“Mission impawssible?”

An analysis of community-specific orthographic variation and lexical creativity online

Masterarbeit

zur Erlangung des akademischen Grades

Master of Arts (M.A.)

der Philologischen und der Philosophischen Fakultät

der Albert-Ludwigs-Universität Freiburg im Breisgau

vorgelegt von

Hanna Mahler

aus Nürnberg

Sommersemester 2020

Englische Sprachwissenschaft

Erstgutachter: Prof. Dr. Dr. h.c. Christian Mair

Table of contents

List of figures

List of tables

1. Introduction

Over the last decades, the internet and its impact have expanded in every possible dimension: more people from all around the globe have access to the world wide web and contribute to it, the number of different online platforms has grown rapidly, the predominance of English is slowly giving way to a multitude of languages being used online, and people are spending more and more time in the online environment (Internet World Stats 2020). With this steady expansion, language practices on the internet are also multiplying and diversifying, producing a variety of interesting linguistic phenomena. Among this fascinating variety one could count the following disparate expressions referring to domestic dogs found in the titles of posts on the popular online platform Reddit, which consists of many topical sub-groups referred to as "subreddits" (1-6, emphasis added):

(1) p292: Cute shelter *dog*

(2) p014: Philosophy *doggo*

(3) p132: Someone's good *boy*

(4) p066: Fast *bois*

(5) p162: Meet our *puppy* Mochi. He's 12 weeks old today!

(6) p071: This *pupper* getting food on the table

These non-standard variants are part of a style used on pet-themed subreddits, referred to as "pupper talk" or "DoggoLingo", which I will classify and describe as an online slang. Treating the slang features as instances of stylistic variation, one can ask about potential factors influencing users' stylistic choices during the composition of these utterances. My study will therefore address the following research question:

> Which factors influence the presence of community-specific features in the comments on pet-themed communities on Reddit?

The study is therefore concerned with stylistic, intra-speaker variation (Bell 1984:145, Biber & Conrad 2009:16) within the context of community formation in the online environment. As a theoretical framework, Bell's audience design (1984) and his proposed audience roles will be adapted to the online environment. To answer the research question, qualitative data collected during an ethnographic pilot study will be combined with a corpus of posts and comments from eight pet-centred subreddits. A logistic regression model will be used to predict the presence of slang features based on several audience and non-audience factors.

My study is a first step towards filling several research gaps concerning the online environment. First of all, research on computer-mediated communication (henceforth "CMC") has often

ignored the highly multimodal environment on many platforms and its potential impact on text production and perception. My study takes into account the multimodal prompts (consisting of a title and a picture or video) that the textual contributions react to (Androutsopoulos 2013:245). Furthermore, my study investigates stylistic variation in the context of online communities. The paper thereby contributes to filling the research gap identified by Androutsopoulos (2006:423): he states that compared to the creation of "virtual identities" on blogs and personal websites, "[l]ess attention has been paid to the processes by which people establish member identities in the frame of an online community" (Androutsopoulos 2006:423).

This paper is structured as follows. Within the theoretical introduction I will introduce the main areas relevant to the paper: the linguistic investigation of computer-mediated communication, the platform Reddit, linguistic conceptualisations of communities online, the slang pupper talk, research on audience design, and finally research on multimodality in the online environment. The following section outlines the methodology of the study and the data collected. After presenting descriptive and inferential statistics in the results section, the discussion contains a description of the slang under investigation, the factors influencing its usage, and the community status of the subreddits in question. Furthermore, I will present some methodological challenges for the quantitative study of stylistic variation online, before moving on to the conclusion.

2. Theoretical background

This section provides an overview of previous research relevant for this thesis. First, computer-mediated communication in general will be addressed, before focusing more specifically on the Reddit environment. Next, research on online communities will be summarised. Afterwards, internet slang will be addressed, which includes non-standard orthography and lexical creativity. This will be followed by a section on the important theory of audience design as well as multimodality.

2.1 Computer-mediated communication

Before moving on to Reddit as an example of an online platform, the following section provides a very brief introduction into CMC and previous research conducted on the topic. Linguistic studies on CMC evolved alongside the technical innovations of the medium over time. While communication on the internet was initially text-based, researchers are now more and more considering the combination of multiple modes that have recently become available, such as the incorporation of audio and video transmissions (Herring & Androutsopoulos 2015:127,

Herring 2015, Thurlow et al. 2020). The complete consideration of these various modes and their interplay during language perception and production is, in my view, one of the central challenges facing linguists working on CMC at this point in time.

Androutsopoulos (2006:420-421) provides a useful distinction between a first and second 'wave' of CMC research. According to him, during the first wave researchers focused on 'the' language of the internet, emphasising how it differed from offline communication and working with comparatively small samples. In contrast to that, the second wave of CMC research takes into account the vast variability of language online and shifts the focus "from medium-related to user-related patterns of language use" (Androutsopoulos 2006:421). The second wave is therefore concerned more with a sociolinguistic perspective on the online environment, addressing questions around "the interplay of technological, social, and contextual factors in the shaping of computer-mediated language practices, and the role of linguistic variability in the formation of social interaction and social identities on the Internet" (Androutsopoulos 2006:421). The focus hereby is on how linguistic resources are employed locally by internet users in the context of community and identity creation (see also Seargeant & Tagg 2014:5). Situated within this second wave, the present paper also does not investigate language use on Reddit as a whole but zooms in on specific communities to see how users strategically employ community-specific features and what influences their usage. The paper is furthermore in line with the recent advancement of quantitative methods being fruitfully applied to CMC data (e.g. Cole et al. 2017, Grieve et al. 2017, Liimatta 2016, Paolillo 2001).

A further helpful distinction that is commonly used is the one between the more "static" web 1.0 and the "dynamic" web 2.0 (Bolander & Locher 2014:16). This study will be concerned with web 2.0, as it is defined by Herring & Androutsopoulos (2015:130, emphasis original):

> The term *Web 2.0* refers to Web-based platforms that incorporate user-generated content and social interaction, often alongside or in response to structures and/or (multimedia) content provided by the sites themselves; such platforms have been ascendant since the turn of the millennium. A common characteristic of Web 2.0 environments is the cooccurrence or convergence of different modes of communication on a single platform.

Reddit is a prime example of a web 2.0 platform, as it consists entirely out of (multimodal) user-generated content (see section 2.2).

Looking at the characteristics of CMC, researchers now agree that computer-mediated language does not fit neatly into the categories of "written" versus "spoken" language, as it combines features of both (Herring & Androutsopoulos 2015:128). With the increasing multiplication and diversification of platforms and channels, researchers have also realised that CMC cannot be

regarded as a homogenous entity. Instead, "as CMC on the internet became more and more diversified, there was a growing recognition that language was used differently in different kinds of CMC" (Herring & Androutsopoulos 2015:129). The linguistic diversity we observe online therefore mirrors the diversity we observe in spoken language, as speakers adapt to their interlocutors and the affordances of each speech situation. In order to classify and categorise instances of CMC, several schemas have been proposed; this paper (see chapter 2.2.1) will employ the "faceted classification scheme" by Herring (2007).

Summing up, one could say that linguistic research has managed to "keep up" and do justice to the ever-changing nature of the online environment and the language on it. However, there are a number of questions still waiting to be thoroughly addressed. They concern, for example, the multimodality mentioned above. Furthermore, in my opinion, it is not yet fully explained how language change and innovation online and offline relate to one another. Can linguistic innovations originating online permeate into spoken language and if yes, how? How do known linguistic processes take place differently in the online environment (e.g. s-shaped curves of language change)? Can language change originating in spoken language be accelerated by adapting it in computer-mediated conversations? Apart from these theoretical questions, some methodological issues also remain to be addressed. Linguists need to agree on standard principles of conduct for retrieving, processing, presenting and storing data from the internet, especially social media data (e.g. Fiesler et al. 2016).

2.2 Reddit

Since this study is based entirely on data from Reddit, this section provides a short introduction into Reddit and previous research conducted on it. As of May 2020, Reddit boasts more than 430 million active users and over 130,000 active subreddits; and is the fifth most visited website in the United States (Reddit Inc. 2020). It therefore warrants linguistic investigation as millions of people consume and produce language on the platform on a daily basis. One thing that should be clarified at the start is how to label the platform Reddit, as several terms are currently used to refer to it. Some studies call it a "news aggregate website" (Bergstrom 2011:2, Cole et al. 2017:2), whereas others use the term "large social networking site" (Golbeck & Buntain 2014:615), "social media website" (Liimatta 2016:5), or "social news website" (Moore & Chuang 2017:2313). Singer et al. (2014) is the only study that also presents users' perspectives on the nature of Reddit. In their survey, 88% of informants would describe the platform as a "Forum / Message board", whereas 71% subscribe to it being an "entertainment site". 56% say it is a "news site" and 54% would also apply the term "Image/Video or file sharing site". Less

popular options included “Portal”, “Educational site”, “Social Network” and “other” (Singer et al. 2014:520). For the remainder of this study, I would like to go along with the very apt description as a “meta-community” by Moore & Chuang (2017:2313, emphasis original):

> Reddit is more than a place to post news content; it has evolved into a massive, thriving, highly influential virtual community [5,6]. With its ever-growing cadre of Subreddits, specialized areas focusing on a wide variety of topics, Reddit is more like a meta-community. [...] Reddit has been referred to as both a culture and many cultures because of the complex interactions across Subreddits.

This definition does not impose that all subreddits must be similar in nature but admits that Reddit is more of a ‘container’ for a variety of communities with a variety of purposes.

At this point it seems adequate to present an illustration of a post from one of the subreddits investigated, so that readers unfamiliar with Reddit can get an impression of the platform. Figure 1 presents a post (consisting of a title and two pictures) taken from the subreddit r/dogswithjobs, a community centred around dogs with a profession, such as guide dogs, search and rescue dogs, or sheepdogs. Above the picture one can see the username (in this case it was deleted), the title, the time of creation, the small tag referred to as “flair” (in this case: “Police Dog”), as well as the “karma” of the post (here: “12.7k”), which is a measurement of its overall success and appreciation by the community.

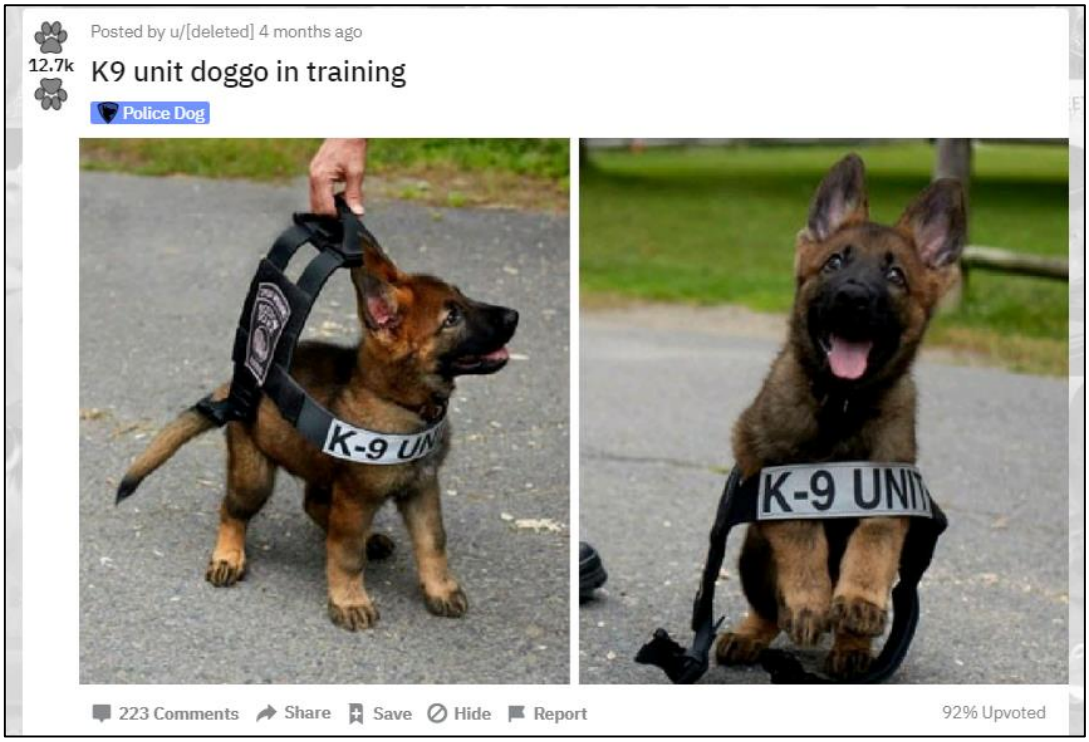

Figure 1: Illustration of a post from r/dogswithjobs (p033).

Below the picture Reddit displays the number of comments (in this case 223) and, in the right corner, the percentage of "upvotes" (so the percentage of users who liked the post). Moving further down the page one can find the comments, a short section of which is depicted in Figure 2.

Figure 2: Illustration of comments below a post (p033).

Here one can see the unique comment structure of Reddit: users can respond directly to whichever comment they choose, which are depicted on different indentation levels. Each comment contains information on the author (redacted), the number of points it received (similar to post karma), the time of creation, as well as the text itself.

2.2.1 A faceted classification of Reddit

The following paragraphs will now describe Reddit as a communication environment according to the "faceted classification scheme" for computer-mediated discourse proposed by Herring (2007). The approach is divided into medium and situational factors and provides a useful starting point for the description of online platforms. However, Herring herself (2007:27) admits that the scheme does not pay enough attention to different modes of communication.

Focusing on medium factors first (as seen in Table 1), communication on Reddit proceeds asynchronously and content (posts, comments and private messages) is transmitted upon completion. Posts and comments are stored without a temporal limit, and both may contain a maximum of 40,000 characters. Reddit supports textual as well as visual communication within the posts in the form of pictures, videos or GIFs (short for "graphic interchange format"), whereas no images can be incorporated in the comment section. Communication is generally highly anonymous, since users can choose any nickname when registering and are not required

to provide any contact details or personal information. In addition to public posts and comments, users are able to send each other private messages and participate in topical discussion groups. Users can also filter out unwanted communities from their home feed, as well as suppress unwanted private messages in several ways. Furthermore, users can easily incorporate quotes from previous comments into their text. The format in which messages are displayed is rather different compared to other platforms. Users can choose the order in which content should be presented to them ("newest", "most popular", "trending", or "controversial") and also specify which content they prefer to receive (topical and geographical).

Medium factors	
Synchronicity	asynchronous
Message transmission	message-by-message
Persistence of transcript	very persistent
Size of message buffer	40,000 characters
Channels of communication	textual and visual
Anonymous messaging	high level of anonymity
Privat messaging	possible
Filtering	possible
Quoting	possible
Message format	variable

Table 1: Medium factors of Reddit (adapted from Herring 2007).

Moving on to situation factors (displayed in Table 2), the participation structure on Reddit is public and can be described as many-to-many communication (with the exception of the private messaging function). However, the percentage of Reddit users actually composing posts and comments is often said to be relatively small compared to the number of people consuming the content, which are referred to as "lurkers" (Golbeck & Buntain 2014:616, Singer et al. 2014:520). The characteristics of the participants are likely to vary considerably between subreddits. For example, the average subscriber to r/teenagers can be expected to be quite distinct from the average member of r/StockMarket. In general, however, Reddit users are often described as young, urban, and male (Duggan & Smith, 2013). Purpose, topic, tone, and activity are also difficult to describe for the platform as a whole due to its vast internal diversity. While communities such as r/relationshipadvice are centred around asking for and providing advice, other communities, such as r/pics, have their main purpose in sharing photographs. I would therefore like to focus only on the subreddits chosen for the analysis at hand in order to allow for a more detailed description.

In all the communities investigated, the main activity appears to be sharing positive images and videos of the users' pets, especially dogs, and discussing this visual input. Commenting activity can include asking questions about the dog and its breed, praising the author, discussing dog-related questions, or engaging in a joke based on the video or picture. The purpose of doing so can range from increasing one's personal karma score, to creating community bonds with other members, or simply entertainment. While doing so, the tone is mostly positive, cooperative, and highly informal, but trolling and other negative behaviour does occur sporadically. Concerning the norms, Reddit has a general behavioural guideline it asks users to comply with (Reddit Inc. 2020). It contains recommendations such as "Please do: Use proper grammar and spelling" or "Please don't: Post someone's personal information". In addition to that, every community has its specific community rules, which are visible on the sidebar of each subreddit. For example, the rules on r/aww include "No 'sad' content", and "No asking for donations or adoptions". The code on Reddit is predominantly English, however there are some smaller subreddits using different languages (e.g. German on r/de, Spanish on r/argentina). Similarly, Roman script is the most widely used writing system, but other systems are also used infrequently (e.g. on r/arabs or r/china).

Situation factors	
Participation structure	public, many-to-many, small percentage of active contributors
Participant characteristics	depends on subreddit
Purpose	Entertainment, news, messaging (depends on subreddit)
Topic or theme	depends on subreddit
Tone	depends on subreddit
Activity	depends on subreddit
Norms	general "Rediquette" and specific community rules enforced by moderators
Code	predominantly English and Roman font

Table 2: Situation factors of Reddit (adapted from Herring 2007).

The previous table again emphasises how fitting the description of Reddit as a "meta-community" (Moore & Chuang 2017:2313) is, as many features cannot be described in general and the individual subreddits may differ considerably in character. Having described Reddit as a platform and its specific affordances and possibilities, we can now move on to previous research investigating language use on Reddit.

2.2.2 Previous linguistic research on Reddit

Considering the vast amount of data provided by the platform and linguistic variability found on it, research on Reddit is still scarce. The studies mentioned below might therefore amount to a nearly exhaustive list on previous linguistic research on Reddit. One of the earliest studies on the platform is Finlay (2014), which attempts the difficult task of analysing the influence of age and gender on communicative behaviour and success on Reddit. As the platform is anonymous, Finlay's information is based on self-disclosure of Reddit users in response to a post requesting demographic information – the accuracy of the information can therefore not be evaluated. Finlay (2014:26) finds that younger users on average have lower karma scores (which indicate community-internal success and appreciation) and write shorter, less complex comments. The gender data appeared to be too scarce to allow for meaningful conclusions. While his results seem plausible, I would like to draw attention to the fact that he did not take differences between subreddits into account: younger users might simply be more involved in subreddits which value short comments than older users. Coming from a similar direction, Flesch (2018) conducted a sociolinguistic analysis on a single linguistic variant: the shortening of <though> to <tho>, which she defines as part of general "internet slang" (Flesch 2018:37). Flesch finds that the non-standard spelling is preferred by younger users and users with Hispanic and Black ethnicity (Flesch 2018:39, her terminology). She concludes that <tho> is therefore more than a mere shortening but is used by the non-white minority of Reddit users to differentiate themselves. Her analysis provides a good example of users employing linguistic means to create their identity on the platform, but the study is subject to the same constraint as Finlay (2014), as it relies on the accuracy of demographic information given by the users.

Coming from a more computational side, Kershaw (2018) constructs a method to model language innovation and spread on Reddit and Twitter. He finds that the diffusion of language change hinges both on the interaction between users as well as between communities. Focusing also on Reddit's internal structure, Cole et al. (2017) investigate the link between subreddit size and spread of linguistic innovations. Through statistical modelling they find that larger and more general subreddits (which they refer to as "supersubreddits") are more open to innovation than smaller, topically specific subreddits. However, I would question the validity of their claim that "larger communities are based on discussing general topics and have weak social ties, whereas small communities are based on discussing specific topics and have strong social ties" (Cole et al. 2017:1). This question will be taken up again in the discussion below on whether subreddits can be classified as distinct virtual communities (see section 5.2). Under these quantitative studies on Reddit we can also count Liimatta (2016). He performs a

multidimensional register analysis on a sample of 37 self-selected subreddits and arrives at three relevant dimensions, which partially overlap with the dimensions previously identified by other scholars. They include the temporal orientation of the utterance (past vs. present), the topic of the conversation (personal vs. factual) and the level of personal attachment (involved vs. informational) (Liimatta 2016:68). This exploratory analysis therefore provides interesting insights into register variation on Reddit, but also has several drawbacks. Especially not distinguishing between titles and comments is, in my opinion, problematic, as the present study reveals important differences between these two text types, e.g. regarding text length (see section 4.1.2).

Other disciplines have also started paying attention to the Reddit environment; even if their topics are not directly relevant for the present study, they still provide valuable background information on the platform and user behaviour on it. Here I can only provide a few illustrations representing the whole range of research being conducted. Vasilev (2018) presents a computational approach to inferring Reddit users' gender from their comments. Coming more from a social perspective, Bergstrom (2011) discusses the relation between anonymity and conflicting expectations of authenticity on some subreddits. Golbeck & Buntain (2014) research user roles on Reddit and find that the "answer person" role is attested, which is probably unsurprising considering their choice of subreddits, including r/IAmA (in which popular or interesting people answer other users' questions). Moore & Chuang (2017) focus on motivational factors influencing usage of Reddit. Community building, status-seeking and entertainment are found to be significant predictors of use. Singer et al. (2014) describe the evolution and increasing diversification of Reddit from its inception in 2008 onwards. They find that "Reddit has transformed itself from a dedicated gateway to the Web to an increasingly self-referential community that focuses on and reinforces its own user-generated image- and textual content over external sources" (Singer et al. 2014:1). Their findings remind us that online platforms are constantly evolving and that research on the characteristics of a certain platform might be outdated within a few years.

Summing up, we saw that there is wide range on previous research on Reddit from a variety of perspectives. Previous linguistic investigations are mainly quantitative and aim to generalise over the platform or over subreddits as a whole, or try to answer general questions on language innovation. Little attention has so far been paid to how users modify their language within specific interactions to display a certain identity or to achieve certain communicative goals. The site therefore still holds a lot of potential for linguists interested in computer-mediated communication.

2.3 Online communities and linguistic communities online

Communication online, just like in the offline context, can take place between two individuals or within a group of people. Various online platforms offer a range of possibilities to form and join different sorts of groups and to pursue activities within them. Researchers have long been interested in how these settings differ from offline groups and how these differences impact the formation and presentation of an individual's identity, as well as his or her communicative behaviour. Within this section, I would therefore like to address two questions. First, how can communities online be characterised in a structural sense? Second, to what extent can a given online community be adequately described using sociolinguistic concepts? This discussion is of great importance, since many studies investigating online groups talk about the "communities" they study without discussing what constitutes a community in an online context. Herring (2004:6) points out that "for some writers, it seems that any online group automatically becomes a 'community'". As an example, we could cite Cole et al. (2017:2), who presuppose that subreddits represent distinct communities, without further discussion of the concept.

The first researcher to discuss the contrast of online and offline communities in depth is Rheingold (1993). In his early work, he defines virtual communities[1] as "social aggregations that emerge from the Internet when enough people carry on those public discussions long enough, with sufficient human feeling, to form webs of personal relationships" (Rheingold 1993:5). Androutsopoulos (2006:421) emphasises the importance of the concept "online/virtual community" for more user-related approaches to CMC, while at the same time admitting that there is no general definition for the term. What appears to be clear is that online communities have different characteristics than 'offline' communities and cannot necessarily be compared with them. More specifically, "besides their lack of physical proximity, Internet-based groups lack the stable membership, long-term commitment, and social accountability that would be needed to qualify as communities in the sociological sense" (Androutsopoulos 2006:422). One of the most extensive discussions on what constitutes virtual communities can be found in Herring (2004). In her paper, Herring (2004:14) collects 6 criteria for virtual communities from previous research:

[1] The terms "virtual community" and "online community" are used interchangeably in the literature and are also intended to refer to the same concept within this paper (see Androutsopoulos 2006:421).

1) active, self-sustaining participation; a core of regular participants
2) shared history, purpose, culture, norms and values
3) solidarity, support, reciprocity
4) criticism, conflict, means of conflict resolution
5) self-awareness of group as an entity distinct from other groups
6) emergence of roles, hierarchy, governance, rituals.

Herring (2004:15) also proposes ways to operationalise these criteria, which will be used in the discussion section to determine whether the subreddits investigated can be termed online communities in a structural sense.

