, sucker, slut, skank

- **Other**

  Midget, gimp, retard

Appendix D: List of Pronouns Investigated

Personal:

I, you, he, she, it, we, they, me, him, her, us, them

Possessive:

My, your, his, her, its, our, your, their; mine, yours, hers, ours, theirs

Relative:

That, who, whom, whose, which, whoever, whomever, whichever

Demonstrative:

This, these, that, those

Indefinite:

Anybody, anything, anyone, everybody, everything, everyone, nobody, nothing, no one, somebody, something, someone, each, either, neither; all, any, most, some, none; both, few, many, several; another, such

Reflexive:

Myself, yourself, himself, herself, itself, ourselves, yourselves, themselves

Interrogative:

What, who, which, whom, whose, whatever, which

Appendix E: Abstract (English)

Like the medium itself, language used on the Internet is constantly evolving and changing. This thesis aims to investigate one particular subgenre of language – that of online storytelling – and especially to investigate its position on the spoken-written continuum, based on the idea that due to the ease of publication for users, the comparatively relaxed situation and oral storytelling traditions, the genre might be posited in between traditional ideas of spoken and written language use. To do this, methods of corpus analysis were used on a corpus of about 700,000 words taken from the online storytelling platform AskReddit. After sections considering existing research and literature on both the differentiation between the two types of language and on corpus analysis as the methodology of choice, seven features that are commonly thought to distinguish spoken and written language use were chosen and analyzed: pronouns use, hedging, lexical density, use of extra-clausal constituents, tense use, taboo language use and the coordination and subordinations. Concerns that delineating a polar system of two opposites (speech and writing) constricts analysis of a topic of such complexity as language were also addressed. Furthermore, it was also investigated whether the genre shows any further particularities that sets it apart from other online language usage. Results show the genre to be rather informal, more so than other, more conventional written language genres; however, while indeed often being positioned between spoken and written language, it is still closer to written language. Interestingly, commonly used Internet language features such as emoticons, emoji and abbreviated "text speak" is also rather rare. The conclusion is drawn that the hypothesis of the genre's increased "spokenness" only holds true to a small extent and that it is rather conservative in comparison to other genres of language used in computer-mediated communication.

Appendix F: Abstract (German)

Als Teil des sich konstant verändernden und weiterentwickelnden Mediums Internet wird in dieser Arbeit ein dort benutztes Subgenre der englischen Sprache analysiert: es handelt sich um das Genre des im Internet stattfindenden Geschichtenerzählens, oder „online storytellings“. Untersucht wird vornehmlich seine Position am Stilkontinuum zwischen gesprochener und geschriebener Sprache, da aufgrund von Faktoren wie der einfachen Zugänglichkeit, der wenig formellen Situation und mündlichen Erzähltraditionen eine Positionierung des Genres zwischen herkömmlichen Arten gesprochener und geschriebener Sprache angenommen werden kann. Dafür wurden Methoden der Korpuslinguistik an einem 700.000 Wörter umfassenden Korpus angewandt, dessen Inhalte von der „online storytelling“-Plattform AskReddit gesammelt wurden. Nach einer Analyse der bestehenden Literatur und Recherche sowohl zum Thema als auch der Methodologie der Arbeit wurden sieben Teilaspekte der Sprache ausgewählt und untersucht, deren Benutzung in gesprochener und geschriebener Sprache unterschiedlich sind: die Nutzung von Pronomen, Hedging, lexikalische Dichte, Nutzung von Extra-Clausal Constituents, Nutzung verschiedener Zeitformen, Nutzung von Tabusprache sowie Ko- und Subordination auf der Satzebene. Es werden in der Arbeit auch Einwände angesprochen, dass die Einstufung von Sprache auf ein bipolares System von mündlicher und schriftlicher Nutzung zu reduktiv ist. Ein weiterer Aspekt des Genres, der analysiert wird, ist seine Unterscheidung von anderen Instanzen des Onlinesprachgebrauchs. Die Resultate der Untersuchung zeigen, dass das Genre informeller als konventioneller Genres geschriebener Sprache ist und es sich durchaus zwischen gesprochener und geschriebener Sprache positioniert – dennoch aber deutlich näher am schriftlichen Sprachgebrauch bleibt. Auch häufig benutzte Aspekte der Internetsprache wie Emoticons, Emoji und abgekürztes „Text Speak“ scheinen nur selten auf. Daraus wird der Schluss gezogen, dass gesprochene Aspekte der Sprache im Genre nur eingeschränkt erscheinen, sowie dass es im Vergleich zu anderen Genres des Internetsprachgebrauchs als konservativ bezeichnet werden kann.