Let us now look at how groups of speakers have traditionally been conceptualised in sociolinguistic studies, in order to determine whether these concepts can also fruitfully be applied to virtual communities. Androutsopoulos (2006:422-423) mentions that many established sociolinguistic concepts have previously been employed by various authors in order to describe online communities; however, their applicability depends on the nature of the respective online group being studied. He notes:

> It seems that the adequacy of these notions for particular online groups will depend both on their collective patterns of online interaction and on types of individual engagement. For example, while the administrators of a discussion board might satisfy the conditions for a community of practice […], an imagined community of like-minded individuals might be more suited to the viewpoint of occasional users of the same board. (Androutsopoulos 2006:422-423)

Let us first look at the concept of speech community. The main focus of this concept is the continuity of a certain social group, its linguistic practices, and its delineation from other, equally stable communities. While earlier definitions also included the internal homogeneity of the speech community, scholars later replaced this notion with internal structure and ordered variation. Thus, speakers within a speech community may not share the same linguistic repertoire, but they share norms of usage (Irvine 2006:691-693). The concept of speech community does not appear very appropriate for the study of online communities, since online groups are in most cases more fluid. To my knowledge, the concept has so far not been applied to the study of a specific online group.

A different perspective on what shapes linguistic practices is presented in the social network approach. Here, the focus lies on individuals' "webs of personal relationships" and linguistic variation is explained by differences in type and density of these networks (Irvine 2006:693-694). Being mainly concerned with the type and structure of social ties (Milroy 2004:552), social network analysis seems primarily useful for CMC sites that provide a technical environment allowing for the creation and maintenance of various social relationships. For

example, Lange (2007) used the social network approach to study public and private behaviour on the video-sharing site YouTube, and Hinrichs (2016) investigates language choices within initial and responding posts on Facebook.

A further construct commonly employed is that of communities of practice, which Irvine (2006:694) defines as "a grouping that is based on participation in some activity or project". In this approach, shared interpretations (concerning both linguistic and non-linguistic practices) arise from shared experience. Similarly, Eckert (2006:683) defines them as "a collection of people who engage on an ongoing basis in some common endeavor". Meyerhoff (2004:527-528) elaborates on three criteria that must be met in order for a group to be termed a community of practice: the members need to be 1. mutually engaged in a 2. jointly negotiated enterprise, using 3. a shared repertoire (see also Wenger 1998:2). The concept of community of practice appears to be a useful one for the description of online communities, as on many platforms users are primarily brought together by a shared interest or activity, without any form of offline contact or socio-demographic similarities between them. To provide two examples, Stommel (2008) successfully applies the concept to a German eating disorder forum to shed light on the reification of community rules, while Graham (2019) employs the concept to analyse the use of emojis in an online gaming community.

In a quick side note I would like to mention that Meyerhoff (2004:534) also comments on the compatibility of the community of practice approach with theories explaining stylistic variation. She deliberately delineates community of practice approaches from concepts such as audience design, saying that:

> linguistic style shifting is neither a function of the attention speakers pay to their speech […], nor of their attention to social characteristics of the addressee or audience […]. Instead, linguistic style is part and parcel of speaker's work to construct a social identity (or identities), which is meaningful to themselves and to others (Meyerhoff 2004:534)

Since audience design serves as the main theoretical foundation for the study at hand, a short comment on this statement is needed. In my view, audience design and creating one's identity as part of a community of practice are not mutually exclusive phenomena but can go hand in hand in online interactions. Through aligning one's utterance with the style previously employed by the addressee, users can intentionally position themselves as part of the same community of practice which they perceive their addressee to be a part of. Tagg & Seargeant (2014:180-181) also explain how the two concepts can be fruitfully combined:

> Audience design is founded on the insight that one constructs an idea of the audience […] for the purpose of giving context to one's utterances. As such, it is an important element in constructing or maintaining the community: it is an aspect of constructing the links between yourself and those in your network, and of building these links around shared cultural and linguistic practices. […] Audience design works by drawing on shared practices which are part of the dynamics which constitute community relations, and at the same time enacting and elaborating upon these practices. In this sense, online audiences are imaginings of the poster's understanding of a community's practices.

Summing up, it appears as if the concept of "community of practice" is the most useful in order to describe virtual communities. Whether the term also applies to individual subreddits, groups of subreddits, or Reddit as a whole, will be discussed later on in the discussion section.

2.4 Pupper talk, DoggoLingo, and internet slang

I would now like to provide a brief introduction into pupper talk and related phenomena, while saving a detailed description based on the data obtained for the discussion section. The variety will be labelled a slang in accordance with the definition provided in Malmkjær (2010:489):

> One useful way of characterising slang is as a style of language occupying, along with intimacies such as 'baby talk' and terms of endearment, the extreme 'informal' position on a continuum representing degrees of formality. Slang is coined, adopted and used, and evolves separately from or in deliberate contrast to what are thought to be the standard and prestige varieties of a language.

This definition seems especially fitting as the connection between pupper talk and endearment as well as child-addressed language was frequently pointed out by informants during the pilot study (see section 5.1.1). Other studies prefer the term "language play"; but while the variety is unquestionably highly playful and creative in nature, this terminology seems not appropriate for the data at hand. The main reason is that many items occur in otherwise standard utterances and might have lost their playful connotation due to their high frequency in the communities concerned. Compare utterances (7) and (8) below: while the first one is clearly playful in imitating a dog's wish to be fed a chicken, the second one employs the form *doggo* in an otherwise non-playful sentence.

> (7) c1253: gib longboye chimken,, / - totally not lomgboye
>
> (8) c1347: My sons's college brings therapy dogs in before finals week. And faculty and staff are allowed to bring their doggos on campus.

At this point I would like to take some time to talk about different internet slangs and how they relate to the features investigated in this paper. It needs to be said that internet slangs have only received little scientific attention so far. Let us first look at LOLspeak, which is a slang that is likely to have inspired and served as a source for the slang pupper talk investigated in this paper. Bury & Wojtaszek (2017) study this phenomenon and classify it as a playful, humorous "micro-

genre" that deliberately uses grammatically incorrect utterances, mostly in combinations with pictures of cats, known as "LOLcat". Their goal is to show that there are regularities to what inexperienced observers might perceive as pure chaos. The main features they investigate are provided in the table below (Table 3).

Feature	Example
"CAN HAS" formula	I can has prom date?
Question formation	AM dis what Squidward luuked like azza babbeh?
Auxiliary verb substitution	I iz pritty
Omission of auxiliary verb	How YOO doin??
Inconsistent verb forms	ai luvs re-reeding deze storees.
Manipulating categories	I has a happy!
Deviant spellings	<oo> for <u>, <b> for <v>
Onomatopoeia and rhyming	…blargh…

Table 3: Features of "LOLspeak" with examples taken from Bury & Wojtaszek (2017).

Many of the above features also appear in the Reddit communities investigated. What appears to be even closer to the variety at hand is what Bivens (2018) terms "Doggo-Speak". Bivens (2018) attempts a description of the features of the variety based on Facebook data (as seen in Table 4) and trains a machine learning tool to distinguish it from standard English based on these properties.

Feature	Example
Do-Rule	You are doing me a frighten
Usage of *Heck*	Aw heck German boii
Pronoun Mismatch Rules	him didn;t get treats when he wanted he awooo
Spelling Transformations	grow big and stronk Protecc bol
Capitalization Rules	oh GOSH what have you DONE chef shoob!>!??! sOME history for you WOuld u wanna see?

Table 4: Features of "DoggoLingo" according to Bivens (2018) with her examples.

Going more into detail, Punske & Butler (2019) focus entirely on the "do a X" construction, which they call "do-rule", and how it relates to the formation of standard verb phrases in English. Golbeck & Buntain (2018) investigate lexical propagation of two common features of "DoggoLingo" (*heck* and the "_/10 would"-construction) on Twitter and Reddit. They find that the usage of these items increased as the user account from which they stem grew in popularity. However, they admit that this does not allow for any conclusions about causation. Apart from these academic works, other journals and media platforms also take an interest in the

"DoggoLingo" variety. Out of the plethora of online material I would only like to mention the article by Jessica Boddy (2017), in which she explains several features associated with the variety. Finally, it is of interest that the word *doggo* has reached such a level of popularity that the Merriam Webster Dictionary even put it on its "Words we're watching" list (Merriam Webster). For the remainder of this paper I will use the term "pupper talk" to refer to the realisation of the slang specifically on Reddit, delineating it from related realisation on other platforms known as "DoggoLingo" or "Doggo Speak".

What needs to be mentioned is that none of the informants in the pilot study described pupper talk as part of a wider internet phenomenon. So even though there is an obvious overlap with other internet slangs, the users might not necessarily be aware of that. This might be related to the user demographics of Reddit (see section 2.2.1): as users are on average rather young, they might not have been on the internet long enough to observe other slangs, or to see how pupper talk was established. Even though this diachronic perspective and questions of source material and influence across platform are interesting, they are not the main focus of this paper. A more detailed description of the features of pupper talk as it is used on Reddit is provided below (section 5.1). The following subsections will now provide some more details on prominent characteristics of slang: first, non-standard orthography, and second, lexical creativity and humour.

2.4.1 Non-standard orthography

One way in which (written) slang can purposefully deviate from the standard variety of a language is through the use of different orthography. This is also one of the salient features of pupper talk. Androutsopoulos (2000:514) defines non-standard spellings as "spellings that diverge from standard (codified) orthography and/or do not occur in formal writing". In the same paper, he proposes a typology of non-standard spellings (developed originally for German, but applicable to other languages as well), including the six sub-types displayed in Table 5.

Sub-type	Description	Examples
Phonetic spellings	representations of standard pronunciation not covered by standard orthography	<wuz> for *was*
Colloquial spellings	representation of reduction phenomena typical of colloquial speech	<gonna> for *going to*
Regiolectal spellings	representation of features typical of a regional variety	German <ick> for *ich*
Prosodic spellings	representation of prosodic patterns, e.g. word stress	<HIII> for *hi*
Interlingual spellings	phonetic spellings of loanwords according to native orthographic rules	German <äktschn> for *action*
Homophone spellings	graphic alterations without a correspondence to phonic alterations	<u> for *you*

Table 5: Typology of non-standard spellings (from Androutsopoulos 2000:520-522).

The answer to the question where the orthographic variation associated with pupper talk fits into this classification scheme is not straightforward and depends heavily on how the origin of these variants is interpreted (see section 5.1.4). A major problem for the interpretation is that the existence of a spoken equivalent, which the spellings could refer to, is questionable. Other cases of orthographic variation online, for example as investigated by Honkanen (in press: 336), have a clear connection to the oral vernacular of the speakers. This leads to the more general question whether spoken language serves as an adequate reference point for the description of online slang, which might be shared by users of various linguistic backgrounds and of various proficiencies in the standard variety of the language (in this case: English). Another question that is of relevance is what counts as 'standard' within the online environment. For example, a study by Paolillo (2001:206) showed that the notion of standard is a relative one and needs be contextualised. For many online contexts it might be more appropriate to consider features common in the online environment as standard (without attempting to homogenize language use online). Androutsopoulos (2011:12) also explains how the standard of a language loses its normative character in the online environment, a process he relates to "destandardisation".

The importance of orthographic practices for the formation of group identity has already been investigated for the offline context (e.g. Sebba 2007, 2012, Androutsopoulos 2000), but research on the use of spelling variants in the process of identity creation online seems to be fairly scarce. Looking at one specific non-standard spelling only, <tho> for <though>, Flesch (2018:39) finds that "nonstandard spellings are some of the linguistic strategies they [Hispanics and Blacks] use to differentiate themselves from the overwhelmingly white Reddit user base." She therefore classifies <tho> as "a marker of affiliation with a social group as well as a sign

of familiarity with memes and internet subculture." In a similar way, the non-standard orthography investigated in the present paper can function as a marker of in-group identity.

2.4.2 Lexical creativity and humour

The internet is a fertile ground for playful interaction and conversational humour (North 2007:538, Herring 1999:9, Chovanec & Tsakona 2018:5). For example, Danet (2001) provides a wide-ranging overview of different forms of playfulness online, both visual and textual. In addition to the internet in general, the platform Reddit as a whole "encourages a playful approach to discourse" (Massanari 2015:1). Several researchers point out the importance of humour in the online environment for the creation of both individual and group identity (Shifman 2014:389, Chovanec & Tsakona 2018:6). In a pioneering study, Baym (1995) investigates the humour used on an online forum focused on television series. By taking up previous contributions by other users within one's own humour and referencing back to previous group discussions and topics, users create a sense of in-group identity. She explains: "The continual invocations of common knowledge assume that the others are familiar with these things and accordingly strengthens the shared bases on which group unity is founded" (Baym 1995:17). Danet (2001:107-152) also shows how joint humour brings strangers together during an online performance of Shakespearean theatre.

This joining force was also shown to be present for the variety known as "LOLspeak", which is a relevant source variety for the slang pupper talk studied here (see section 2.4). Miltner (2014) conducted focus groups with users consuming and producing the variety and found that "Lolspeak's main function was creating and enforcing group boundaries" (Miltner 2014:8). LOLspeak and associated varieties are also frequently classified as a humorous form of "speech play" (Bury & Wojtaszek 2017:31, Gawne & Vaughan 2011:103) as defined by Sherzer (2014:727): "Speech play is the playful manipulation of elements and components of language". Speech play and language play are of great interest to linguists, as they shed light onto "linguistic structure by revealing the ways in which various elements of language can be manipulated in different contexts" (Sherzer 2014:728). For the discipline of sociolinguistics, especially the relation between speech play, stylistic variation, and community affiliation is of interest.

Humour is often classified as subset of the broader term creativity (Kozbelt 2014:181). Gerrig & Gibbs (1988:3) claim that the main function of creativity is to expand the expressive potential of language beyond the established form-meaning pairings. While young children use creativity to name objects and action for which they have not yet acquired to conventional expression

(1988:4-5), adult speakers employ creativity to express their "greater range of experiences" and "to differentiate the world into finer categories than those established by convention" (1988:5). Gerrig & Gibbs also point out important social aspects of using creative language: its use establishes intimacy between those able to understand the utterance, it expresses personal opinion and stance, it can be used as a persuasive tool, it is helpful to avoid taboos and impoliteness, and it can be used to elevate one's personal status within a group (1988:7-13). The discussion section will elaborate to what extent these features also apply to the slang at hand.

2.5 Accommodation and audience design

Before looking at the theories of speech accommodation and audience design, which attempt to explain stylistic variation, it is first important to outline how style is understood within this paper. Fundamentally, this study will follow the definition of style as formulated by Rickford and Eckert (2001:2):

> The traditional delimitation of style in the variationist paradigm has been any intra-speaker variation that is not directly attributable to performance factors (in the strict sense) or to factors within the linguistic system.

The theory used in this paper tries to explain differences in style mainly through the speaker's relation to his or her interlocutors and the groups they belong to. As Bell (2001:141) puts it: "Style it what an individual speaker does with a language in relation to other people". At the same time, I also agree with Eckert (2004:43) when she states that style is not simply stable but acquires its meaning through ongoing construction. By combining my quantitative analysis with an ethnographic pilot study and discussing stylistic choices in concrete examples taken from the corpus within the discussion part later on, I hope to pay due attention to this characteristic of style (see also Schilling-Estes 2004).

An important theoretical foundation for the theory of audience design is Communication Accommodation Theory. It describes how interlocutors "use specific communication strategies (in particular, convergence and divergence) to signal their attitudes towards each other and their respective social group" (Giles & Ogay 2007:294). For this paper, especially convergence of linguistic features is of importance: according to the theory it is motivated by the desire to receive social rewards through similarity attraction. In mediated communication (such as CMC), such feedback is often not as readily available as in face-to-face conversation. Interestingly, on Reddit, feedback on one's utterance is provided by the whole community, not only be the previous commentator, through the use of up- or downvoting. In general, the lack

of immediacy on many online platforms might theoretically discourage users from using accommodating behaviour, as for example Danescu-Niculescu-Mizil et al. (2011:745) note. However, they find no such tendency in their analysis of accommodation on Twitter.

Building on Giles' model of accommodation theory, Bell (1984) proposes a more fine-grained model of "audience design", which distinguishes between different audience roles. He offers a classification distinguishing between addressee, auditor, overhearer, and eavesdropper (1984:159), which have varying degrees of influence on a speaker's stylistic choices. Applying this taxonomy to computer-mediated communication is not straightforward, as Androutsopoulos (2014:65) points out. For his research on audience design by bilingual Facebook users, he distinguishes the following three audience roles: 1. Addressees: "are those members of the networked audience who are directly addressed in a contribution", 2. Bystanders: "those members of the networked audience who are actively engaged in a particular […] exchange", and 3. Overhearing audience: "the entire social network" (see also the similar categorisation by Tagg & Seargeant 2014:172). To Androutsopoulos' statement I would add that not only does the online environment create new types of audiences, but different online platforms with their diverging participation structures also create different types of audiences. While his terminology is appropriate for the Facebook environment, the situation on Reddit is notably different. When responding to another user's comment or directly to a user's post, it can be assumed that this user is the immediate addressee. However, this might not necessarily be the case, as it is frequent practice to "hijack" a comment that is topically unrelated but has a high visibility in order to increase the visibility of one's own contribution. It is also less clear who should be regarded as an auditor: the group of people commenting under the same post and probably reading each other's comments beforehand? Should "lurkers", who only consume content and do not produce text themselves, be regarded as overhearers or auditors (Marcoccia 2004:131)? It is furthermore likely that users have diverging expectations concerning who reads (and responds to) their contributions (see also Marwick & boyd 2011).

When asked about their decision-making process during composing posts and comments, two informants in the pilot study made explicit reference to the desired reaction to their input: being upvoted by fellow Reddit users. Their frame of reference therefore seems to be those people that not only consume content, but that participate through voting. For example, informant 07 writes: "If you are among people you believe will react better to puppy talk you use it" (see Appendix 1). For the remainder of this study I will therefore distinguish between three audience roles only:

- Formal addressee: the user who composed the textual unit that the utterance technically responds to. In most cases this formal addressee will also be addressed content-wise, but this does not need to be the case.
- Active audience: users who actively engage in up- and down-voting on the platform and that therefore decide on the fate of a contribution (and ultimately the user's status within the community).
- Passive audience: all other people who consume the content without performing other actions on the platform. These people must not necessarily have a Reddit account.

What might further distinguish audience design on Reddit from the original theory is the importance assigned to the different audience roles. According to Bell (1984:159-160), the immediate addressee has the strongest influence on the design of a speaker's utterance, followed by auditor and the overhearer. How this translates to the Reddit environment will be taken up again in the discussion section.

Coming back to the theory, Bell (1984:151) relates stylistic, intra-speaker variation to broader inter-speaker variation between groups of speakers. He demands that: "Any model for style shift must account satisfactorily for that relationship" (Bell 1984:158). To make this connection explicit in the online environment is a difficult task, as social categories commonly used to explain variation (such as age, gender, or socioeconomic situation) are blurred and often unknown to the other users and the researchers alike. A further important point to discuss is the distinction between responsive audience design and initiative referee design (Bell 1984:186), in which the speaker orientates his or her style towards an absent group of speakers and their style. While Bell originally treated referee design as secondary, he later acknowledges that "these may be two complementary and coexistent dimensions of style, which operate simultaneously in all speech events" (2001:165). Similarly, Androutsopoulos (2014:64) states that style shift can always be both responsive and initiative: "any instance of human communication potentially is a combination of responsive and initiative style". The focus of the present study will, however, be on stylistic choices in response to previous utterances. Primarily because the responsive dimension is easier to measure and quantify, as Bell himself admits (2004:166-167), and because of the fact that initiative style shifts are previously been recorded in the context of longer conversations between two individuals. The communication structure on Reddit is noticeably different and people seldom engage in lengthy exchanges (similar to the newsgroups described by Marcoccia 2004:119).

Within responsive variation, Bell (1984:161-162) further distinguishes between "audience design", which he regards as primary, and other factors not directly relating to the audience,

such as “setting” and “topic”. Brown & Fraser (1979:34-35) lay out how this “scene” is equally important to consider when attempting to explain stylistic variation. They distinguish between “setting” (comprising bystanders, locale, and time) and “purpose” (consisting of activity type and subject matter). Transferring this classification to the online environment is not straightforward. In order to explain variation between different platforms, a faceted classification of the medium and situation factors (as provided by Herring 2007) seems appropriate in order to capture the characteristics of the communication environment. However, since this study is concerned with stylistic variation within the same platform, other factors need to be chosen. For this paper, special attention will be paid to the multimodal prompts as instances of subject matter. The analysis in this paper therefore mirrors Bell’s two-fold approach, as both audience and non-audience factors are taken into account: the audience will be represented by the stylistic choices within the previous textual unit (representing the formal addressee) and the subreddit (representing the active audience), whereas the scene will be represented by the features of the multimodal prompts.

The theory of audience design has been applied to computer-mediated communication in a number of studies. As already mentioned, Androutsopoulos (2014) researches audience design on Facebook within the context of language choice of teenagers. He proposes changes to the delineation of audience roles and suggests that audience design is always both responsive and initiative. Tagg & Seargeant (2014) apply a comparable classification of audience roles to investigate language choices of multilingual adult Facebook users, finding that the immediate addressee has the largest impact on language choice. Their finding is confirmed by Hinrichs (2016). In a similar vein, Honkanen (in press:160) identifies another set of audience roles relevant to the users of the Nigerian online forum “Nairaland”. Especially relevant to the study at hand is the paper by Pavalanathan & Eisenstein (2015), which investigates audience design of American Twitter users. Applying a similar methodology, their regression model reveals that local and non-standard features are employed more often when addressing the tweet to a local audience (indicated through explicitly addressing local users), compared to a broader audience (indicated through the use of hashtags). This tendency might, however, also be related to topical choices, as Shoemark et al. (2017:59) note: “users may use more hashtags when discussing political events than when discussing daily routines”. Shoemark et al. (2017:63) therefore take both the intended audience and the topic into account for their own study on Scottish regional features on Twitter. They find that both have an influence on the use of Scottish lexical variants. All these previous studies have confirmed that, in various CMC platforms, users are very much aware of who their (intended) audience is and take that into account when designing the style

of their utterance. However, most of these studies focus on Bell's (1984) audience factors only and pay little attention to setting and topic (with the exception of Shoemark et al. 2017). This study wants to take both aspects into consideration to compare their actual influence on stylistic choices on Reddit.

2.6 Multimodality

In the previous chapter we saw that non-audience factors play an important role for stylistic variation. While the overall setting of Reddit was already described in section 2.2.1 above, we will now focus on another aspect that might hold insights for the explanation of platform-internal variation: the multimodal prompts. Let us therefore now take a look at the combination of different modes in the online environment and its relation to language production and perception. According to van Leeuwen (2015:477), the term *multimodal* "indicates that different semiotic modes (for instance language and image) are combined and integrated in a given instance of discourse or kind of discourse" (see also Bateman 2017:7). Previous research on multimodality in language has mainly focused on oral communication, investigating the role and interplay of different modes, such as pitch, gestures, or gaze direction (e.g. Wahlster 1994 on deictic gestures). But not only is spoken language highly multimodal, the way people interact online is also increasingly characterised by multimodality. For example, the platform central to this study, Reddit, can be classified as an "interactive multimodal platform" according to the definition by Herring (2015:2), as many posts not only consist of plain text, but employ videos, pictures, or GIFs as well. Especially interesting are posts that combine multiple images or that incorporate further text into the image. It therefore seems appropriate to take the interplay of the various modes into account when researching language use in communities on Reddit.

Previous research on multimodality and CMC has mainly focused on graphical means being employed during language production. For example, Herring & Dainas (2017) focus on the use of graphicons (under which they subsume emojis, GIFs, stickers and others) and their functional specialisation within public Facebook threads. Focusing more on the integration of text and visual modes is the study by Bourlai & Herring (2014), in which they compose and analyse a corpus of Tumblr posts combining text and images. They find that posts containing images were overall more emotional and more positive then posts containing plain text only (Bourlai & Herring (2014:4). Similarly, Tolins & Samermit (2016) analyse the usage of GIFs in text messages (either replacing or accompanying plain text) and describe them as "embodied reenactments". The present study is, however, concerned with multimodal input and its impact on utterance production. The only study done in this direction is Lee & Barton (2011): in their

qualitative analysis of multilingualism on the image-sharing platform Flickr, informants named the content of the picture as one of the factors influencing their choice of language for the title and the description of the image. Other factors were their imagined audience, their "situated language ecology", and their perceived purpose of the platform (Lee & Barton 2011:52-54).

In addition to this scarcity of (especially quantitative) studies on multimodality online, Herring (2015:401) also points out how most methods and frameworks proposed for the analysis of CMC are not well equipped to handle multimodal environments. One of the first steps in this direction is O'Halloran (2011), who proposes "Multimodal Discourse Analysis" as a framework that might be useful for future qualitative studies of multimodal phenomena. One can therefore conclude that there is still much work to be done for linguists in order to accurately describe the multimodal nature of communication in the online environment. The constant evolution of new platforms and means of interaction is not facilitating this endeavour, but the present study still hopes to make a contribution in that direction.

3. Data and Methodology

This section will now provide a detailed description of the methods and data used for the present study. First, the ethnographic pilot study will be summarised, before moving on to the aspects of the quantitative analysis. This encompasses data collection, coding, and the statistical analysis of the data. Finally, the important question of research ethics needs to be discussed.

3.1 Ethnographic pilot study

Prior to the main analysis of this thesis, a qualitative pilot study was conducted focusing on orthographic variation only, which helped to inform the design of the quantitative analysis. The method chosen was discourse-centred online ethnography, as outlined by Androutsopoulos (2008). As suggested by him, the pilot study consisted of an initial phase of systematic, regular observation of the subreddits in question and a second phase of contact with the informants. The subreddits were selected due to their topical similarity and a sufficient overlap in the stylistic features employed. Instead of contacting the regular subscribers to the fora, the moderators of each subreddit were approached, as they were considered to be experts on their specific community. Out of eight subreddits and thirty-eight moderators contacted, a total of eight exchanges emerged (the anonymised transcripts of which can be found in Appendix 1).

The results gleaned from these conversations helped to inform the main analysis in several ways. First of all, participants provided several cover terms for the use of these spelling variants: the terms proposed are "doggo speak", "pupper speak", "pupper talk", "puppy talk" and "doggo

lexicon". Furthermore, participants pointed to a number of factors that might influence the choice of spelling variants: individual preferences, the subreddit, the type and content of the post, and the "cuteness" or mood of the post. Upon inquiry, many also admitted that the language used in the post or in previous comments might have an impact. The connotations named for the non-standard spellings included loving and endearing behaviour, acting "goofy" or "memey", as well as cuteness. Participants disagreed, however, on the origins of the non-standard spellings and on potential modelling after child-addressed language. While some said that pet-addressed language was inspired by child-addressed language, others argue that it is supposed to imitate animals' (assumed) thoughts (see the discussion in section 5.1.1). Demographic characteristics seemed not to be important to the informants.

3.2 Quantitative analysis

This section will now present the details of the quantitative analysis. In-depth description of each methodological decision taken during data-collection and processing is of vital importance to ensure transparency of the results (Berez-Kroeker et al. 2018:8). Complete reproducibility of this analysis, which should be the goal of every linguistic study (Berez-Kroeker et al. 2018:4, Flanagan 2017), can, unfortunately, not be achieved due to concerns about the privacy of the users contributing the data (see section 3.3).

3.2.1 Data collection

Let us now look at where the data for the corpus was taken from. The eight subreddits selected for the analysis are presented in Table 6 below, together with their size and the amount of material collected from them. In total, 356 posts and 4472 comments were sampled, which amounts to a total of 58,965 words produced by 3922 different contributors. The subreddits were chosen due to their topical similarity (they all focus on animals and the majority of them specialise in dogs) and their overlap with regard to the use of slang features. Furthermore, they represent communities with a wide range of numbers of subscribers; this allows for an analysis of the effect of group size on the use of the variants (Liimatta 2016:15). What makes the selection of subreddits especially interesting is that the same content would often be posted to several fora, often with the same or with a very similar title. This provides the opportunity to observe the influence the different subreddits might have on spelling choices in the comments.

Subreddit	Number of subscribers (as of 11.05.2020)	Number of posts collected	Number of comments collected
r/aww	24,706,818	45	942
r/rarepuppers	2,215,765	45	763
r/goodboys	7,018	45	120
r/Eyebleach	1,997,819	45	765
r/longboyes	58,473	45	374
r/barkour	223,676	44	407
r/dogswithjobs	659,699	45	741
r/dogs_getting_dogs	75,855	42	360
total		356	4472

Table 6: Overview of subreddits and amount of data collected.

Posts were collected on a weekly basis, i.e. once a week on a set day for each subreddit. The corresponding comments were not collected immediately, but in the following week. This decision was taken to allow the up- and down-voting of comments to settle before collecting the comments' karma (similar to the procedure in Liimatta 2016:18). Furthermore, this increased the number of comments being sampled, as more comments would accumulate over the week. The evaluation of the posts via the karma score and the total number of comments responding to a post were also collected in the subsequent week for the same reason.

A maximum number of five posts was collected per subreddit each week (in some communities the posts were less frequent). For the purpose of selecting posts that were appreciated by and representative of the communities, posts were sorted by popularity (instead of by time of creation). This way of sorting the posts also ensured that the number of comments was high enough, as newer posts might not have received that many comments at the time of collection. The comments, on the other hand, were sorted with the oldest comments first. This allowed for an analysis of whether the choice of community-specific features might also be linked to the overall success of a comment. In order to increase comparability of the subreddits, only posts about dogs were selected (there had to be at least one dog depicted in the visual mode). Whereas some subreddits are focused entirely on dogs, others (such as r/aww) can also feature other animals and even humans. Post collection started on 1st October 2019 and lasted until 9th December 2019. The presence of a non-standard spelling variant or other slang feature was no criterion for the collection of a post or a comment.[2]

[2] There was no minimum word limit imposed on the posts, as was done by Liimatta (2016:25), since the posts typically feature only short titles.

Concerning the comments, many options for selection presented themselves: for example, it would have been possible to go down the complete "comment tree" beneath a post. However, it was decided to only consider comments from the initial level and the three most popular replies to that comment, as well as two more responses on the third level. This allowed for the focus to remain on the direct influence of the initial comment. Working with replies to replies to replies would have decreased the comparability between the comments, as well as increased the potential inter-relatedness of the texts with other comments. To sum up, the five most popular posts about dogs were collected in each subreddit each week, along with the five oldest comments, accompanied by a maximum of three responses each (on the second level) and a maximum of two responses to those responses (on the third level).

Messages by bots were disregarded in general, as well as meta-comments from the moderators, which are automatically attached at the beginning of a comment section.[3] Reposts and cross-posts were included in the analysis, since the ensuing discussions might still differ. While a "repost" is the re-submission of content that is already known to the subreddit members, a "cross-post" is submitted to multiple fora at the same time. Comments that were hidden due to their unpopularity were still included. Occasionally, the username had been deleted, but the comment itself was still present; in those cases, the comment was nevertheless included in the analysis. Deleted comments (at the time of collection) were disregarded, as well as the comments replying to the deleted comment. If a comment was deleted after it had been collected it was not erased from the data set. Edits on the original comments were included, mainly because there is no way of reconstructing the original version (editing is, however, a rare phenomenon).

Several challenges came up during the process of data collection that had to be addressed. The first one was how to represent emojis and pictograms within the Excel sheet. Even though it would have been possible to copy and paste them, this would have posed challenges for transferring the data into R and for the word count in Python. Therefore, emojis and other occurring pictograms were rendered as descriptive lexemes enclosed in square brackets. An example of an original comment and the version entered into the data sheet is represented below (9-10). The same holds true for special formatting of the text, such as cursive or bold print, which had to be disregarded in general.

[3] For an example, see post p031 (https://www.reddit.com/r/dogswithjobs/comments/ddj0x8/the_goodest_boye_with_the_bestest_job/, last accessed 10.06.2020).

(9) ♪♪Pup-on! *boop boop!* Pup-off! *boop boop!* Pup-on, pup-off... The Pupper!♪♪

(10) c0129: [notes] Pup-on! *boop boop!* Pup-off! *boop boop!* Pup-on, pup-off... The Pupper! [notes]

A second challenge that came up was how to display comments that included line breaks. As for the emojis, Excel would have allowed multiple lines within one cell, but this would have hindered the further analysis. Line breaks were therefore represented as a forward slash / within the text. It was important to keep the distinction between the lines observable, as they play an important role for the poems that are occasionally composed by the users. Similar to the reconstruction of a word's pronunciation in the history of English through rhymes, the same method can be applied to non-standard spellings and newly coined lexemes in order to find out about their spoken realisations. An example is provided in (11), containing the slang lexeme *heck* (emphasis added).

(11) c0583: as we laying down to sleep,
my pup on top of me i keep
i love this little babe like *heck*
n as he slumbers on my *neck* [...]

In a concluding remark it should be mentioned that all the data was collected manually despite the fact that a corpus of Reddit submissions and comments exists: the Pushshift Reddit Corpus (Baumgartner et al. 2020). While this collection presents a valuable resource and was already used for a number of studies focusing on Reddit, it is unsuitable for the present analysis as it samples contributions from the platform as a whole by their time of creation. This makes it very difficult to recreate the original structure of the comment tree and to observe the influence of previous textual units.

3.2.2 Coding

This section will now list and justify the factors used for coding the posts and the comments. Focusing on the posts first, each post received a post identification number within the Excel sheet and was saved with the direct URL, the date of publication (in GMT +2h time zone format), the subreddit it was posted on, the username of the author, the exact title, and the exact flair (if applicable). To evaluate the post's popularity, the overall number of comments replying to it as well as the percentage of upvotes were collected. Furthermore, it was recorded whether the post was a crosspost from another subreddit (in which case it would be displayed with the title of the original post).

Concerning the visual modes, it was recorded whether the title was accompanied by a video or picture. The factor "mood", with the options positive and negative, was included since

informants in the pilot study pointed out how the overall mood of the post influenced their spelling. Three other factors to approximate cuteness were chosen based on previous research on animal characteristics influencing human behaviour. While it is known that the "Kindchenschema" (baby schema) and the associated facial features holds true for animals as well as for humans (Borgi & Cirulli 2016), it was beyond the scope of this paper to estimate, for example, the size of the mouth in relation to the size of the head (Borgi et al. 2014). Therefore, the factor "face" includes only the overall visibility of the dog's facial features, since this is a requirement for responding to facial features associated with the baby schema. Furthermore, the factor "age" was included since the vulnerability of animals is also known to affect human responses (Serpell 2004:147). Serpell (2004:147) also mentions how physical and behavioural similarity to humans impacts human responses, which informed the third factor, "anthropoidness". An overview of the coding criteria for the visual modes can be found in Table 7 below.

Feature	Values	Criteria
Age	young	dog identified as a puppy in the title, clear physical features
	adult	all other dogs
	both	depiction of a young and an adult dog
Anthropoid	humanoid	performing action not associated with animals, wearing clothing
	non humanoid	performing action associated with animals
Face	visible	both eyes, nose, and mouth visible
	partially	face visible partially or for parts of the video
	averted	face not visible or mostly averted
Mood	positive	default
	negative	mention of the animal's death or serious illness

Table 7: Coding of visual modes.

One should note, however, that the coding of the visual material is not as straightforward as the table suggests. While it is easy to assess the visibility of the face for a motionless picture, moving images make the task a lot more complex. Therefore, the visual features were only coded for the videos if the answer was very clear, e.g. if the face was visible throughout the whole duration of the video. A further problem was how to assess visual material depicting more than one animal at a time, as well as unusual visual material (such as a screenshot from an animated computer game showing a dog, or an animated comic strip featuring a dog). It was therefore decided to only apply the visual coding to material that allowed for a clear decision.

I also want to emphasise that I am very aware that the features I chose will not be able to catch the overall (subjective) cuteness of a post but can only be an approximation.

Moving on to the comments, they also received a comment identification number and were supplied with information about the subreddit, the author, and the date of creation. In addition, the level of the comment was recorded, so whether it was an independent comment (level 1), a response to an independent comment (level 2), or a response to a second level comment (level 3). For all comments on level 2 and 3, the comment they technically replied to was also stated. It was furthermore noted whether the comment contained a quote from a previous comment or a hyperlink to another post, subreddit, or website.

In the next step, a word count was assigned to both comments and posts, first manually and then with the help of a Python script (Python Core Team 2015). Diverging results of these two procedures were resolved manually. This double approach was deemed necessary due to the idiosyncrasies of the data set, such as the frequent insertion of blank spaces within words to emphasise length (such as <L O N G> in c3738). For the word count, smileys and emoticons were counted as one word (in line with Herring & Dainas 2017:2185)[4], as were hyphenated compounds and clitics together with the root. Words separated by a slash were counted as two words. Words crossed out ~~like this~~ and quotes from previous comments were still included in the word count and further analysis.

Both posts and comments were then coded for whether they contained any features associated with the pupper talk slang. The decision which features to include was based on the impression of the slang gained through the observations of the fora during the pilot study and the interviews with the moderators. Pupper talk was measured on a binary level: every comment and every post received a "yes" if they contained at least one feature associated with pupper talk; otherwise, a "no" was assigned. In addition, a numeric measurement, the "Doggo-score" was also implemented. Hereby, every word within the utterance received a value of 0 or 1. The value 1 was assigned if the word contained a non-standard spelling, if it was a lexeme specific to the slang, or if it was part of a non-standard grammatical construction. If a non-standard variant was part of a link or quote, it was coded as 0, since the user had no option of modifying it. Through dividing the count of features by the number of words in each post and comment, an overall ratio was calculated, which ranged from 0 to 1. A "Doggo-score" of 0 indicates that no

[4] This decision could be argued against, as smileys and emoticons do not present a context of potential presence of slang features. Including them in the word count can therefore lead to inaccurate measurements of the Doggo-score. However, since the ratio was not used within the regression model, this was considered to be of no mayor impact.

community-specific features were used within the textual unit, whereas a score of 1 shows that every word either contains community-specific features or is part of a community-specific grammatical construction. For the flair of the posts, only a binary distinction (pupper talk present: yes or no) was used, due to their overall shortness.

At this point it needs to be mentioned that in some cases it was not easy to distinguish whether a certain orthographic realisation was produced intentionally or should be regarded as a typing error or learner error. I decided on an inclusive approach to this problem, so in a case of doubt the instance was included in the analysis. The same holds true for the question whether a certain feature should be regarded as community-specific or as a general non-standard feature common on the internet. For example, the repetition of letters is frequently used in all sorts of online communication but was also mentioned as one of the features of pupper talk. Therefore, it was only counted if it was used on dog-related terminology or on dog-related topics, and if the same letter was repeated more than twice. The complete coding scheme (containing an account of which feature were counted and which were not counted) can be found in Appendix 2.

3.2.3 Statistical analysis

The coded data was read into RStudio (RStudio Team 2019), using R version 3.6.2 (R Core Team 2019) in order to produce visualisations and the logistic regression model. As a first step, the assumptions of logistic regression models needed to be tested (Levshina 2015:271). This includes:

1. Independence of the observations
2. Linear relationship between any quantitative explanatory variable[5] and the categorical response variable
3. No multicollinearity between the explanatory variables

The first assumption immediately stands out, since we can reasonably assume that the response variable of one comment is not independent from the linguistic features of the comment it responds to. This problem should, however, be mitigated by including the level of the comment and the presence of pupper talk in the previous textual unit as factors in the model. The posts, on the other hand, can be assumed to be independent. The second assumption does not need to be tested, since all explanatory variables are categorical in nature (Levshina 2015:17). Moving on to the third assumption, potential multicollinearity was investigated by calculating the "Variance Inflation Factor" (VIF).

[5] While there are different terminologies being currently used, this study will go along with the terms used by Levshina (2015:139): "response" and "explanatory" variables.

To calculate a binary logistic regression in R, Levshina (2015:257) presents two options: glm() and lrm(). For the present analysis, lrm() was chosen due to its versatility. Outliers and overly influential cased were investigated using the influencePlot() function described by Levshina (2015:153). Identified cases were only disregarded if they presented genuine "coding or sampling errors" and not if they could be interpreted as "inherent variability in the data" (Levshina 2015:155). The significance value was set in advance at 5 percent and all tests performed are two-tailed. As Levshina (2015:274) suggests, the model was also tested for overfitting by using bootstrapping with the validate() function. If previous research points towards a potential interaction between two or several predictors, Levshina (2015:268) also recommends testing for interactions. For example, if research on lexical variants includes different dialects as predictors, it is reasonable to assume that the constraints influencing the use of each variant are different depending on the dialect. Due to the high number of predictors used in the present study, testing for interactions between all of them was deemed unfeasible.

A further point that requires discussion is the method used for selecting the explanatory variables (Levshina 2015:149-152). There are several options that all have their benefits and drawbacks, which do not need to be discussed here. Since the study at hand is mainly concerned with the testing of hypotheses, the forced entry method was deemed the most appropriate (Field et al. 2012:457). In this approach, all explanatory variables that are considered theoretically relevant are entered into the model simultaneously. Table 8 below contains the explanatory variables that can be expected, on basis of the literature discussed before and the insights gained in the pilot study, to influence the value of the response variable. All explanatory variables are categorical and were subject to the default treatment coding (Winter 2020:131).

Explanatory variables	Response variable
Presence of pupper talk in previous textual unit, level, mood of picture, age of dog, anthropoidness, visibility of face, mode type, presence of pupper talk in flair, subreddit	Presence of pupper talk in comment

Table 8: Overview of explanatory and response variables.

One further comment concerning the predictor "presence of pupper talk in previous textual unit" is in order: for comments on level 1, this refers to the usage of slang within the post title, for comments on level 2 and 3 it refers to the usage of slang in the comment the utterance responds to. Readers might also notice that not all the information collected was included as explanatory variables. For example, the date of production of the text was not included as a potential factor, since the dates were collected in middle European summer time format.

Therefore, comments that might have been posted just a few minutes from each other, but posted from different time zones, might be represented as being posted on two separate days.

3.3 Research ethics

Opinions on research ethics concerning CMC studies vary to a great extent, as do the corresponding practices (Henderson et al. 2012:1). Many studies on CMC do not pay sufficient attention to research ethics or do not discuss them altogether. Bergstrom (2011:2), for example, explicitly states that no attempt was made to anonymise the Reddit users that are the subject of her analysis. This thesis, on the other hand, intends to discuss ethical implications in an appropriate manner. According to Fiesler et al. (2016), several ethical aspects deserve attention when research is conducted on online communities. These include the design of the research project, the informed consent of the participants, the collection and analysis of data, and the dissemination of the results. Each aspect will be discussed in turn.

The first question concerns how ethical principles can be secured during the design of a new research project, including the validation by ethical review boards (Fiesler et al. 2016:457-458). Since human subjects are not directly impacted in the study of CMC, as compared to medical studies, this step is often skipped in linguistic approaches. However, Fiesler et al. (2016:458) make the valid point that humans are still "implicitly involved" and may be harmed through large-scale harvesting of their online data, which is not the case in the study at hand.

Fiesler et al. (2016:458) also lay out the many challenges that researchers face when they want to obtain informed consent in large-scale online studies. Hutton & Henderson (2015:178) distinguish between two main types of consent: Whereas "secured consent" describes subjects giving their consent at a single point in time (often when subscribing to the "Terms and Conditions" of a platform), "sustained consent" includes participants being constantly asked about their approval of single pieces of data being shared. The "Privacy Policy" (valid as of June 2018), which Reddit users have to subscribe to upon creating an account, includes a section that can be interpreted as guaranteeing "secured consent". Reddit states:

> When you submit content (such as a post or comment or public chat) to the Services, any visitors to and users of our Services will be able to see that content, the username associated with the content, and the date and time you originally submitted the content. Reddit allows other websites to embed public Reddit content via our embed tools. Reddit also allows third parties to access public Reddit content via the Reddit API and via other similar technologies. (Reddit 2018)

However, Hutton & Henderson argue that "secured consent" is not sufficient, since participants' opinions might change over time (2015:185). They also subscribe to the statement that "simply

because a social network user chooses to share data [...], this does not give a researcher the right to collect such data" (2015:185). However, I would argue that this question requires drawing a distinction between different online platforms and the associated privacy expectations. Hutton & Henderson (2015) are mainly concerned with research on Facebook, a platform that arguably is less anonymous and more centred around authentic self-presentation. On Reddit, however, the anonymous users create nicknames and provide no further personal information.

As a third point, the collection and analysis of the data mainly requires that the researchers guarantee the anonymity of the people that produced the data. This is a difficult question to answer for the Reddit data at hand, since even if usernames are anonymised (as it is done in this paper), their posts and comments will still be traceable when typed into a search engine. However, instead of focusing on a user's individual linguistic profile throughout their posting or commenting history, only single utterances are collected for the study at hand. This is in line with the user preferences identified by Fiesler & Proferes (2018:6).

As a last point, the dissemination of the results is not of great relevance for this thesis, since the paper will only be accessed by a small number of people. While the fora themselves are openly accessible, the specific corpus collected for this thesis can only be viewed by myself and my supervisors.

4. Results

This section will now move on to the results of this study. The first subsection will provide descriptive statistics on the data set collected, while the second subsection will present the logistic regression model.

4.1 Descriptive statistics

Let us now first look at the overall composition of the data set. Inspecting one's data graphically is an important first step that should always be taken before any inferential statistics are applied (Field et al. 2012:231). This chapter will therefore look at each variable and measurement in turn to describe its distribution and how it relates to the pupper talk slang. The significance of these factors and their contribution to predicting the occurrence of pupper talk will then be investigated by the regression model in section 4.2.

4.1.1 Doggo-score

When looking at the distribution of the Doggo-score (the ratio measuring the amount of slang features within a contribution), we first need to distinguish between posts and comments. Table 9 displays the number and percentage of posts containing at least one community-specific feature, as well as the mean Doggo-score for each subreddit. One can see that the subreddit r/longboyes features the highest mean score, whereas r/dogswithjobs has the lowest mean score.

	Absolute number of posts		**Percentage of posts**		**Mean Doggo-score in posts**
	without pupper talk	with pupper talk	Without pupper talk	with pupper talk	
aww	37	8	0.82	0.18	0.05
barkour	27	17	0.61	0.39	0.12
dogs_getting_dogs	32	10	0.76	0.24	0.07
dogswithjobs	37	8	0.82	0.18	0.03
Eyebleach	35	10	0.78	0.22	0.08
goodboys	32	13	0.71	0.29	0.09
longboyes	16	29	0.36	0.64	0.25
rarepuppers	28	17	0.62	0.38	0.13

Table 9: Overview of frequency of pupper talk in posts of each subreddit.

Moving on to the comments, table 10 displays the same information as for the posts above. Again, r/longboyes has the highest mean Doggo-score and r/dogswithjobs the lowest mean score.

	Absolute number of comments		**Percentage of comments**		**Mean Doggo-score in comments**
	Without pupper talk	With pupper talk	Without pupper talk	With pupper talk	
aww	797	145	0.85	0.15	0.05
barkour	350	57	0.86	0.14	0.05
dogs_getting_dogs	303	57	0.84	0.16	0.04
dogswithjobs	644	97	0.87	0.13	0.03
Eyebleach	658	107	0.86	0.14	0.05
goodboys	97	23	0.81	0.19	0.06
longboyes	267	107	0.71	0.29	0.09
rarepuppers	611	152	0.80	0.20	0.06

Table 10: Overview of frequency of pupper talk in comments of each subreddit.

Let us now put the mean scores from both tables into relation (Figure 3). One can see that, overall, posts tend to have a higher mean score compared to comments. The differences between the subreddits are considerable.

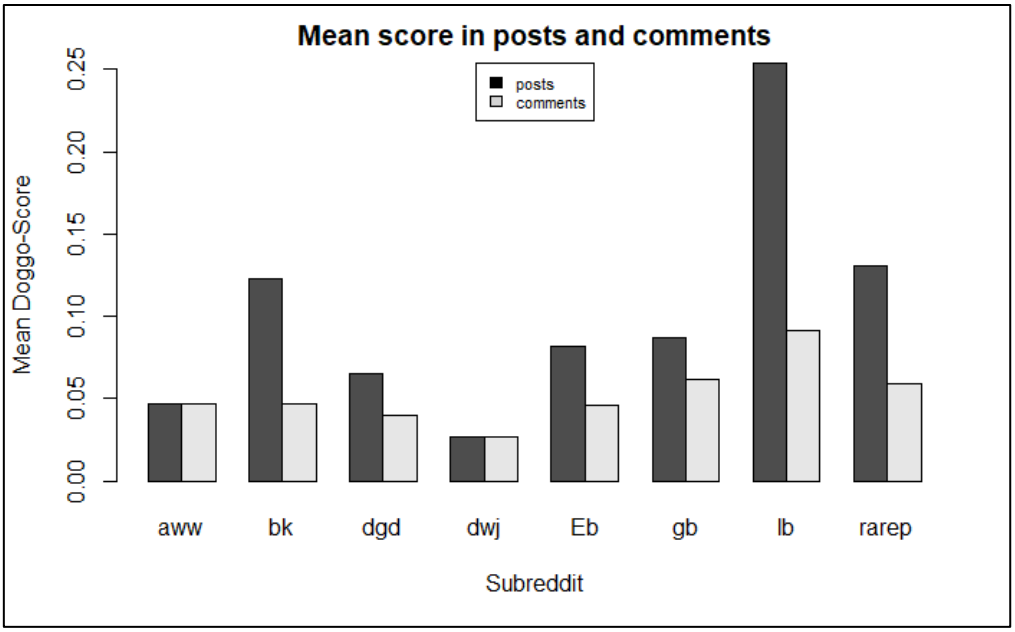

Figure 3: Mean Doggo-Score in posts and comments.

4.1.2 Word count

Moving on to the word count, the overall mean post length is 9.65 words (median: 7 words), whereas the mean length of the comments is 12.42 words (median: 8 words). Figure 4 displays the distribution of word counts for posts in all the subreddits. One can see that most posts are rather short with fewer than 20 words, while longer posts are more of an exception.

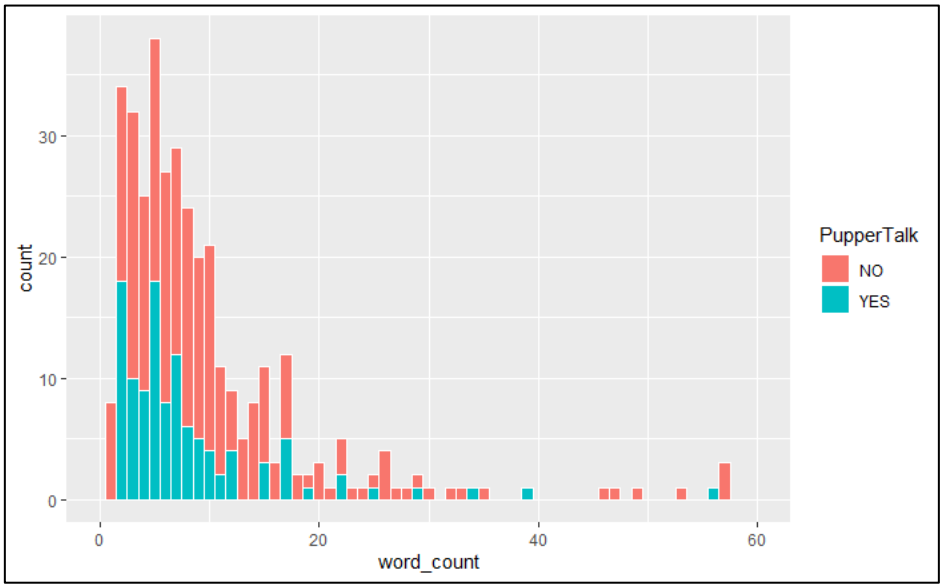

Figure 4: Word count in all posts divided by presence of pupper talk.

Figure 5 displays the same distribution for the comments. Compared to the posts we can observe a greater variability here, which might in part stem from the higher number of comments collected. Note that the diagram was cut off at a word count of 60 to aid comparability with the diagram for posts; there are only a few outliers with a count beyond that limit.

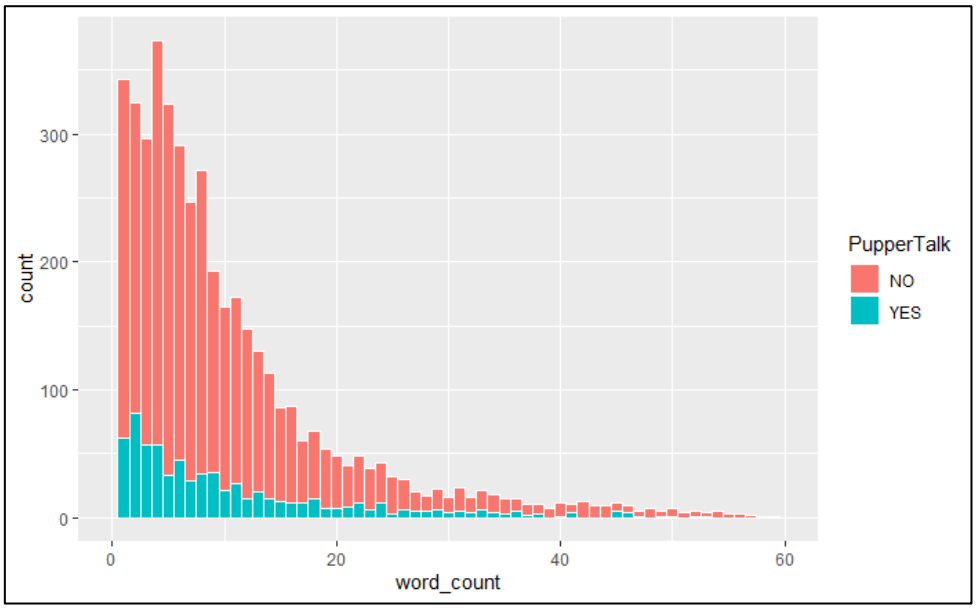

Figure 5: Word count in all comments divided by presence of pupper talk.

In both diagrams the colours indicate the number of textual units that contain pupper talk. Due to the overall smaller number of posts, the distribution is less even; but for the comments we can observe that the overall trend (as the word count increases the number of comments decreases) for comments containing pupper talk is the same as for comments without these features.

4.1.3 Mode used in the post

Let us now investigate the overall distribution of modes used in the posts. Figure 6 displays the proportions of pictures and videos used per subreddit. One can see that in some subreddits moving images prevail (and are even prescribed in r/barkour). Others, such as r/longboyes and r/goodboys mainly feature pictures.

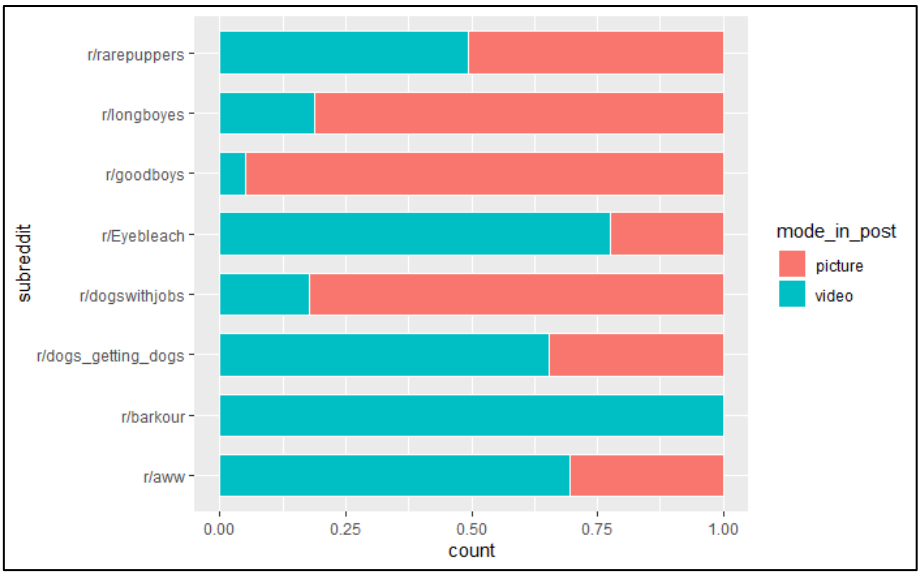

Figure 6: Distribution of pictures and videos per subreddit.

It is furthermore interesting to see whether this choice of mode has an impact on the usage of slang in the comments below the post. Figure 7 depicts the distribution of the slang in the comments divided by the choice of mode. Around 15% of all comments responding to videos contain pupper talk, while 19% of all comments replying to pictures contain it. However, one needs to be careful when interpreting this difference, since it is probably influenced by the preference for a particular mode in certain subreddits which either encourage or discourage the use of the slang.

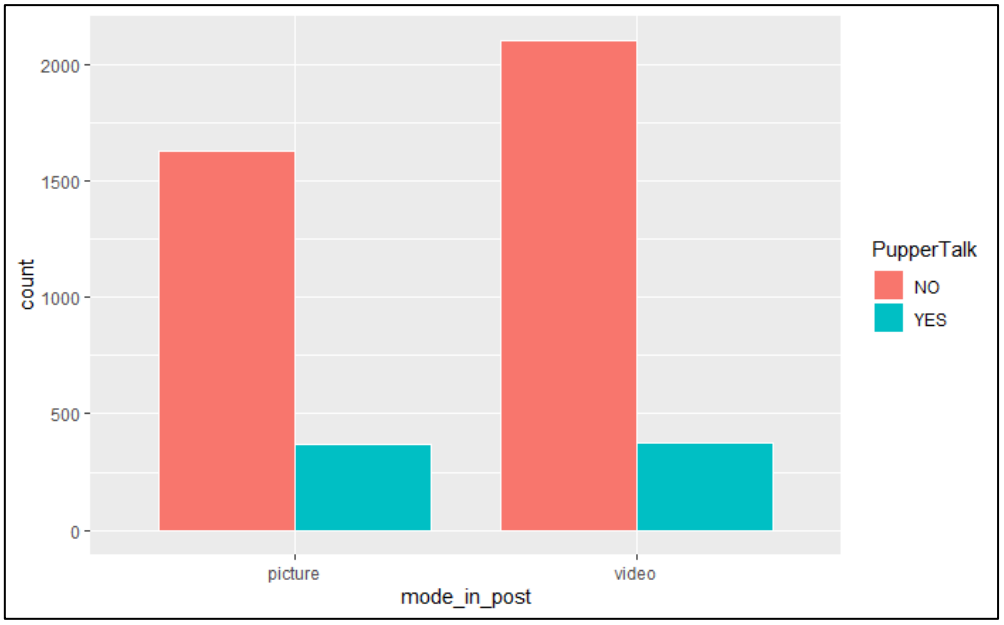

Figure 7: Number of comments containing pupper talk divided by mode in post.

4.1.4 Links and quotes

Providing links to other websites or subreddits as well as quoting previous comments serve important functions in the process of community-building (see section 5.2). It is therefore informative to inspect the distribution of these features over the range of subreddits investigated (Figure 8).

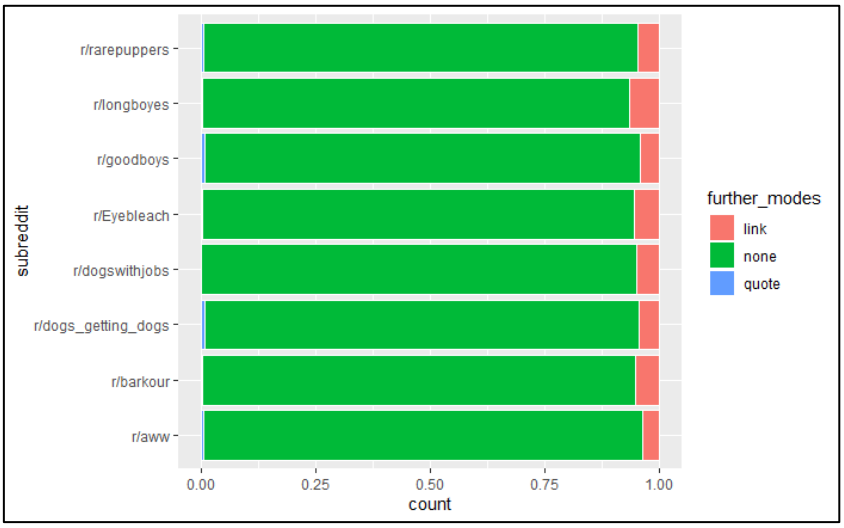

Figure 8: Percentage of comments containing a further mode for each subreddit.

We can see that providing links is overall an infrequent phenomenon – around five percent of all comments employ it. Quotes are even more infrequent: they occur in less than one percent of all comments. There appear to be no noteworthy contrasts between the subreddits. What is interesting is that comments containing a quote or link are less likely to also contain pupper talk, as Figure 9 shows.

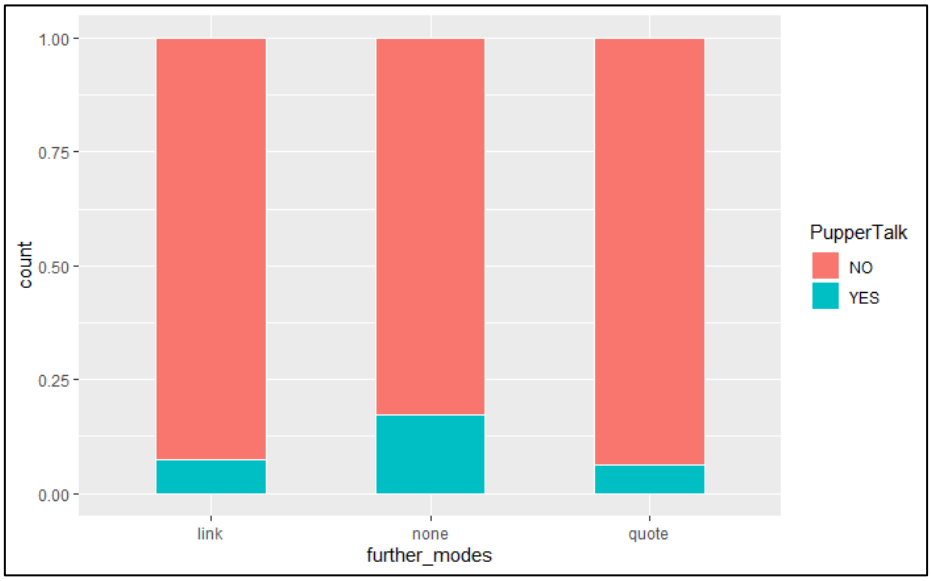

Figure 9: Percentage of comments with or without pupper talk containing a further mode.

While both, the inclusion of further modes and the use of pupper talk, can serve a community-building purpose, it seems that these methods are not frequently combined. This might have to do with the more humorous character of the slang, which users might feel does not go along well with providing additional information on a topic.

4.1.5 Number of comments

The number of comments is an informative measurement as far as it represents community activity. Figure 10 below shows that the number of comments below each post in the corpus varies considerably and is highly correlated with the overall size of the community. To aid readability, the x-axis was cut off at 1500 comments.

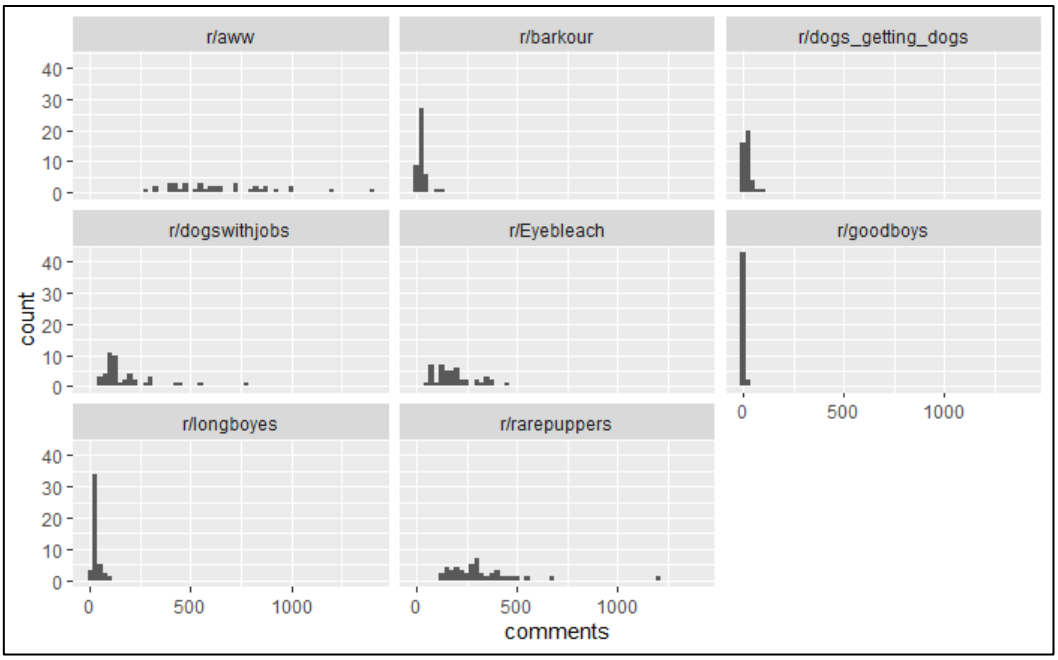

Figure 10: Number of comments for each post on the different subreddits.

The by far largest subreddit, r/aww (with over 23 million subscribers), also has the highest mean number of comments, followed be the second largest subreddit, r/rarepuppers, and so on.

4.1.6 Levels

Of further interest are also the levels of the comments. Figure 11 displays how many comments were sampled for each level, divided by subreddit. While the sampling method restricted the maximum number of comments that could be sampled from each level, the actual numbers obtained provide further insight into the participation structure of each subreddit.

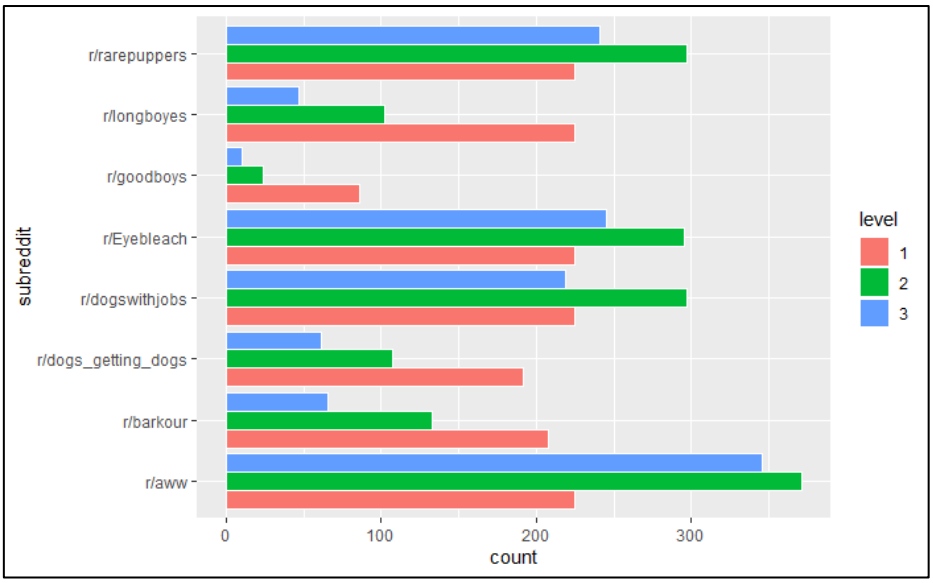

Figure 11: Number of comments sampled from different levels for each subreddit.

On the smallest subreddit, r/goodboys, only a small percentage of level 1 comments receive a reply on the second level and even fewer on the third level. In contrast, the largest subreddit, r/aww, shows more active involvement as there are more contributions on the second and third level than on the first level. What is even more interesting is the presence of pupper talk depending on the level of the comment, as depicted in Figure 12, averaged over all subreddits.

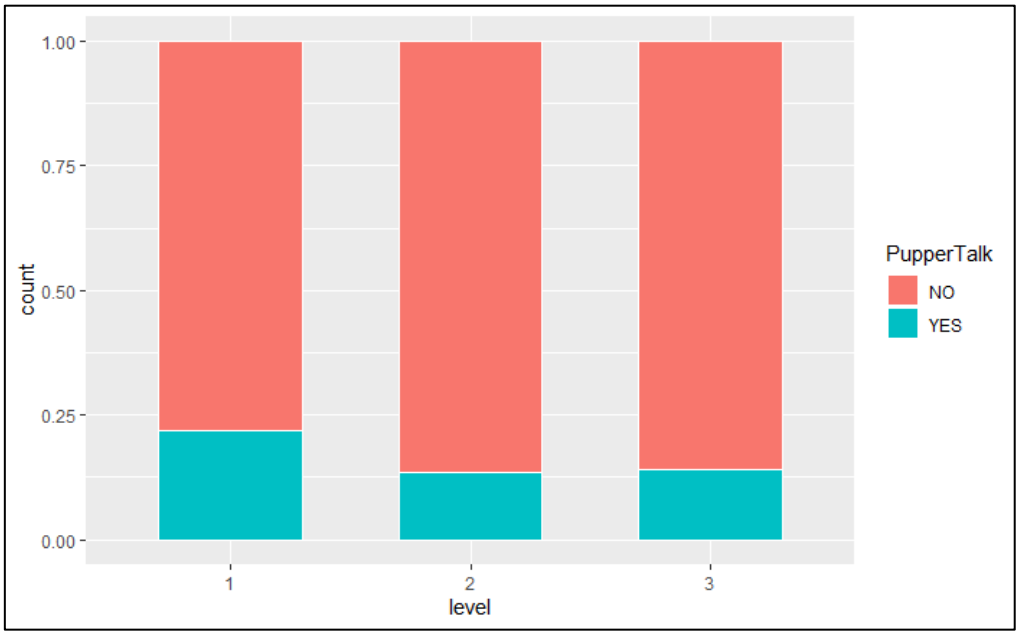

Figure 12: Percentage of comments containing pupper talk by level.

We can see that while 21.8% of all comments on level 1 contain pupper talk, the percentage drops to 13.5% for level 2 and to 14.1% for level 3. A possible explanation for this phenomenon will be laid out in section 5.2 below.

4.1.7 Cuteness of the visual mode

Three variables were used to approximate the cuteness of the visual modes: visibility of the face, anthropoid behaviour, and age of the animal. Looking first at the visibility of the animal's face, Figure 13 displays the distribution of pupper talk for the three levels "visible", "partially", and "averted".

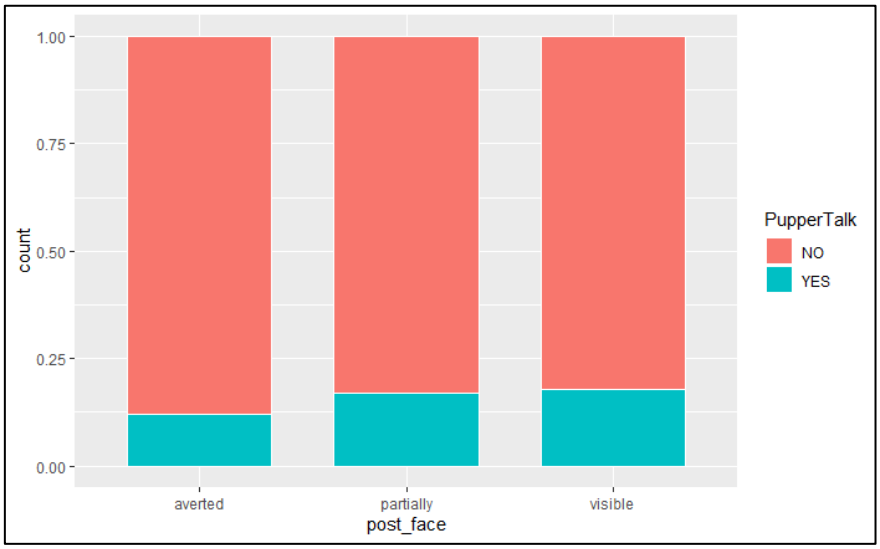

Figure 13: Percentage of pupper talk in the comments, divided by visibility of the face.

The figure shows that comments replying to a post with an averted face have the lowest percentage of pupper talk (12.0%). Posts in which the face is partially visible feature a higher percentage (16.9%), and posts with a completely visible face have the highest percentage of pupper talk in the comments (17.9%). Moving on to the age of the dog depicted, Figure 14 displays the same distribution divided by age.

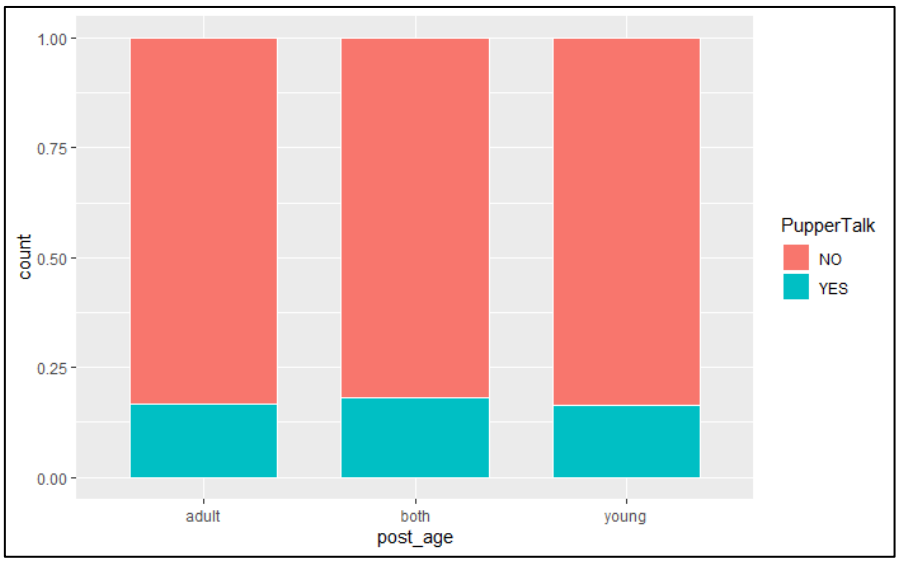

Figure 14: Percentage of pupper talk in the comments, divided by age of dog.

In this graph, the percentages are more alike. For the categories “young” and “adult”, 16.3% and 16.8% of all comments contain the slang, while there is a slight increase to 18.0% for the category “both”. This distribution indicates that age is probably not a relevant factor in the users’ decision. The last criterion applied is whether the dog performs an action usually associated with humans. For example, p055 features a video of a dog sitting on a stool in a crowded bar, looking at a television screen mounted on the wall – this qualifies as behaviour untypical for a dog. Figure 15 below again present the percentage of comments using pupper talk divided by this factor.

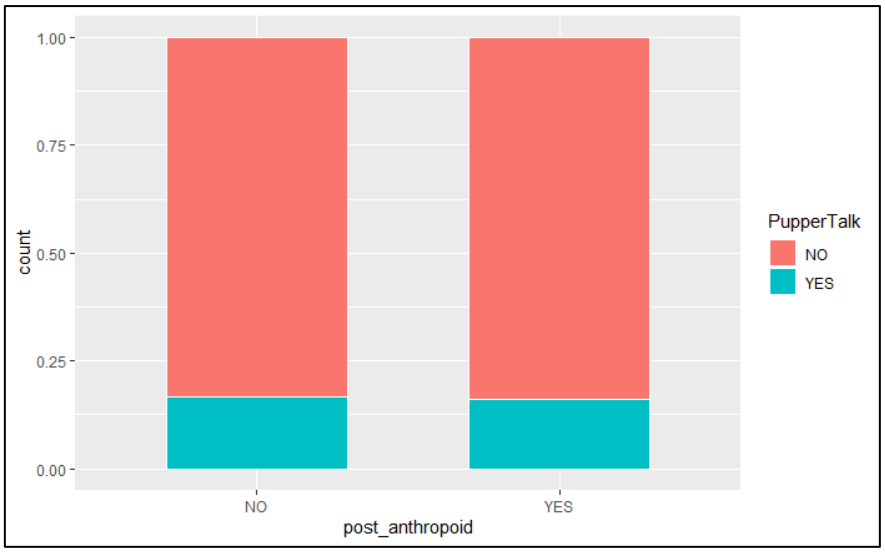

Figure 15: Percentage of pupper talk in the comments, divided by anthropoidness.

Here, again, there is little difference between the two categories: comments responding to an anthropoid post feature pupper talk in 16.0% of all cases, and comments responding to a non-anthropoid post feature pupper talk in 16.8%. This difference is marginal and points in the opposite direction than expected. Overall, we could therefore say that the factors used to approximate cuteness of the posts do not have a large influence on the usage of pupper talk within the comments.

4.1.8 Mood of the post

Let us now move on to the mood of the post, which, according to the informants, influenced their stylistic choices. Looking only at those comments that contained the slang, Figure 16 displays the relationship between the mood in the post and the Doggo-score within the responding comments (the dashed line representing the average Doggo-score for all comments containing pupper talk).

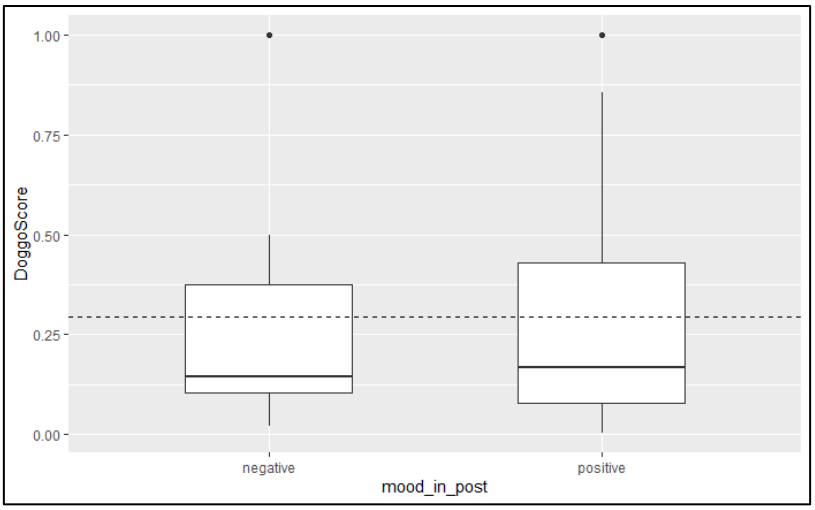

Figure 16: DoggoScore for comments containing pupper talk divided by mood in the post.

One can see that the mean score is slightly lower for comments responding to a negative post. However, one should note that overall the number of posts with a negative mood is relatively small: only 7 out of 356 posts fall into this category. One should therefore be careful to draw conclusions concerning this factor.

4.1.9 Appreciation of comments

For comments, the appreciation was measured by the "comment karma". Since karma distribution heavily depends on subreddit, the following plot (Figure 17) depicts the relation between karma and the Doggo-score divided by subreddits. For this plot only comments with a karma score below 6,000 were considered in order to aid readability (28 data points were excluded, all from r/aww).

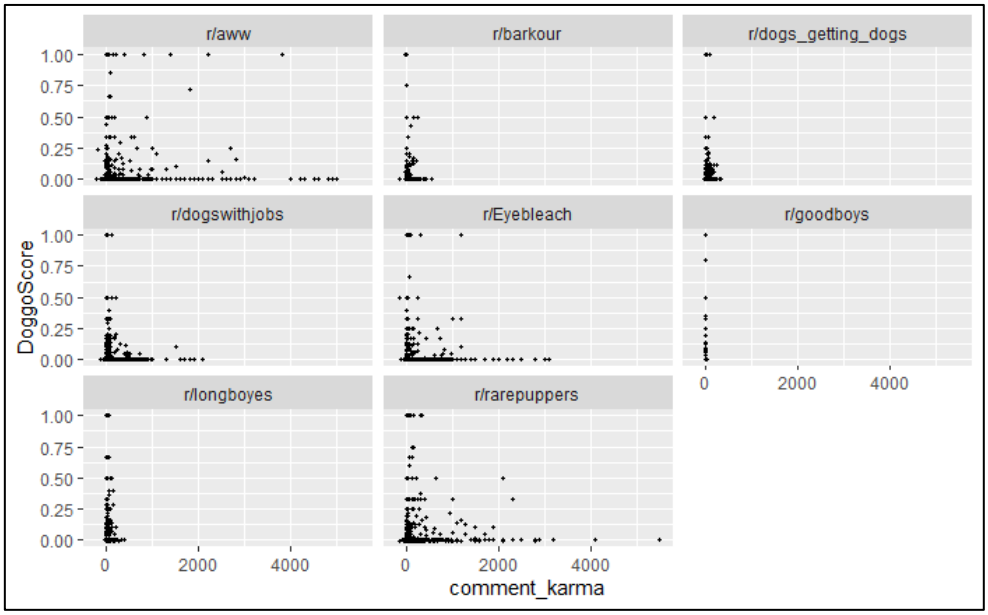

Figure 17: Relation between comment karma and Doggo-score, divided by subreddit.

The plots show that a higher Doggo-score is not related to a higher comment karma, even in the subreddits that explicitly encourage using the slang, such as r/rarepuppers. The plot furthermore provides insights into the activity on the subreddits, showing that larger subreddits also have higher karma scores (compare r/aww and r/goodboys).

4.1.10 Appreciation of posts

For posts, appreciation by the community was measured by the percentage of upvotes. Since the average percentage of upvotes varies considerably between the different subreddits, the following diagram (Figure 18) depicts the Doggo-score and the percentage of upvotes for each post divided by subreddit. We can see that there is no direct link between the Doggo-score of a post and its appreciation.

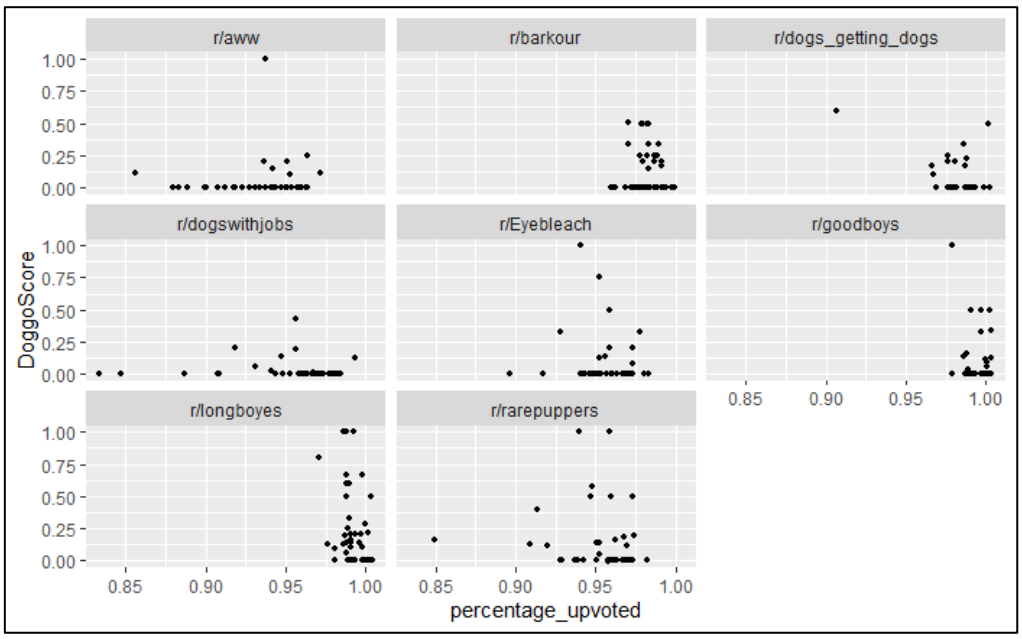

Figure 18: Percentage of upvotes and Doggo-score for each post divided by subreddit.

Due to the sampling method, which collected the most appreciated posts of each week, the overall upvote scores are rather high. The subreddits display interesting differences: the larger subreddits, such as r/aww and r/Eyebleach, feature a wider dispersion of upvote scores, whereas smaller subreddits, such as r/barkour and r/goodboys, have a higher average. This might well be caused by Reddit's inbuilt structure: if a post receives a lot of attention within the subreddit it was posted to, it will be featured on the main Reddit frontpage. There it will also be viewed by people who do not subscribe to the subreddit and who might not share the interest in the topic. This greater exposure to non-subscribers might lead to a lower percentage of upvotes. On the other hand, we can assume that only people who intend to do so will see the content on small subreddits such as r/goodboys.

4.2 Regression model

After presenting the descriptive statistics, let us now inspect the output of the logistic regression model predicting the probability of pupper talk occurring within a comment. Below we see the model statistics and the relevance of the individual predictors for the logistic regression model (Figure 19).

```
Logistic Regression Model

 lrm(formula = PupperTalk ~ level + previousPTbinary + mood_in_post +
     post_face + subreddit + mode_in_post + post_anthropoid +
     post_age, data = comments_total)


                       Model Likelihood     Discrimination    Rank Discrim.
                          Ratio Test           Indexes           Indexes
 Obs           4344    LR chi2     175.23    R2        0.067    C       0.658
  NO           3621    d.f.            17    g         0.564    Dxy     0.316
  YES           723    Pr(> chi2) <0.0001    gr        1.757    gamma   0.318
 max |deriv| 7e-09                          gp        0.081    tau-a   0.088
                                            Brier     0.133

                                Coef    S.E.   Wald Z Pr(>|Z|)
 Intercept                     -1.5139 0.1486 -10.19 <0.0001
 level=2                       -0.4360 0.1010  -4.32 <0.0001
 level=3                       -0.3616 0.1110  -3.26 0.0011
 previousPTbinary=YES           0.8113 0.0930   8.72 <0.0001
 mood_in_post=negative         -0.0708 0.3286  -0.22 0.8295
 post_face=partially            0.0996 0.1054   0.95 0.3445
 post_face=averted             -0.2125 0.1511  -1.41 0.1595
 subreddit=r/barkour           -0.1511 0.1867  -0.81 0.4186
 subreddit=r/dogs_getting_dogs -0.1408 0.2347  -0.60 0.5487
 subreddit=r/dogswithjobs      -0.2636 0.1654  -1.59 0.1111
 subreddit=r/Eyebleach         -0.0511 0.1461  -0.35 0.7265
 subreddit=r/goodboys          -0.0530 0.2696  -0.20 0.8441
 subreddit=r/longboyes          0.3179 0.1721   1.85 0.0648
 subreddit=r/rarepuppers        0.2026 0.1380   1.47 0.1422
 mode_in_post=video            -0.1816 0.1128  -1.61 0.1073
 post_anthropoid=YES           -0.0039 0.1192  -0.03 0.9742
 post_age=young                -0.0121 0.1228  -0.10 0.9218
 post_age=both                  0.1694 0.2101   0.81 0.4199
```

Figure 19: Output of logistic regression model.

Let us first inspect how well the model overall explains the presence or absence of pupper talk features. Since the value Pr(>chi2) is smaller than 0.0001 we know that the model overall is significantly better than a model without any predictors. However, the concordance index C only has a value of 0.658, which implies that the model is only able to explain a small part of the overall variation that we find in the data.

Next, the individual contribution of the predictors can be assessed. Readers might notice that not all the predictors originally coded were included in this model: one predictor, “pupper talk in flair of the post”, had to be excluded since it only applied to a small fraction of the data set, which resulted in too many data points being dropped from the model. In total, three predictors reach significance: level 2 ($p<0.0001$), level 3 ($p<0.005$), and the presence of pupper talk within the previous textual unit ($p<0.0001$). A further predictor that comes close to statistical significance is the subreddit r/longboyes ($p=0.0648$). Out of these predictors, level 2 and level 3 lead to a smaller likelihood of pupper talk occurring. On the other hand, the presence of pupper

talk in the previous unit, as well as the subreddit r/longboyes, lead to a higher likelihood of pupper talk being used.

Concerning the validity of the model, all assumptions are met: observations are independent, there are no numeric predictors that need to be tested for linearity, and testing for multicollinearity reveals that all values are well within the acceptable limits (maximum 2.28). Checking for overfitting also shows no reason for concern. While looking for outliers and overly influential cases, two data points stand out: comments c1650 and c1651, both responding to a post on a dog with unusual colouring, which are displayed below (12-13). Since both appear to be authentic and valid utterances, there is no reason to exclude them from the data set.

(12) c1650: Insane Clown Puppy

(13) c1651: Congrats you just made me dislike a dog for the first time ever in my life

The theoretical implications of the model output will be discussed below in section 5.2.

5. Discussion

After presenting the results, this section will now discuss the implications of the findings. First, a linguistic description of the slang pupper talk as found on pet-centred subreddits will be presented. Afterwards I will address whether the subreddits investigated qualify as virtual communities and as communities of practice. Following this important terminological clarification, we will move on to the factors influencing the presence or absence of the slang. Within the last section some methodological questions relating to the quantitative study of stylistic phenomena online will be discussed.

5.1 A linguistic description of "pupper talk"

I will now provide an overview of the features of pupper talk as it is used on dog-centred subreddits in the year 2019. Grouping these features into categories such as lexical and grammatical proves difficult in some cases, as many of them play with different levels of language at the same time. For example, the word *pawsome* is classified as the result of blending here, but as an orthographic alteration in Leppänen (2015:15). The following lists include both the slang observed within the data, as well as the features pointed out by the informants, even if not captured during the data collection process. For items occurring more than twice, frequency counts (per million words, henceforth "pmw") were calculated. Within the first section, I will discuss which varieties might have inspired the pupper talk slang.

5.1.1 On potential templates for “pupper talk”

Before discussing the features associated with pupper talk, it is worth considering what potential templates the slang might be modelled after. When asked about what might have inspired the irregular orthography, participants in the pilot study had varying explanations. In total, there are four potential templates: how humans speak to small children, how humans speak to pets, how infants speak, and how humans imagine dogs to think or speak. The differences between some of those are quite marginal: researchers have previously confirmed that pet-directed speech has considerable similarities to infant-directed speech (Ben-Aderet et al. 2016:2). Therefore, these two will be considered as one template only. Similarly, the imagined language of pets is likely to be similar to certain stages of first language acquisition. Let us first look at the evidence for each of the options. Several informants mentioned that they see the irregular orthography as an imitation of pet-addressed or child-addressed speech (14-15):

> (14) I08: I'd say it's probably modeled after how people speak to their pets.
>
> (15) R: […] Do you happen to know how this type of spelling came about? What the motivation behind it was?
> R: For example, were people trying to imitate how they spoke to their pets in real life?
> I01: Yes, that's exactly it! A lot of the time people speak to animals like toddlers, or of the sort, but it would be weird to type "wHOUSAgud bouoy" or something, so instead, it has evolved to be typed with minor mispells, like a toddler would

Informant 01 alludes to an important caveat: one of the most salient features of infant-directed speech is its phonology and prosody. Caregiver speech is characterised by slower pace, higher pitch, exaggerated intonation, and longer pauses (O’Grady & Cho 2001:353). All of these properties are difficult to transfer to a primarily written CMC environment. Research on pet-directed speech has so far also focused exclusively on phonology and prosody: within their study on dog-directed speech, Ben-Aderet et al. (2017) show that humans employ a higher pitch when addressing dogs of all ages – despite the fact that only puppies react better to this type of speech. To my knowledge, there are no studies on the grammatical or lexical features of pet-addressed speech. This poses the question whether the plethora of non-standard grammar and the lexical inventions listed below really have a spoken equivalent in dog-directed speech. Informant 03 explicitly denounces this possibility (16):

> (16) I03: […] I've heard people say doggo in real life, but I think that happened after the term popped up online.

Apart from infant- and pet-directed speech, there were also references to first language acquisition. Informant 04 recounts how people disliking the slang derogatorily refer to it as “baby talk” (17):

> (17) I04: Some people hate it and often leave comments saying "it's dog, not doggo" or "puppy, not pupper, stop using baby talk", and it's like they're instantly out-group.

Investigating the same question for the variety LOLspeak, Gawne & Vaughan (2011:102-103) admit that some grammatical features are reminiscent of first language acquisition errors, but due to the overall complexity of the language used they doubt that imitation of learner errors is the only inspiration. As a final potential template, some informants mentioned that the style is used to imitate how one imagines a dog to think or speak (18):

> (18) I03: The alternate spellings, the way I see it, try to “mimic” how our dogs might talk or think.

This explanation is supported by the fact that many comments are composed from a dog’s perspective. The slang used within these can therefore be interpreted as being an imitation of what the author imagines a dog’s utterances to be like. This becomes apparent in the comment below (19), in which a user formulates the imagined words of an older dog towards a newly adopted puppy (emphasis added):

> (19) c0422: Cody: Listen baby brother, you have the luck. We have the best *hoomans* here. They give us *smackos*, take us on *walkies*, rub our bellies, and call us good *bois* all the time. Just one rule. / Jax: What? / Cody: Do business outside, never in house.

In her study on DoggoLingo on Facebook, Bivens (2018:3-4) also mentions how DoggoLingo includes both humans as speakers as well as pets as speakers. One should furthermore note that humans writing from their dog’s perspective is productive in the genre of dog diary-blogs discussed by Leppänen (2015). The “doggielect” occurring in her data could be interpreted as an earlier and attenuated version of pupper talk, and is described as an instance of “stylization” (Leppänen 2015:16).

In sum, there are several challenges for identifying a template for the slang pupper talk. First of all, users themselves disagree on what the slang might be modelled after. Secondly, there is no complete description of pet-directed speech as a variety that would allow for a thorough comparison. What further complicates the question is that pupper talk combines elements of several distinct stylistic phenomena stemming from different platforms, which might be modelled after conflicting templates. A detailed qualitative analysis of all slang utterances collected and their speaker might hold further insights. Since we cannot arrive at a definite

answer at this point, the following sub-chapters will point out any similarities to the potential templates whenever they seem plausible.

5.1.2 Lexical features

Since slang is frequently considered "primarily lexical" (Malmkjær 2010:489), let us first focus on how the lexicon is expanded in pupper talk. The slang uses various word formation processes to create novel lexical items. As mentioned by Mattiello (2005:19), slang employs word formation processes that we would also find in standard English; among these we can count compounding, conversion, and shortenings. On the other hand, slang incorporates word formation processes that are relatively uncommon in standard English, such as infixation (which is not found in the data at hand), reduplication, or onomatopoeia. Some word formation processes that are common in general are realized with unusual morphemes, such as derivation with <-o> (Mattiello 2005:12). The following list now presents the strategies used within the data set to expand the lexicon together with a selection of examples:

- Blending:
    - With *bark*: *starbarks*/*starborks* (*Starbucks* + *bark*), *Cerebork* (*Cerberus* + *bark*), *barkourist* (*bark* + *parkourist*), *barkour* (427.17 pmw, *bark* + *parkour*), *borkday* (*birthday* + *bark*), *barkista* (*bark* + *barista*), *Cybork Dogs* (*Cyborg* + *bark*)
    - With *pup* or *pupper*: *puppoccinos* (*cappuccino* + *pup*), *pup-kin* (*pumpkin* + *pup*), *half-pup* (*half-pipe* + *pup*), *pup club* (*fight club* + *pup*), *pupendicular* (*perpendicular* + *pup*), *pupset* (*pup* + *upset*), *Puplates* (*Pilates* + *pup*), *meerpups* (*meerkats* + *pup*), *puppervisor* (*pupper* + *supervisor*), *pupdate* (*pup* + *update*), *parapupper* (*paratrooper* + *pupper*), *telepuppy* (*telepathy* + *puppy*), *pupperoni* (*pepperoni* + *pupper*), *American Ninja Pupper* (*American Ninja Warrior* + *pupper*), *Cyperpup* (*Cyberpunk* + *pup*)
    - With *dog*: *dorse* (*dog* + *horse*), *dogtor* (*dog* + *doctor*), *dogist* (*racist* + *dog*), *meerdogs* (*meerkats* + *dog*)
    - With *paw*: *therapawtic* (*therapeutic* + *paw*), *pawsome* (*paw* + *awesome*), *paw-se* (*paw* + *pause*), *mission impawssible* (*paw* + *impossible*), *pawsitive* (*positive* + *paw*)
    - With *woof*: *American Ninja Woofier* (*American Ninja Warrier* + *woof*), *Wooftradamus* (*Woof* + *Nostradamus*)
    - With *fur*: *furever* (*forever* + *fur*), *Affurmative* (*affirmative* + *fur*)
    - With breed names: *recorgnition* (*recognition* + *Corgi*), *Shiberus* (*Cerberus* + *Shiba*), *Shusky* (*shepherd* + *Husky*), *corgopractor* (*chiropractor* + *corgi*), *whipper schnauzer* (*whippersnapper* + *schnauzer*)
    - With *poop*: *spoopy* (*spooky* + *poop*), *poopervisor* (*poop* + *supervisor*)
    - Others: *Han Solong* (*Han Solo* + *long*), *tailcopter* (*tail* + *helicopter*), *Luke Sky water* (*Luke Skywalker* + *water*), *koalifications* (*koala* + *qualifications*), *salivation* (*saliva* + *salvation*), *Assassin's breed* (*Assassin's creed* + *breed*), *ear-*

ectile dysfunction (*erectile dysfunction + ear*), *ginormous* (*gigantic + enormous*), *Rex Games* (*X Games + Rex*), *cowraffe* (*cow + giraffe*), *koala-ty* (*quality + koala*)

- Shortenings: *lab* (279.30 pmw, 'Labrador'), *dobe* ('Doberman'), *pup* (1412.94 pmw, 'puppy'), *pit* ('Pitbull'), *great pyr*/*pyrs* ('Great Pyrenes'), *Boston* ('Boston terrier'), *blacklab* (98.58 pmw, 'Black Labrador'), *grey* ('Greyhound'), *Pems* ('Pembroke Welsh Corgis'), *Newf* ('Newfoundlander')
- Initialisms: *BC* ('Border Collie'), *GSD* ('German Shepherd')
- Conversion: *fuzzies* (from *fuzzy*, plural, 'fluffy dog')
- Reduplication: *tippy taps*, *lookie-likies*, *noodle poodle* ('greyhound'), *wiggle biggle*, *spinner winner*
- Onomatopoeia: *boop* (295.73 pmw), *blep*, *blop*, *mlem* (180.72 pmw), *floof*, *floofy*, *nom*, *aoow*/*awoo*, *pawp*, *woof* (197.15 pmw), *bop*, *omnomnoming*, *nomf*, *plop*, *doot*, *sproing*
- Compounding: *sky water* ('rain'), *grass dog* ('cow'), *noodle horse* ('greyhound'), *velvet hippo* ('pitbull'), *long girl* (98.58 pmw, 'female greyhound'), *long boy* (558.60 pmw, 'male greyhound'), *danger noodle* ('snake'), *meow pupper* ('cat')
- Neologisms: *splooting* ('lying with the belly flat to the ground, hind legs stretched apart')
- Derivation:
    - Suffixation with <o>: *doggo* (1330.79 pmw), *treatos*, *ear floppos*, *wolfdoggo*, *druggos*, *huggo*, *schmackos*, *smackos*, *puppo*, *friendo*, *boyos*
    - Suffixation with <er>: *pupper* (1215.76 pmw), *kitters*, *woofer*, *napper*, *snooter*, *borker*
    - Suffixation with <y> or <ie>: *doggie* (164.30 pmw), *doggy* (115.01 pmw), *tuckies* ('legs tucked in'), *zoomy*/*zoomie* (82.15 pmw, noun, from the verb *zoom*), *doby*/*dobbie*/*dobie* ('Doberman'), *bosties* ('Boston terrier'), *walkies*, *chessie* ('Chesapeake Bay Retriever'), *boxy* ('Boxer'), *pittie*/*pitty* (Pitbull), *dutchie* ('Dutch shepheard'), *sheltie* ('Shetland sheepdog')

We can see that blending is among the most popular word formation processes in pupper talk, even taking lexemes that were previously modified as input (such as the shortening *pup* or the alternative spelling <bork>). Shortenings are mainly used for dog breeds. Within the area of compounding we find the interesting tendency to create new terms for concepts that already have an established name in standard English. This reminds us of the tendency of children during first language acquisition to create novel terms for concepts or objects for which they do not know the established term, as described by Gerrig & Gibbs (1988:4-5) and O'Grady & Cho (2001:344).

What is striking is the overwhelming absence of word formation processes that Herring (2020:4) describes as particularly productive in the online environment, such as acronyms and alphabetisms – with the exception of the two instances of initialism. This is not to say that lexemes resulting from these processes did not appear in the data (for example *lol* for "laughing out loud" or *op* for "original poster"), but they were not classified as being an integral part of the slang. Apart from the lexemes mentioned above, pupper talk also employs some terms that have been associated with slang for a longer time:

- Slang terminology: *heck* (82.15 pmw), *hecking/heckin* (115.01 pmw), *bamboozle* (98.58 pmw, 'mislead'), *X/10* (230.01 pmw), *tootsie* ('paw')

The frequent use of the noun *heck* and the adjective *hecking* as a replacement for swear words is already mentioned by Punske & Butler (2019:2) and Mattiello (2005:14). Furthermore, Bivens (2018:1,2,11) also lists the lexemes *bamboozle* and *heck*, as well as the *X/10*-construction as distinct features of the DoggoLingo-slang she studies on Facebook. Both *heck* and the *X/10*-construction have reached such salience that Golbeck & Buntain (2018) devote their whole study to investigating the propagation of these two features across different online platforms. A further interesting aspect of pupper talk is the tendency to replace lexemes within fixed phrases. The data contained three instances of *paw* being used instead of *hand* and three other cases.

- *paw* for *hand*: *the situation at paw*, *in safe paws*, *left pawed*
- others: *through stick and thin*, *through tick and thin*, *once in a whale*

Summing up, pupper talk employs a variety of different word-formation processes to expand the lexicon and to create unique terminology that is difficult to understood outside of the communities. While doing so, users show great creativity by using already modified lexemes as input for word formation processes.

5.1.3 Grammatical features

Apart from the vast creativity we can observe in the lexical domain, pupper talk also contains a number of grammatical features, including both non-standard morphology and non-standard syntax. One of the most salient is probably the irregular comparative morphology:

- Suppletion in combination with suffix: *bestest* (180.73 pmw), *betterest*
- Use of suffix instead of suppletion: *goodest* (262.87 pmw), *goodestest*

This irregular morphology does not appear with any other adjectives, which might indicate that this is not perceived as a productive rule by the users, but this tendency is probably also related

to suppletion being a limited option within the English comparative system. This irregular use of the comparative and superlative reminds us of the overgeneralisation of irregular morphology during first-language acquisition (O'Grady & Cho 2001:341). Apart from this we also find word classes used in a syntactic position that they do not normally assume in standard English. Within the data set, three types of changes occur:

- Adjectives used as nouns: *such cute, distract with cute, such happ, most cute, too much long*
- Verbs used as nouns: *more enjoy, much startle, many scared, visible shook, invisible shook, much smile, such dancing, much guarding, such sproing*
- Verbs used as adjectives: *very scare, very spook*

All of these were not taken out of context but occurred as these isolated fragments. Bury & Wojtaszek (2017), in their study on LOLspeak, also mention a phenomenon they call "manipulating categories", in which "the distinction between adjectives and nouns is blurred" (2017:36). However, the examples they quote, such as "I has a happy", have no equivalent in the data set at hand and are quite different from the changes listed above. It therefore seems as if the 'flexibility' of word classes is a shared phenomenon in online slang, but the specific realisations differ.

In addition, there is a small number of irregular grammatical constructions within the data. Apart from the "do-a-verb" construction, all of them occur only once or twice, and are probably better described as relatively fixed (often quoted) phrases than as wide-ranging rules. The first construction could arguably also be listed under the word-class changes but appears to be more elaborate than those in that it is also used to form complete sentences.

- Do-a-verb construction: *longboye does a dilemma, She's doing a shrink, He do a gentle excite, Did a good sit, Doing sit!, did me a scare there*
- Can-has-construction: *I can has cheezburger?, you can has hugs*
- Irregular determiner use: *i have ball too, you have the luck, Balto is best boy*
- Omission of *BE*: *she a good girl, she building a house, He looooonnnnggggg, Why tiny human no pet me?, Why tiny human hide?, He a big good boy, He muscley, she a n g e r e y, He very muscle, He a fun boi*
- Irregular plural formation: *stuffs, foods, foots* ('feet')
- Pronoun mismatch: *him loves bred* ('he loves bread')

Many of these constructions are also mentioned in previous studies. Bivens (2018:7-10) discussed the "Do-Rule" at length and proposes four "transformations" that need to be applied to convert a sentence into this formula. The construction is also treated in detail by Punske & Butler (2019), assuming that "speakers are using underlying grammatical principles to conduct the language game" (2019:4). The construction seems to be considerably more productive in the forum they study, so it is likely that this feature, again, was borrowed from other websites and was only partially successful in being incorporated as a productive rule into the slang pupper talk on Reddit. The pronoun mismatch is also mentioned by Bivens (2018:14-17), whose examples (such as "him didn;t [sic] get treats when he wanted") seem to fit the one instance of irregular pronoun use well. Again, this feature is not as frequent on Reddit as it is in Biven's data from Facebook. The same seems to be the case for the two instances of the "can-has" formula, which is mentioned by Bury & Wojtaszek (2017:33-34) and by Gawne & Vaughan (2011:115) as one of the central features of LOLspeak. Gawne & Vaughan (2011:113,117) furthermore draw attention to irregular plural formations (such as "waters" and "earths") and omission of determiners (as in "Ceiling Cat rode invisible bike").

One last feature I would like to present poses challenges for linguistic categorisation and reconstruction. The pronoun-verb combinations presented below could be the result of either the omission of the inflectional ending, or the omission of both the verb *BE* and the present participle suffix. For example, "she construc" could be a variation on either "she constructs", "she constructed", or "she is constructing". More complex verb phrases as origins (such as "she will have constructed") seem implausible. Without further insight into how these phrases were composed one cannot completely account for their origin.

- *he snac, she construc, she destruc, she nom, he s a c r i f y, he s t r e t c h, HE FLUFF, he attac, he defen, he E X T E N, He S C E N T, He I N V E S T I G A T E, He S P L O S H*

In some of the above cases the word class of the second element is ambiguous. For example, "he snac" could mean both 'he is a snack' as well as 'he snacks'/'he is snacking'. These utterances could be an imitation of the two-word stage during first language acquisition. O'Grady & Cho (2001:346-347) explain how the juxtaposition of two lexemes at this stage, such as "Mummy busy" and "Mummy push", can express various semantic relations and often lacks explicit marking for syntactic categories. On the surface, there also appears to be a similarity to *BE*-deletion, which is a prominent feature of African American Vernacular English (e.g. Bender 2001). While there are no further apparent connections to this variety in the data

at hand, Callier (2016:245) mentions how racist language ideologies might play a role in the use of "DH-stopping" (rendering <the> as <da>) in LOLspeak.

In sum, this section has listed the irregular grammatical features associated with pupper talk. Most of the features are not productive and are probably best regarded as relatively fixed phrases. That grammatical features are less central than orthographic and lexical features was also emphasised by the participants of the pilot study (20):

> (20) I08: As for the spelling vs. grammar, I think it's strictly spelling. Aside from the occasional outliers, I don't really see people changing their sentence structure. It's pretty much just word substitution.

This section also emphasised how a number of features overlap with related online slang phenomena such as "Doggo speak" (Bivens 2018) and "LOLspeak" (Bury & Wojtaszek 2017) and were probably inspired by them. This again underlines how interesting it would be to trace the diachronic development of these features on various online platforms to observe their propagation, as was attempted by Golbeck & Buntain (2018) for Reddit and Twitter. However, due to the apparent divergence from other types of slang and the use of novel constructions, it is justified to regard pupper talk as a distinct variety.

5.1.4 Orthographic features

One last area in which pupper talk deviates from standard English is its orthography. We find instances of letters being omitted, letters being added, and letters being substituted. In addition, letters can be repeated, or blank spaces inserted, to provide emphasis. Some of these spelling changes appear to be so productive that one of the informants even referred to them as "pupper misspelling logic":

- Insertion of blank spaces: <the L E N G T H>, <so m a n y>, <he s t r e t c h>, <l o n g c o m m i t t e e>, <he E X T E N>, <Such l e m g t h>, <H O W D Y>, <He S C E N T>, <Ear F L O P P O S>, <she a n g e r e y>, <make her L O N G again>, <L E M G T H Y>, <L O N G and C U T E>, <l o r g e>
- Repetition of letters: <He looooonnnngggg>, <boyyy>, <cuteee>, <after soooooo l o n g>, <Loooong girl>, <Goooooooood boy>, <Let's goooooo>, <Elloooo>, <Awwwww>
- Consonant alterations:
  - Change to <c> or <cc>: <snac> ('snack'), <bac> ('back'), <attac>/<attacc> ('attack'), <fucc> ('fuck'), <licc> ('lick'), <hecc> ('heck'), <socc> ('sock'), <intac> ('intact'), <protec>/<protecc> ('protect')
  - Change to <m>: <lemgth> ('length'), <lomg> ('long'), <grampa> ('grandpa'), <lomgboy> ('longboy')

    - Change to <f> or <ff>: <everyfing> (‘everything’), <teef> (‘teeth’), <fanks> (‘thanks’), <ruff> (‘rough’), <tuff> (‘tough’)
    - Change to <b>: <deserbs> (‘deserves’), <gib> (‘give’)

- Vowel alterations:
    - Change to <o> or <0>: <lorge> (‘large’), <smol> (328.59 pmw, ‘small’), <smort>/<sm0rt> (‘smart’), <br0ve> (‘brave’), <bork> (82.15 pmw, ‘bark’), <chonky> (‘chunky’), <monch> (‘munch’), <gorl>/<g0rl> (‘girl’)
    - Change to <oo>: <coote> (‘cute’), <hooman> (345.02 pmw, ‘human’), <foock> (‘fuck’), <snoot> (262.87 pmw, ‘snout’)
    - Change to <e>: <slep> (‘sleep’), <bred> (‘bread’), <fren> (‘friend’), <ples> (‘please’), <ded> (‘dead’), <snek> (‘snake’)
    - Change to <i>: <boi> (1035.06 pmw, ‘boy’), <curli> (‘curly’)
    - Insertion of <e>: <bige> (‘big’), <girle> (‘girl’), <doge> (‘dog’), <boye> (722.90 pmw, ‘boy’)

- Others (selection): <chimken> (‘chicken’), <dawg> (‘dog’), <happ> (‘happy’), <gurl> (‘girl’), <bouy> (‘boy’), <souper> (‘super’), <henlo>/<hewwo> (‘hello’)

Non-standard orthography was already commented upon by previous studies investigating related slang phenomena. Bivens (2018:18-20) lists a number of “spelling transformations” occurring in the “Doggo Speak” that she observes on Facebook, but none of them occurs in the data at hand, except for the repetition of letters for emphasis and the <-cc> pattern. The same holds true for the “deviant spellings” observed in LOLspeak by Bury & Wojtaszek (2017): the only similarity is the replacement of <v> with <b>. In their description of the orthography of LOLspeak, Gawne & Vaughan (2011:109) mention one feature that also occurs in the data at hand: the replacement of <o> with <0>. According to them, this feature originated in “leet speak”, a variety associated with hackers; this shows how users draw on a variety of resources to create a new variety. As before, some of the features can be related to the speech during first language acquisition. The replacement of the interdental fricatives (represented through <th>) might be related to the fact that the interdental fricatives are among the last consonants that children acquire (O’Grady & Cho 2001:331).

Let us now come back to the question posed in section 2.4.1 above: where does the non-standard orthography used in pupper talk fit into the classification provided by Androutsopoulos (2000)? What can be ruled out at the beginning are regiolectal spellings, as there is no apparent link to any regional or national variety of English. Furthermore, prosodic spellings do appear in the data set (e.g. “You NEED to buy him a tuxedo” in comment c1436) but only repetition in relation to pets and the display of affection were considered a part of pupper talk. Interlingual spellings can also be discarded, as no loanwords appear as part of the slang. Some of the spelling

alteration that we find seem to fit into the category of "phonetic spellings", representing standard pronunciation that is not represented by the standard orthography (Androutsopoulos 2000:521). As examples one could cite the omission of <k> in <bac> or the omission of <a> in <bred>. Moving on, "colloquial spellings" representing colloquial speech also appear. Among those are <dawg>, <gurl> and potentially also the replacement of interdental fricatives as in <fanks> and <teef>. For many spellings, however, the classification remains ambiguous as it is unclear how (and if) they are realised in oral language. For example, the variant <boi> could be classified as a homophone spelling if the oral realisation was the same as for <boy> but would have to be termed a colloquial spelling in case the pronunciation differed (e.g. /boi:/). The same holds for <bige>, <girle>, <curli> and others. It therefore seems as if a complete classification of orthographic phenomena can only be achieved if oral data is collected in combination with written data. This was done by Miltner (2014), who conducted focus groups with users producing and consuming the slang LOLspeak online. By recording the discussions Miltner was able to observe how participants used the slang in their oral communication and how they realised the orthographic variants. Whenever possible, such a procedure seems to hold great advantages for the description of online slang.

5.2 Subreddits as communities?

In the theory section we saw how online communities are defined in a structural sense (section 2.3). Based on that discussion, especially on the criteria defined by Herring (2004), let us now assess to what extent subreddits can be classified as online communities. Afterwards, the relation to the concept "community of practice" and its applicability will be addressed, which provides an important background for the factors influencing the usage of the slang features.

5.2.1 Virtual communities

While examining the conditions for online communities, I will follow the suggestions to operationalise these criteria proposed by Herring (2004:15, emphasis original), which are given below. Each aspect will be examined in turn. First, let us look at "active, self-sustaining participation" and "core of regular participants":

1) *Participation* can be measured over time, and *core participants* identified on the basis of frequency of posting and rate of response received to messages posted […]

Due to the method of data collection, no adequate account of the posting and commenting frequency on the individual subreddits can be presented. What was measured, however, was the number of comments that each post in the data set received, as depicted in section 4.1.5 above. The number of comments is higher for subreddits with more subscribers and lower for

smaller subreddits. Also informative is the number of comments divided by their level, as shown in section 4.1.6: the larger communities not only have more comments per post, but also more active involvement through replying to comments on higher levels. The question of core participants also cannot be readily answered by the data collected for this study. However, as pointed out in the description of Reddit as a whole (section 2.2), only a very small percentage of Reddit users actively contribute content, while the vast majority consumes content only (Golbeck & Buntain 2014:616, Singer et al. 2014:520). Golbeck & Buntain (2014:619) claim that their data provides evidence that users do not normally contribute to more than one subreddit, which would strengthen the assumption that they represent distinct communities. But one has to take into account that they are researching comparatively large subreddits, on which enough content is posted that could theoretically occupy and entertain an individual user. The subreddits studied in this paper are in some cases rather small, with sometimes less than five new submissions within a week. This is hardly enough content for a user to not also consume content on other subreddits and pursue activity there. I therefore assume that it is common for Reddit users to be engaged in more than one subreddit. Investigating this first criterion would therefore suggest that larger subreddits are more community-like than smaller subreddits.

> 2) Shared *history* can be assessed through the availability and use of archives […]. *Culture* is indexed through the use of group-specific abbreviations, jargon and language routines […]. *Norms and values* are revealed through an examination of netiquette statements […] and verbal reactions to violations of appropriate conduct […].

Moving on to the next aspect, the subreddits provide no structured archive or account of their history, but older submissions can be searched for. Some of the moderators interviewed during the pilot study showed an awareness for the development of their subreddit over time. For example, informant 02 explains (21):

> (21) I02: What can happen is as a sub gets bigger, it loses it's uniqueness. I definitely remember /r/RarePuppers having a different feel to it a few years ago

While the moderators, who intentionally chose to take responsibility for a certain subreddit, might have such an awareness, it is unclear to which extent normal subscribers to the subreddit have a sense of its diachronic development. A common culture can be indexed by shared linguistic routines, under which we can include the slang investigated in this paper. Pupper talk is explicitly encouraged in some of the subreddits; this is most obvious in r/rarepuppers and r/longboyes, by stating it in their subreddit rules or by using the slang to compose the subreddit's description. As another form of shared culture one could also count the taxonomy of dogs enforced by the subreddit r/dogswithjobs: submissions are obliged to tag their post with

a flair, which makes the grouping of the post more easily accessible (the most popular flairs include “Service Dog” and “Police Dog”). Furthermore, the subreddit r/dogswithjobs encourages the use of specific tags marking users as experts, such as “Sheepdog Trainer” or “Livestock Guardian Owner”. For the topic of norms and values (which was already mentioned in the faceted classification above, see section 2.2.1) it seems sensible to make a distinction between implicit norms and explicit rules (Chandrasekharan et al. 2018:5). Every subreddit has its own rules which are depicted on the sidebar. As an illustration, Figure 20 presents the rules for posts and comments on the subreddit r/longboyes.

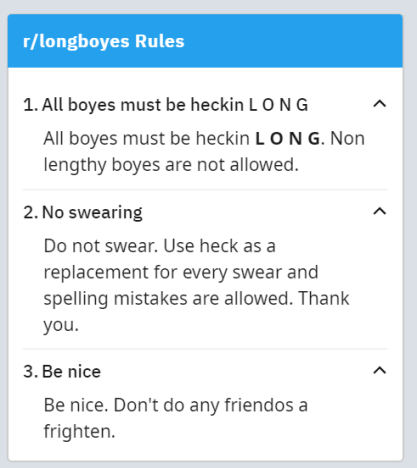

Figure 20: Example of explicit community rules, taken from r/longboyes.

This qualifies as the explicit rules of a subreddit. In addition to these, we can expect the presence of implicit norms of how communication is normally conducted. For example, from the r/longboyes rules one could infer that an implicit rule of this subreddit is that pupper talk and other humorous wordplay is highly encouraged. Looking at the users’ view as represented in the pilot study, informant 02 (22) made explicit reference to community norms (emphasis added). It is unclear whether he or she is hereby referring to the explicit subreddit rules or to the implicit norms.

> (22) I02: If I'm posting to /r/Aww I'll probably avoid using these words simply because some people will downvote you just for saying "doggo" in your post title. On the hand, if I'm posting to /r/RarePuppers, using doggo *is to be expected* and is *usually the norm*.

One also has to take into consideration that Reddit has general rules of conduct applying to the site as a whole (Chandrasekharan et al. 2018:2), known as the “Reddiquette” (Reddit Inc. 2020). For example, the user composing comment c0013 complains about the post being re-submitted by a person that is not the original author. The user hereby violates one of the rules in the Reddiquette: “Please don’t: Complain about reposts”. The user is penalised for his or her behaviour by being downvoted by the other users. Summing up, we saw that every subreddit

has its own explicit rules and allows access to earlier submissions, but a sense of history can only be assumed for the moderators and core, long-term members. Slang and group-specific taxonomies are employed by some subreddits, making them more prone to community-building.

> 3) *Solidarity* can be measured through the use of verbal humor […]; *support* through speech act analysis focusing, e.g., on acts of politeness […]; and *reciprocity* through analysis of turn initiation and response […].

Showing solidarity through humour is undoubtedly a salient feature in the subreddits investigated, which might be fostered by the informal tone of the platform as a whole. Larger subreddits, such as r/rarepuppers, feature more humour and wordplay than smaller subreddits due to the overall higher activity. Typical forms of humour practiced on the subreddits analysed are the imitation of the animal's thoughts in the situation depicted: (see comments c0092 or c0355), including pop-culture references to movies (c0232) or music (c0236), and playing with the words or word order of previous comments (c0304, c0309). As humour was not coded, no quantitative statement about its pervasiveness can be made. Support can be found through answering questions other users posed: frequently comments beneath a post ask about further information on the dog depicted, mostly its breed. Such questions are usually answered by fellow users or the person who originally contributed the picture or video. As can be seen in quotes (23) to (25), even after an answer is given to the question, a further comment with more specific information was inserted by another user.

> (23) c0572: What breed of dog was he? That face has a lot of personality!
>
> (24) c0574: Pibble
>
> (25) c0575: Specifically, looks like an American bully/american Staffordshire bull terrier mix of some sort.

As another form of support one could also count the inclusion of links to other, topically related subreddits within one's comment, as this provides an opportunity for fellow users to easily access similar content. As was shown in section 4.1.4, links are included in around 5 percent of all comments, with little difference between the subreddits. To provide an example, post p005 on r/aww features a video of a dog meeting young kittens. Within his or her comment (26), a user included a link to a subreddit devoted entirely to kittens.

> (26) c0100: The dog looks like a friend to kitties. Also, adorable r/pointytailedkittens!

Reciprocity does not seem to apply to the comment sections under investigation, since they only seldom feature lengthy conversations between two users. Summing up, solidarity and

support is a frequent feature found in all the subreddits analysed but both are naturally more frequent in subreddits that feature more participation.

4) *Criticism* and *conflict* can be analyzed through speech acts violating positive politeness […]. *Conflict resolution* might usefully be considered as an interactive sequence of acts […]

The most salient form of conflict on the subreddits is probably embodied by the behaviour known as "trolling" (Bergstrom 2011:1). It does not show up systematically in the data, as the comments in question are often deleted after a short while. But some instances of trolling could still be recorded. For example, in the utterances (27-29) below, a user is called out as a troll, and responds with another condescending comment.

(27) c0850: Disgusting.

(28) c0859: Yes. Useless trolls like you always are.

(29) c0860: What a democratic and low effort comment. I hope you felt a little more powerful today.

For further examples see comments c0547 or c0724. In such cases, conflict resolution is mainly performed via intervention by the moderators, deleting the offensive comments and frequently also banning the author from further contribution on the subreddit. But there are also other minor instances of conflict recorded in the data: in most cases these concern comments that are felt not to match the light-hearted tone of the platform. Disapproval of such remarks is mainly visible through a low comment karma score. To provide an example, post p084 depicts a dog breed that is known to have health problems. As a user points this out in c1014, he or she is downvoted by the other users and is advised by the author of c1027 to "keep it adorable" (30-31):

(30) c1014: I hope she doesn't have/get brain damage. Cavaliers are a breed that suffers a huge number of potential problems, including a brain too big for their skull resulting in brain damage for a large number of dogs :(

(31) c1027: You have a valid point, however this is r/aww and its definitely not the platform for that discussion. Keep it adorable, yo.

In sum, criticism and conflict can be found on almost all subreddits studied. Similar to the humour and support discussed before, they occur more often in the larger subreddits compared to the smaller ones due to the higher overall traffic.

5) A group's *self-awareness* can be manifested in its member's references to the group as a group, and in 'us vs. them" [sic] language […].

Also important for defining a virtual community is the users' self-awareness as a distinct group. In the pilot study, some informants referred to their subreddit as a community, for example informant 08 (32, emphasis added):

> (32) I08: "In the /r/longboyes *community* (and really *any dog community on Reddit*) this is implicit, so it's not as obvious what's going on when people use it elsewhere on the internet.".

This tendency might be influenced by the terminology proposed by Reddit, as the frontpage of every subreddit features a section entitled "About community" and subreddits are also referred to as "communities" elsewhere on the site, e.g. in the privacy policy (Reddit 2018). Apart from this terminology, moderators also referred to a distinct sense of identity for certain subreddits, as informant 02 describes (33, emphasis added):

> (33) I02: /r/RarePuppers is the sub where pupper speak really got big. It was probably in use around the internet and reddit before that sub, but when you think about "doggo" and "pupper" on reddit you'll probably think of /r/RarePuppers. One issue that the sub was having lately is that the mods, and some users, felt *the sub was losing it's identity* and was basically becoming /r/aww 2.0. They made some mod posts setting out new rules to try to *get the sub back to its roots*.

Here the "identity" of r/rarepuppers is explicitly defined as intentionally different from the bigger, more popular subreddit r/aww. Apart from these two utterances by the informants, we can also find further hints on the subreddits themselves. A clear indication that r/rarepuppers has self-awareness as a distinct community is provided by two ongoing events at the time of writing: a subreddit-internal contest awarding prizes in a variety of categories, such as for the user who contributed the best posts[6]; and a sale of limited-edition merchandise for the subreddit as a celebration of the subreddit reaching two million subscribers[7]. Self-awareness can, however, also be made explicit through the subreddit rules: while elaborating on the rules for r/dogswithjobs, the moderators explicitly delineate their platform from r/aww, in an attempt to position themselves as a more serious platform (34):

> (34) "For the purposes of this sub, your pet dog who also guards your house is not considered a dog with a job. This rule is mostly to prevent a surplus of posts of dogs looking out a window or door. […] No offense, but your Chihuahua sitting by the window is better for /r/Aww".[8]

For this criterion, we can therefore conclude that not for all subreddits we can find evidence that users perceive them as a distinct community. The evidence found, including self-

[6] https://www.reddit.com/r/rarepuppers/comments/fmhc0p/rarepuppers_best_of_the_year_awards/ (last accessed: 07.06.2020).
[7] https://www.reddit.com/r/rarepuppers/comments/f24xcj/2_million_subscribers_fundraiser_and_limited/ (last accessed: 07.06.2020).
[8] https://www.reddit.com/r/dogswithjobs/wiki/rules (last accessed 07.06.2020).

description by moderators during the pilot study and within the subreddit rules, supports the community status of r/dogswithjobs, r/longboyes, and especially r/rarepuppers.

> 6) Evidence of *roles* and *hierarchy* can be adduced through participation patterns […] and speech act analysis […]. The study of *governance* and *ritual* would appear to require an ethnographic approach […].

The last relevant aspect concerns roles, hierarchies, and governance. Through Reddit's inbuilt function of moderators there is a clear structural distinction between users who have power to control content and textual submissions, and users who do not have these rights. This applies to all the investigated subreddits equally, despite the number of moderators varying. During the ethnographic pilot study, it became apparent that these moderators frequently made use of their power and removed utterances that did not comply with the rules of the subreddit. While no quantitative statement about the frequency of moderation can be made, one subreddit nevertheless stood out through a higher sense of regulation: on r/dogswithjobs (which was noted earlier as attempting a position as a more serious platform), a comment reminding users of the subreddit rules is automatically posted as the first comment beneath each post. So, while moderation is a ubiquitous phenomenon, it is perceivable to a higher extent on r/dogswithjobs.

Summing up the six aspects investigated, we saw that the features proposed by Herring (2004) apply more to some subreddits than to others. While all subreddits have the structural prerequisites to develop a virtual community, such as explicit rules, access to previous content, and moderators, some subreddits use additional tools to foster a sense of community. Especially relevant seems the slang pupper talk, which is employed especially by both r/longboyes and r/rarepuppers to create an in-group feeling. Other community-creating mechanisms include the 'profession-taxonomy' developed by r/dogswithjobs, or community-internal events on r/rarepuppers. It therefore seems reasonable to assume a scale of subreddits that are least like a virtual community (r/goodboys) and subreddits that exhibit many properties of virtual communities (r/rarepuppers and r/longboyes). It is also highly likely that the degree to which a subreddit qualifies as a virtual community differs from user to user. For core members, such as the moderators, the sense of community might be stronger through frequent interaction and close acquaintance with regular contributors. On the other hand, peripheral members, which might only occasionally click on a post they saw on the main Reddit frontpage, might not perceive a subreddit as a coherent group altogether. In a similar vein, Honkanen (in press:72) concludes concerning the forum she studies: "One could argue that NL [Nairaland] might be a community for some members though definitely not for all". At this point I would like to mention again that the data set used in this study was not collected to answer the question

whether subreddits can be classified as online communities. It is therefore possible that other studies using a different approach arrive at more detailed results.

The finding that some subreddits do not fulfil all the criteria for virtual communities has remedies for other studies using this concept. It is central that the criteria are investigated thoroughly before the term virtual community is applied. For example, in their study on norm and rules on Reddit, Chandrasekharan et al. (2018:2) start off with the assumption that each subreddit is a distinct online community. Danescu-Niculescu-Mizil et al. (2013) proceed in a similar fashion in their investigation of users' response to linguistic innovation during various stages of their user lifecycle. In both cases, a further level of detail could have been added to the study if the community status of the platforms investigated had been taken into account. Are rules enforced differently on platforms that are more community-like than others? Does a user's response to linguistic innovation also depend on the community status of the platform he or she is participating in?

A quick concluding note on the methodology employed seems in order here. Many of the features listed by Herring (2004:15) seem to be correlated with the size of the group under investigation: the higher the number of people involved and the higher the interaction on a given platform, the more instances of humour, reciprocity, and conflict can obviously be found. However, as group size increases this does not necessarily imply that a group is also more community-like. It therefore seems appropriate for future studies to measure these characteristics in relation to the overall activity of a certain platform.

5.2.2 Communities of practice

Having established that subreddits can, to a certain extent, be classified as online communities, we will now investigate whether the concept of community of practice can be applied to them. The three criteria used in order to make this decision are adapted from Meyerhoff (2004:527-528):

> First, there must be *mutual engagement* of the members. That is, the members of a CofP [community of practice] need to get together in order to engage in their shared practices. (Meyerhoff 2004:527, emphasis original)

This first criterion already proves to be problematic when researching an asynchronous online platform such as Reddit. Compared to other CMC platforms, users do not need to be present at the same time in order to consume content and react to other users' comments (Boland & Locher 2014:15, Herring 2007:13). The underlying question therefore seems to be: does the mutual engagement of the members have to take place in real time? I would argue that communities of

practice can also function if communication takes place asynchronously. Johnson (2001:53-54) discusses this question in the context of corporate learning and argues that communities of practice can indeed also be formed in asynchronous, text-based online environments. As an example of this from the discipline of linguistics, one could cite Stommel (2008), who analyses an asynchronous German forum using the community of practice approach.

Defining communities of practice through their mutual engagement clearly excludes the users known as "lurkers" (Preece et al. 2004), who only consume content without pursuing any visible activity on the platform. The more difficult question is whether people whose only activity is voting on a post or comment should be included, or whether mutual engagement only covers the practice of contributing actual text in the form of posts or comments. Since the informants of the pilot study frequently made reference to voting as a relevant behaviour, I propose to include both voting and commenting/posting under the term "engagement". Looking at the subreddits at hand, there is little difference concerning this first criterion: all platforms feature asynchronic communication and have the same forms of activities (posting, voting, and commenting) in which active members are engaged. The extent to which these practices are performed varies between subreddits and users. Larger subreddits feature a higher proportion of all three activities compared to smaller subreddits. Core users might post, comment, and vote on a regular basis, while less central members might only comment and vote, and peripheral members might only vote on other's contributions.

> The second criterion for a CofP is that members share some *jointly negotiated enterprise*. […] It is the pursuit of this enterprise that creates relationships of mutual accountability among the participants (Meyerhoff 2006:528, emphasis original)

While for the whole group of animal-centred subreddits, the main purpose could be summarised as "sharing (and, in most cases, discussing) pictures or videos of animals", every subreddit has its own, specific enterprise. For example, while the purpose of r/goodboys can be summarised as "sharing personal pictures of one's own dog", r/dogswithjobs has a quite different focus: "sharing and discussing pictures and videos of working dogs". These enterprises are "jointly negotiated" in the sense that each team of moderators can alter the community purpose and intentionally steer a subreddit in a certain direction. Such re-orientations are informed by discussions with and among the users of the subreddit (for an example, see the announcements on new subreddit rules for r/rarepuppers[9]). The question of accountability is a more difficult one, as the online environment is characterised by a lack of "long-term commitment"

[9] https://www.reddit.com/r/rarepuppers/comments/ao0si9/official_new_rulesies_and_a_new_banner/ (last accessed 07.06.2020).

(Androutsopoulos 2004:422). Even though there are community rules and members are expected to follow them, rule violations are not always called out, and if so, the person violating the rules can easily avoid further retribution by deleting the offending comment or the whole user account (which is a frequent phenomenon). Also, regular participation is neither enforced nor expected. But, since many online platforms (apart from identity-focused sites such as Twitter or Facebook), are characterised by this low level of accountability, I would not take it as a counterargument rendering the formation of online communities of practice impossible. One could therefore conclude that all subreddits in question fulfil the criterion of having a jointly negotiated enterprise.

> Third, a CofP is characterized by the members' *shared repertoire*. These resources (linguistic or otherwise) are the cumulative result of internal negotiations. (Meyerhoff 2006:528, emphasis original)

The communication observed on the subreddits draws on a variety of resources, some of which are subreddit-specific, some are shared by a group of subreddits, some are shared by the platform Reddit as a whole, and some are drawn from more wide-ranging cultural phenomena. As a practice that is performed on one subreddit only, one could name the taxonomy of dog-professions developed by r/dogswithjobs, which was already mentioned above. Then there are some resources that the whole group of animal-centred subreddits draw upon, but that are not used by other subreddits. As an example, one could cite the "dog tax" ritual: when a user talks about his or her own dog within the comment section of a post, other users might respond with a request for dog tax (see examples 35-36), which means requesting a picture or video of the user's dog. This request can be responded to with a link to the visual material. Within the data we also find instances of users anticipating these requests and including links to pictures within their comment. This ritual was recorded once on r/aww, twice on r/dogswithjobs, and three times on r/rarepuppers.

(35) c4063: I never liked sharpei then ended up with a sharpei beagle mix. Best Dog Ever!!

(36) c4073: Dog tax

Puppertalk as a slang is a resource that users draw on to create comments on every one of the subreddits investigated (which was partially the reason for their selection). What distinguishes them is the frequency and importance of the slang (see section 4.1.1). Within the slang we can find no further noticeable diversification: even though certain items are preferred on some subreddits (such as the spelling <boye> on r/longboyes, the blend *barkour* on r/barkour, and the term *pupper* on r/rarepuppers), there are no features that are used exclusively on one subreddit. In addition, there are also Reddit-wide routines and repertoires that users draw upon.

This includes, for example, the uppercase-lowercase-orthography (as seen in comments c0631, c4107, or in post p117), or congratulating other users on their “cake day” (the anniversary of them creating their Reddit user account, see comments c1975-c1977, and c2002-c2003). Moving beyond the platform, users also make a variety of references to pop culture by quoting songs (Queen’s “Bohemian Rhapsody” in c1775, c1780 to c1781), movies (“Fight Club” in c2482-c2486, “Star Wars” in c0112), or TV series (“The Office” in c1322 to c1343). These references are in many cases taken up and elaborated upon by other members but might also go unnoticed if no other user shares the necessary background knowledge.

In conclusion, besides the shared repertoires and similarity of the engagement across the subreddits investigated, it seems justified to classify each subreddit as a community of practice on its own, as they each have a unique enterprise. The status as a community of practice is especially salient for those subreddits that have specialised routines (such as r/dogswithjobs) or that heavily draw on the pupper talk slang to encourage the creation of a distinct community identity. However, the same words of caution are in order here as were mentioned concerning virtual communities: core members are more likely to perceive the subreddits as distinct communities of practice due to their higher exposure to the activities and the communication taking place on the different subreddits. This is illustrated nicely by the following explanation by informant 02 (37), who is frequent poster and who shows great awareness for the distinct community practices of different subreddits:

> (37) I02: Actually, here's a good example of deciding what words to use.
> I posted this to /r/DogsWithJobs: <link to post>
> "Doggo" in the title. I figure most people in DogsWithJobs wouldn't care and I liked the way sky doggo sounded.
> I also crossposted that post to /r/MilitaryGfys using the same title: <link to post>
> I specifically remember thinking I should change the title to "sky dog" for the post to /r/MilitaryGfys. The logic being "This is a sub for a bunch of dudes, probably some military dudes, and I'll get downvoted for saying doggo because they'll see it as baby talk or whatever."

This study therefore emphasises the importance of slang as a tool to create communities within the anonymous online environment. As Seargeant & Tagg (2014:10) emphasise: “shared language practices are an important part of the broader range of shared social practices which comprise group membership.”

5.3 Factors influencing the usage of pupper talk

After describing the linguistic profile of the slang pupper talk and assessing the community status of the subreddits in question, let us now return to the main research question: Which

factors influence the presence of the slang within the comments? The regression model reveals that the level of the comment as well as the presence of pupper talk in the previous textual unit are highly significant predictors. Let us look at each of these in turn.

Compared to the comments on the first level, which respond directly to the post, comments on both level two and level three are less likely to contain pupper talk. This tendency is probably linked to the topic of the comments. While comments on the first level are in most cases topically related to the visual mode, comments on the other two levels are to a certain extent topically more removed from the original stimulus. Herring (1999:6-7) suggests that this "topic decay" is characteristic of computer-mediated communication in general. An illustration of this phenomenon is provided in examples (38-42, emphasis added) below. Beneath post p075 we can observe how the conversation topically drifts away from the original post: the new topic, in this case address lines, is not discussed using the slang:

(38) p075: Was annoyed about the sheep blocking the road because I'm doing a outta state job for Optus and I'd like to be back home ASAP. But then I saw 3 *doggos* hard at work and suddenly I didn't care how long they took! Brought a smile to my face watching them working hard! (*Western Australia, Australia*)

(39) c0877: I would have had no idea without the second Australia /s

(40) c0884: I cringe when I get to the "New York, New York" part of my address when I'm writing it down… so I felt this even though it was /s lmao

(41) c0885: Haha, we also have South Australia as a state name, and Australian Capital Territory as the name of the 'state' (territory) where Canberra our capital city is located. / When you 5 states and two territories, in the main land mass you don't need creative names.

(42) c0886: My favorite is: West New York, New Jersey

As we can see in the examples above, the topically removed discussion might still be joking in nature but does not draw on the pupper talk repertoire. How can this apparent restriction to the topic of animals be explained? That slang typically only covers certain semantic fields is not surprising, given that it "generally originates within small self-defined communities of practice or communities of circumstances […] where it is used to rename aspects of shared experience and environment" (Malmkjær 2010:489). In many cases of offline slang, these shared experiences resolve around drug consumption or sexuality (see Mattiello 2005). Since for the communities in question the shared experiences centre around animals, the lexical inventory also mainly covers the semantic field of animals and endearment. What is noteworthy, however, is that pupper talk (compared to other forms of slang) is not only lexical, but also incorporates non-standard grammar and orthography (see sections 5.1.3 and 5.1.4). Even though the lexical inventory is restricted to a certain semantic field, the grammatical constructions and spelling

transformations could theoretically be employed to talk about different topics as well. However, this does not seem to be a frequent phenomenon within the data collected. We can conclude that non-standard grammatical constructions, which are not semantically specified, are nevertheless felt to be only appropriate when discussing the same semantic field as covered by the lexical inventory of the slang. This restriction therefore provides evidence for the importance of topic as a non-audience factor influencing stylistic variation. Bell (1984) originally claimed that non-audience factors were less important than audience factors. This also seems to be the case here, as the influence of the level as a predictor is smaller compared to the influence of the previous textual unit.

Another factor that is highly significant is the presence of pupper talk in the previous textual unit. This predictor relates back to the discussion of audience roles in section 2.6. Let us first look at an example of the presence of pupper talk within the initial textual unit as well as the following units (43-46, emphasis added). Responding to post p001, the very first comment c0001 choses double marking for the comparative in the form *bestest*. Responding to comment c0001 on the second level, both comments c0006 and c0007 take up the irregular morphology. This even continues to the third level: comment c0009, responding to c0007, also employs the form used above.

(43) c0001: This might qualify as the *bestest* boyyy ever…

(44) c0006: *Betterest* Boy!!!

(45) c0007: Everyone thinks they have the *bestest* boy ever; and everyone is right

(46) c0009: And if they don't they have the *bestest* girl!

This strong tendency to go along with the style used by the formal addressee seems to support Bell's (1984:160) claim that the immediate addressee exerts the biggest influence upon the stylistic decisions of a speaker. That the direct addressee has considerable impact on stylistic choices online was already demonstrated by Pavalanathan & Eisenstein (2015) and Shoemark et al. (2017) for the platform Twitter, and by Hinrichs (2016) for Facebook.

Let us now inspect the comments made by the informants in the pilot study on what affects their stylistic choices. Two participants explicitly mention the style of the previous unit as relevant for their own composition, directly supporting the statistical significance of this predictor (47-48, emphasis added):

(47) I01: Defiantly! a lot of the time on the Internet, if one person uses a particular type of spelling, or a meme, or anything at all, really, *usually people will follow through*. this can be seen more literally on r/me_irl sometimes, if someone posts a comment of a particularly sad, disappointing or otherwise of a negative event, someone is bound to write the word "F" as a reply, thus starting a *chain of comments* saying the letter "F". i think that is a good example of how *typing something in one style would lead to others writing in a similar caliber.*

(48) R: Could the choice also be influenced by the type of post or by which spelling was used in previous comments?
I08: Oh yeah definitely. Sometimes I see *comment chains* where it's just people coming up with increasingly goofy ways of spelling things, lol.

However, we also find other factors mentioned. Two informants also refer to what could be summarised under the term "active audience" defined above (49-50, emphasis added):

(49) I07: It's sort of like code switching in diverse communities. *If you are among people you believe will react better to puppy baby talk you use it.* No one has long conversations in puppy talk, though.

(50) I02: I'd say it often depends on what *subreddit* I'm in. I'll be more likely to say doggo, pupper, boye/boi, etc in subs like /r/RarePuppers or /r/Longboyes then in /r/Aww or /r/DogsWithJobs.

These comments point to a more complex answer than simply the style used by the formal addressee. I would like to put forward the theory that the significance of the predictor "presence of pupper talk in previous textual unit" does not only indicate the relevance of the formal addressee but could also show the importance of the active audience. This proposition is based on the participation structure and the technical environment of Reddit: we already established that the majority of Reddit users mainly performs up- and down-voting and that receiving upvotes is the desired outcome for many people that contribute comments (see section 2.2.1). Since the pupper talk slang is not valued by every Reddit user, I assume that commenters would only use it if they expect the people reading and voting on their comment to appreciate the style. This is obviously the case on subreddits that explicitly encourage using the slang. But, following this argumentation, how could the usage of the slang on more adverse subreddits (such as r/aww) be explained? I propose that users enjoying the slang and intending to contribute on such a subreddit have only a fraction of the overall people consuming the comments in mind. This has to do with the technical prerequisites of the platform: after a certain number of comments has accumulated beneath a post, Reddit will no longer display all of them, but collapse the replies from the second level onwards beneath the initial comment. In order to read these responses, users need to actively click on the option "XY more replies" (as shown in Figure 21).

Figure 21: Illustration of option to expand a comment tree (showing c2426).

My hypothesis is that only users who enjoy the initial comment and its stylistic choices will expand the comment tree to also see the responding comments, and that contributors are aware of this tendency. The assumption that only people enjoying an initial comment composed in pupper talk will read on to the other comments, while people disapproving of the slang will already jump to the next initial comment, might minimise the risk for commenters to use the slang themselves. Therefore, the main influence on a commenter's stylistic decision would be the active audience, which is indicated by the stylistic behaviour of the formal addressee. One could compare this to the following situation within oral communication: if speaker A produces an utterance (formally addressed to speaker B) within a group setting that elicits appreciation and laughter, speaker B might be inclined to take up the style of A's utterance to prompt the same reaction. So, while formally there is great similarity between the utterances of A and B, the driving force behind the stylistic decision is the audience. It needs to be emphasised that this proposed thought process and risk management is my personal hypothesis. It is not directly apparent from the quantitative results, but it is supported by several comments made during the pilot study. Informant 02 in particular mentions the awareness of the audience and the risk of using the slang (51-52, emphasis added):

> (51) I02: I think whether you say "dog" or "doggo" you're talking about the same thing, but they *will be received in very different ways*. If I'm posting to /r/Aww I'll probably *avoid using these words simply because some people will downvote you* just for saying "doggo" in your post title. […]
> Most people use the "pupper/doggo" words in a "memey", goofy kind of way that isn't meant to be taken too seriously. *Then there's some people who get angry when they see "doggo" or "pupper".* In fact there's a whole sub dedicated to it: /r/DoggoHate lmao. It's a pretty active sub.
>
> (52) I02: Actually, here's a good example of deciding what words to use.
> I posted this to /r/DogsWithJobs: <link to post>
> "Doggo" in the title. I figure most people in DogsWithJobs wouldn't care and I liked the way sky doggo sounded.
> I also crossposted that post to /r/MilitaryGfys using the same title: <link to post>

> I specifically remember thinking I should change the title to "sky dog" for the post to /r/MilitaryGfys. The logic being "This is a sub for a bunch of dudes, probably some military dudes, and *I'll get downvoted for saying doggo because they'll see it as baby talk or whatever."* Well, it ended up being fine. […]
> So I guess the conclusion is that most people really could care less whether you say dog or doggo. *One worry is that the minority of people who despise the words will downvote you just for using them.* Sometimes it only takes a handful of downvotes to kill off your posts visibility. Especially if those downvotes come in the few minutes after you post. So you *ask yourself if it's worth using the pupper speak words and potentially getting downvoted for it.*

This theory of the hidden relevance of the active audience, however, raises a number of questions. Initially, active audience was thought to be represented by the predictor subreddit. If the active audience really is that important, then why did the different subreddits not reach significance within the regression model? Potentially because the active audience on each subreddit is not homogenous, as explained above. Each subreddit is browsed by both people that appreciate the slang and by people who do not. Even on the subreddits that endorse pupper talk, such as r/longboyes, only 64% of the posts and 29% of the comments collected contained it (see section 4.1.1). In the end, this is a question of operationalisation of the proposed audience roles and how their influence can be measured reliably in non-experimental settings.

The importance of the use of pupper talk in the preceding textual unit also points towards another interesting aspect: the responsive, interactional nature of humour in the online environment (Baym 1995:16, Chovanec & Tsakona 2018:8). North (2007:547) explains how humour online is often a "joint construction": "Successful humour builds on previous contributions, and is recognized and often elaborated by other participants, thus playing a role in developing a sense of social cohesion". This becomes very apparent when taking a closer look at the data, for example the following comments responding to a picture of a corgi standing on a man's back (53-56, emphasis added):

(53) c3520: *Corgopractor*

(54) c3531: Not real *dogtors* though!

(55) c3533: *Therapawtic*

(56) c3534: I'm glad there's *recognition* of their massages

We can see how the comments beneath c3520 take up the theme of blending medical terminology with dog-related terminology to continue the joke. The users thereby position themselves as competent users of the slang and strengthen their ties to those users who also appreciate and understand the style used here.

Let us now take a look at the remaining predictors in the regression model. One of the goals of this study was to take into account the multimodal environment of Reddit and therefore pay due attention to Bell's (1984:178) non-audience factors. However, none of the predictors related to the visual material of the post (visibility of the face, mode, anthropoidness, age) had a significant impact on the probability of pupper talk being used within the responding comments. This result therefore provides further evidence for Bell's (1984:178) claim that audience factors play a more important role compared to non-audience factors. This was already confirmed by other studies for the offline environment (see Schilling-Estes 2006:385), but never for the online environment. However, this results also contradicts some of the statements made by informants during the pilot study (57-59):

> (57) I01: Well, it differs from person to person, the type and content of a post, and the overall mood of a person.
>
> (58) I01: The cuter the post (in one's opinion, obviously) and the happier the story behind it, the more likely it is the spelling would be there. If the story behind the post is sad (Ex: this is my dog, died yesterday, etc.) The likelyhood [sic] of this spelling appearing is lower.
>
> (59) I07: Type of post - anything cute really.

There are two possible explanations for this discrepancy between the participants' opinion and the outcome of the regression model. Either their subjective view of their linguistic choices is not in line with their actual language use, or the relevance of the multimodal environment was not adequately captured by the predictors chosen in this study. The second option is again a question of operationalisation: can the predictors adequately measure the influence of the visual mode and its 'cuteness'? Several alternative methods present themselves for capturing this property of the visual modes. Cuteness could have been measured by investigating the textual reactions to the post, e.g. how often words like *cute* or *adorable* appear within the comment section. To do this, a complete set of all the responding comments would have been necessary. Even more appropriate seems the option to ask community members to rate pictures and videos by their cuteness. While this would lead to a more fine-grained measurement, it would also require considerable additional effort.

We have now discussed the three predictors that reached significance in the regression model as well as their theoretical implications. However, it is worth keeping in mind that the model was only able to explain a small proportion of the overall variation found. This raises the question what other factors, which were not considered here, could help to explain the remaining variation. One aspect that comes to mind is inter-speaker variation: it seems plausible that some users are generally more prone to use the slang than others. Within the pilot study,

informant 06 admits to disliking the slang, while informant 05 points out users that are especially fluent in pupper talk. A fitting illustration of this is a particular user known in the community for his or her pupper talk poems, who was also pointed out by the informants. Several of these poems were sampled in the corpus, as the example (60) below illustrates:

> (60) c2480: we golden pups - we like to hide / so humans can't see us inside! / we undercover puppers who / know what the other's thinking, too / we secret club! we plot n plan / (the other dogs don't understand…) / n tho we're pups, not very old / we share our 'secret' / pot of gold! / [heart]

These user preferences could be taken into account by fitting a mixed-effects model (Levshina 2015:192-196) instead of a regression model, which provides an interesting avenue for further research. Previous researchers have also attempted to apply socio-demographic categories that are known to influence variation to the offline environment: there are several studies that incorporate age and gender as factors influencing language variation online (e.g. Finlay 2014, Flesch 2018). Retrieving this information is, however, very time-consuming, and its accuracy cannot be guaranteed. As an alternative, researchers have also considered environment-specific categories, such as participant roles (Androutsopoulos 2013:245), social roles (Golbeck & Buntain 2014:616), or user age, meaning the time an individual has been a member of a certain platform (Flesch 2018, Danescu-Niculescu-Mizil et al. 2013). Retrieving this information would have been possible, but equally laborious. Another potential explanatory factor could be the speech act performed within a textual unit. While exploring register variation between different subreddits, Liimatta (2016:63) comes to the conclusion that the "function of a comment (e.g. a joke, informational content, factual statements, personal experience)" might play a larger role than the subreddit a text is taken from. This seems plausible when looking at the data at hand, as a number of functions are performed repeatedly. These include describing one's personal reaction to the visual material (example 61), evaluating the visual material (62), imitating the animal's thoughts (63), narrating personal experiences related to the topic (64), asking for information related to the visual material (65), or word play and joking (66). Since answering this question would require a detailed speech act analysis, it seems more suitable for future qualitative studies.

> (61) c0204: I laughed way too hard at this [smiley]
>
> (62) c0163: That's so precious
>
> (63) c0165: "Don't talk about it human"
>
> (64) c0271: I had a german shepherd named Koda. Its a solid name.
>
> (65) c0221: What's she doing with them, what's the story here?

(66) c3622: I'll take 2 Cerberuses…. Cerberi… / I'll have 2 please.

Summing up, this study provides evidence of the importance of the formal addressee for the stylistic decisions made by users on the online platform Reddit. The potential indirect relevance of active audience is also discussed but cannot be proven with certainty. While explaining the influence of the level on the occurrence of the slang, the semantic restriction of both the lexical and grammatical features of the slang are elaborated upon, and the importance of topic as a non-audience factor is discussed. Taking together the influence of topic and the multimodal prompts, the study provides proof that in the online environment non-audience factors exert less influence on stylistic variation compared to audience factors. As other potential explanatory variables inter-speaker variation, speech act, and participant status are proposed.

5.4 Challenges for quantitative studies of stylistic variation online

While conducting this study, several methodological challenges came to light that seem to not have been addressed at length by previous studies. I would therefore like to quickly discuss some of the issues in the hope that future researchers can take them into account when conducting quantitative research on stylistic variation in the online environment. One fundamental question regards the operationalisation of stylistic variation. For the phenomenon at hand, usage of slang, two possibilities present themselves. One could either choose a binary measurement (slang being used – slang not being used), or a numeric measurement, describing the amount of slang being used within each linguistic unit (in this case: within one comment), which is realised in the form of the Doggo-score, a ratio between 0 and 1. The second option intuitively appears to be the more appropriate one, as the amount of slang used within one's utterance has an impact on the communicative effect and the overall tone. However, even though a numeric measurement might seem more accurate, it holds some difficulties, which are related to the sampling method and to the text length.

First of all, since the data set is sampled by time and not by phenomenon (see Androutsopoulos 2013:238), 83% of all comments do not contain any slang and receive a Doggo-score of zero. Therefore, when using a numeric measurement for the presence of slang, the disproportionately high number of data points with a value of zero makes fitting a linear regression model nearly impossible. A second problem is related to the average text length of the linguistic units under investigation. In the subreddits at hand, the comments are overall very short: on average, posts contained 9.65 words and comments 12.42 words (see section 4.1.2). This overall shortness leads to an unnatural distribution of values for the Doggo-score, which is measured as the number of words containing the slang divided by the total number of words. For example,

values such as 0.5 are disproportionally favoured. This problem is also discussed by Liimatta (2016:25), who solves it by only choosing Reddit posts with a minimum length of 400 words. While this leads to more normal relative frequencies, this cut-off point also produces a data set that is no longer an accurate representation of the actual language use on the platform under investigation. Both problems are illustrated well by the graph below (Figure 22), which plots the Doggo-score in the previous unit against the Doggo-score in the current unit. The line shows how the regression model tries to fit the data but is unsuccessful as there is no linear relationship between the two variables.

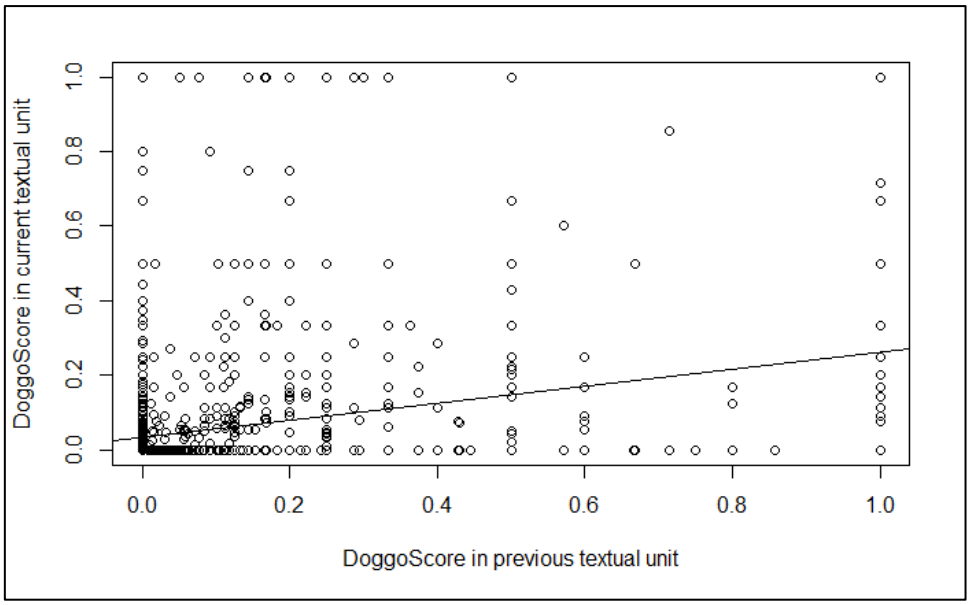

Figure 22: Illustration of challenge posed by numeric measurement.

So, in conclusion, numeric measurements of stylistic phenomena need to be handled with caution if one intends to fit a regression model to the data. Sampling by phenomenon can be chosen to obtain more suitable data for linear regression, but that way no accurate picture of the overall language use can be collected. Furthermore, text length should be inspected in advance to see whether the average word count is sufficient to not skew the relative frequencies. Otherwise, a binary measurement can be chosen to fit logistic regression instead, as was done in the study at hand. This naturally implies a loss of information and is especially problematic for samples with a large variance in text length. Since the range of the word count in the data set at hand is limited (see section 4.1.2), this is not a large drawback for the present study. For other studies using such a binary measurement for measuring variation, see Callier (2016:247-248), or Clarke & Grieve (2017).

A further aspect worth discussing is related to the identification of stylistic phenomena by the researcher. On a globally popular platform such as Reddit, the variety of resources from various

cultural backgrounds that users draw upon within their utterances (see 5.2.2) proves to be quite a challenge. Not every intended (humorous) reference will be identified as such, and not every speech act will be understood as intended. With increasing global connectivity, finding such a wide range of repertoires and resources within a single platform will in the long run develop to be the norm instead of an exception. This challenge could be addressed by working closer with members of the subreddits in question and inspecting textual contributions together with them. Within qualitative, detailed studies this is a feasible solution, but it is not an option for quantitative studies dealing with large numbers of utterances. However, even via close interaction with group members not every utterance will be able to be interpreted in the intended way. The data shows that users are also aware of this multitude of resources. For example, the author of comment c0450 explicitly prides himself or herself in understanding the reference to a video made in the previous comment (67-68).

(67) c0449: Bruh on the bottom is like "Yeah" "Yeah""The maple kind?" / "Yeah"…."What was in there" "Yeah" "Yeah" / [smiley] / [link]

(68) c0450: I understood that reference / Clark g the talking dog is fun

This phenomenon appears partially related to what Marwick & boyd (2011) define as "context collapse". Androutsopoulos, for example, shows how Facebook users navigate that the platform collapses a number of previous social contexts into one context only. In contrast to Facebook, where one's audience is typically made up of offline acquaintances, on Reddit there is no pre-existing social context that could be collapsed. Instead, every user is confronted with a vast audience of people from various social, geographic, and cultural backgrounds. Within this complexity, the only suitable way to properly identify and categorise stylistic variation for quantitative studies appears to be combining quantitative methods with a qualitative, ethnographic approach. By spending time on the platform in question, observing the contributions by the various users, and through interviewing expert users (such as the moderators), researchers can gather the necessary knowledge and create the ethnographic familiarity to later adequately conduct their quantitative analysis. I would therefore like to sum up this section with an endorsement of combining qualitative with quantitative studies to face the challenges posed by the diverse repertoires and backgrounds within the online environment.

6. Conclusion

This study investigates the online slang called “pupper talk” as it is used on pet-centred subreddits. The slang is of interest both in itself, and as an instance of stylistic variation in the online environment, offering insights into what influences users’ choices during utterance composition. The study therefore contributes to ongoing linguistic discussions in several ways. First of all, the study is able to provide empirical proof on the weighing of audience and non-audience factors, as distinguished by Bell (1984), for stylistic variation on the online platform Reddit. Audience factors are found to have a higher impact on the probability of the slang occurring than non-audience factors, including the topic and the multimodal prompts. This is in line with earlier studies emphasising the importance of audience factors on other online platforms, such as Shoemark et al. (2017) and Pavalanathan & Eisenstein (2015). My study is also one of the first to take into account the multimodal environment on Reddit but finds that the subjective relevance of the properties of the visual material, as voiced by the informants, is not borne by the statistical model. This does not imply that the visual material should be discarded in future studies, but provides an interesting starting point for further investigation of this apparent discrepancy between subjective relevance of predictors in contrast to their statistical relevance.

On a methodological level, the study emphasises how essential the combination of qualitative and quantitative methods is, especially in the online environment, in order to correctly interpret the output of statistical modelling. The insights gained during the ethnographic pilot study and the comments made by informants add considerable value, for example, to the interpretation of the relevance of the different audience roles indicated by the predictor “use of slang in the previous textual unit”. This combination of methodologies was already advocated, among others, by Androutsopoulos (2006:42) and Bolander & Locher (2014:20). Furthermore, this study illustrates some methodological issues for the quantitative study of stylistic variation online, which will hopefully be of use to other researchers. Among these challenges are the level of measurement of stylistic variation and its implications for logistic or linear regression modelling, as well as the problematic impact of short text length. The international and inter-cultural span of global platforms such as Reddit, and the attached difficulties for researchers interpreting textual contributions, are also discussed.

This study also contributes to the topic of communities and their formation in the online environment. While discussing the applicability of the terms “community of practice” (Meyerhoff 2004) and “virtual community” (Herring 2004) to the subreddits at hand, the usage

and endorsement of slang emerges as one of the central tools to foster a sense of community within a given online platform. Through strategic audience design, joint word play, and employing insider terminology, users are able to create affiliation and strengthen the ties within their communities of practice. For the linguistic investigation of slang, this study contributes a description of the slang "pupper talk" as it is used in animal-centred subreddits in the year 2019, including its orthographic, lexical, and grammatical features. The relation to other online slang phenomena, such as LOLspeak (Bury & Wojtaszek 2017), as well as to first-language acquisition (O'Grady & Cho 2001) is illustrated. This is a further step into the relatively recent effort to describe and explain online slang from a linguistic perspective (Malmkjær 2010, Shifman 2014).

At this point it is necessary to draw attention to the limitations of this study, which are already mentioned in the discussion of methodological challenges (section 5.4). First of all, measuring the stylistic variation with a ratio, instead of the binary distinction between presence or absence of slang, would have considerably improved the accuracy of this study, since it is likely that the amount of slang used within one textual unit makes a difference to the users. A second important limitation concerns the correct identification of the slang within the data collected. Even with the help of ethnographic familiarity with the virtual communities in question, it is still a challenge for the researcher to determine whether an instance of variation should be considered a part of the slang or not. However, this limitation is relativized when considering that every user has a different perception of the slang and its scope (based on their personal exposure and familiarity with the phenomenon), so there can never be a neutral, objective account of which features belong to the slang and how central they are. It should also be emphasised that the description presented is only a snapshot of the phenomenon at a certain point in time. Due to the ever-changing nature of language in the online environment (Danescu-Niculescu-Mizil et al. 2013:307, Golbeck & Buntain 2018:587), it is to be expected that the slang evolves further, acquiring new features and dropping others. The description at hand nevertheless provides a useful account for further investigations on diachronic developments.

Based on the results of this study, several interesting avenues for further research present themselves. As was already alluded to several times, it would be very interesting to investigate how different online slangs develop over time and incorporate or modify each other's features, especially across platform boundaries. A large research gap is also apparent in the area of pet-directed speech, its grammar and lexicon, and the extent of its similarity to child-directed speech. Furthermore, the data collected could be used to provide insights into other questions as well. For example, the data might prove helpful in assessing user's stylistic consistency

across different online groups. To what degree do users vary their style when posting to different subreddits? Summing up, this paper is able to highlight only a small fraction of the linguistic complexity found on the platform Reddit and hopes to inspire further investigations of the platform.

List of references

Androutsopoulos, Jannis. 2014. Languaging when Contexts Collapse: Audience Design in Social Networking. *Discourse, Context & Media* 4-5. 62–73.

Androutsopoulos, Jannis. 2000. Non-Standard Spellings in Media Texts: The Case of German Fanzines. *Journal of Sociolinguistics* 4(4). 514–533.

Androutsopoulos, Jannis. 2006. Introduction: Sociolinguistics and Computer-Mediated Communication. *Journal of Sociolinguistics* 10(4). 419–438.

Androutsopoulos, Jannis. 2008. Potentials and Limitations of Discourse-Centred Online Ethnography. *Language@Internet* 5. 1–20.

Androutsopoulos, Jannis. 2011. Language Change and Digital Media: A Review of Conceptions and Evidence. In Tore Kristiansen & Nikolas Coupland (eds.), *Standard Languages and Language Standards in a Changing Europe*. Novus Press. 145–161.

Androutsopoulos, Jannis. 2013. Online Data Collection. In Christine Mallinson, Becky Childs & Gerard van Herk (eds.), *Data Collection in Sociolinguistics*: *Methods and Applications*. New York: Routledge. 236–249.

Attardo, Salvatore (ed.). 2014. *Encyclopedia of Humor Studies.* Los Angeles: Sage References.

Bateman, John, Janina Wildfeuer & Tuomo Hiippala. 2017. *Multimodality*: *Foundations, Research and Analysis. A Problem-Oriented Introduction.* Berlin: Walter de Gruyter.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire & Jeremy Blackburn. 2020. The Pushshift Reddit Dataset. In *Proceedings of the Fourteenth International AAAI Conference on Web and Social Media*. 830–839.

Baym, Nancy. 1995. The Performance of Humor in Computer-Mediated Communication. *Journal of Computer-Mediated Communication* 1(2).

Bell, Allan. 1984. Language Style as Audience Design. *Language in Society* (13). 145–204.

Bell, Allan. 2001. Back in Style: Reworking Audience Design. In Penelope Eckert & John Rickford (eds.), *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press. 139–169.

Ben-Aderet, Tobey, Mario Gallego-Abenza, David Reby & Nicolas Mathevon. 2017. Dog-Directed Speech: Why Do We Use it and Do Dogs Pay Attention to it? *Proceedings. Biological Sciences* 284(1846). 1–7.

Bender, Emily. 2001. *Syntactic Variation and Linguistic Competence: The Case of AAVE Copula Absence.* Stanford: Stanford University Dissertation.

Berez-Kroeker, Andrea, Lauren Gawne, Susan Kung, Barbara Kelly, Tyler Heston, Gary Holton, Peter Pulsifer, David Beaver, Shobhana Chelliah, Stanley Dubinsky, Richard Meier, Nick Thieberger, Keren Rice & Anthony Woodbury. 2018. Reproducible Research in Linguistics: A Position Statement on Data Citation and Attribution in our Field. *Linguistics* 56(1). 1–18.

Bergstrom, Kelly. 2011. "Don't Feed the Troll": Shutting down Debate about Community Expectations on Reddit.com. *First Monday* 16(8).

Biber, Douglas & Susan Conrad. 2009. *Register, Genre, and Style* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Bivens, Jennifer. 2018. *Describing Doggo-Speak*: *Features of Doggo Meme Language*. New York: City University of New York Master Thesis.

Boddy, Jessica. 23.04.2017. *Dogs are Doggos: An Internet Language Built Around Love for the Puppers.* https://www.npr.org/sections/alltechconsidered/2017/04/23/524514526/dogs-are-doggos-an-internet-language-built-around-love-for-the-puppers?t=1571242326746. (last accessed: 07.06.2020)

Bolander, Brook & Miriam Locher. 2014. Doing Sociolinguistic Research on Computer-Mediated Data: A Review of Four Methodological Issues. *Discourse, Context & Media* 3. 14–26.

Borgi, Marta & Francesca Cirulli. 2016. Pet Face: Mechanisms Underlying Human-Animal Relationships. *Frontiers in Psychology* 7. 1–11.

Borgi, Marta, Irene Cogliati-Dezza, Victoria Brelsford, Kerstin Meints & Francesca Cirulli. 2014. Baby Schema in Human and Animal Faces Induces Cuteness Perception and Gaze Allocation in Children. *Frontiers in Psychology* 5. 1–12.

Bourlai, Elli & Susan Herring. 2014. Multimodal Communication on Tumblr: "I have so Many Feels!". In *Proceedings of the 2014 ACM Conference on Web Science*.

Brown, Keith (ed.). 2006. *Encyclopedia of Language and Linguistics,* 2nd edn. Amsterdam: Elsevier.

Brown, Penelope & Colin Fraser. 1979. Speech as a Marker of Situation. In Howard Giles & Klaus Scherer (eds.), *Social Markers in Speech*. Cambridge: Cambridge University Press. 33–62.

Buntain, Cody & Jennifer Golbeck. 2014. Identifying Social Roles in Reddit Using Network Structure. In Chin-Wan Chung, Andrei Broder, Kyuseok Shim & Torsten Suel (eds.), *Proceedings of the 23rd International Conference on World Wide Web*. New York: ACM Press. 615–620.

Bury, Beata & Adam Wojtaszek. 2017. Linguistic Regularities of LOLspeak. *Sino-US English Teaching* 14(1). 30–41.

Callier, Patrick. 2016. Exploring Stylistic Co-Variation on Twitter: The Case of DH. In Lauren Squires (ed.), *English in Computer-Mediated Communication: Variation, Representation, and Change* (Topics in English Linguistics 93). Berlin: De Gruyter Mouton. 241–259.

Chambers, Jack, Peter Trudgill & Natalie Schilling-Estes (eds.). 2004. *The Handbook of Language Variation and Change* (Blackwell Handbooks in Linguistics). Malden: Blackwell Publishing.

Chandrasekharan, Eshwar, Mattia Samory, Shagun Jhaver, Hunter Charvat, Amy Bruckman, Cliff Lampe, Jacob Eisenstein & Eric Gilbert. 2018. The Internet's Hidden Rules: An Empirical Study of Reddit Norm Violations at Micro, Meso, and Macro Scales. *Proceedings of the ACM on Human-Computer Interaction* 2(CSCW). 1–25.

Chovanec, Jan & Villy Tsakona. 2018. Investigating the Dynamics of Humor: Towards a Theory of Interactional Humor. In Villy Tsakona & Jan Chovanec (eds.), *The Dynamics of Interactional Humor*: *Creating and Negotiating Humor in Everyday Encounters* (Topics in Humor Research 7). Amsterdam, Philadelphia: John Benjamins. 1–26.

Clarke, Isobelle & Jack Grieve. 2017. Dimensions of Abusive Language on Twitter. In Association for Computational Linguistics. *Proceedings of the First Workshop on Abusive Language Online*. 1–10.

Cole, Jeremy, Moojan Ghafurian & David Reitter. 2017. Is Word Adaptation a Grassroots Process?: An Analysis of Reddit Communities. In *International Conference on Social Computing, Behavioral-Cultural Modeling and Prediction and Behavior Representation in Modeling and Simulation*. 236–241.

Danescu-Niculescu-Mizil, Cristian, Michael Gamon & Susan Dumais. 2011. Mark My Words! Linguistic Style Accommodation in Social Media. In *International World Wide Web Conference Committee (IW3C2)*. 745–754.

Danescu-Niculescu-Mizil, Cristian, Robert West, Dan Jurafsky, Jure Leskovec & Christopher Potts. 2013. No Country for Old Members: User Lifecycle and Linguistic Change in Online Communities. In *International World Wide Web Conference Committee (IW3C2)*. 307–318.

Danet, Brenda. 2001. *Cyberpl@y*: *Communicating Online* (New Technologies/New Cultures). Oxford: Berg.

Duggan, Maeve & Aaron Smith. 2013. *6% of Online Adults are Reddit Users*: *Young Men are Especially Likely to Visit the "front page of the internet"*. Pew Research Center.

Eckert, Penelope. 2004. The Meaning of Style. In Texas Linguistic Forum. *Proceedings of the Eleventh Annual Symposium about Language and Society*. 41–53.

Eckert, Penelope. 2006. Communities of Practice. In Keith Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn. Amsterdam: Elsevier. 683–685.

Eckert, Penelope & John Rickford (eds.). 2001. *Style and Sociolinguistic Variation.* Cambridge: Cambridge University Press.

Field, Andy, Jeremy Miles & Zoë Field. 2012. *Discovering Statistics Using R.* London: Sage.

Fiesler, Casey & Nicholas Proferes. 2018. "Participant" Perceptions of Twitter Research Ethics. *Social Media + Society* 4(1). 1–14.

Fiesler, Casey, Pamela Wisniewski, Jessica Pater & Nazanin Andalibi. 2016. Exploring Ethics and Obligations for Studying Digital Communities. In Stephan Lukosch, Aleksandra Sarcevic, Myriam Lewkowicz & Michael Muller (eds.), *Proceedings of the 19th International Conference on Supporting Group Work.* New York: ACM Press. 457–460.

Finlay, Craig. 2014. Age and Gender in Reddit Commenting and Success. *Journal of Information Science Theory and Practice* 2(3). 18–28.

Flanagan, Joseph. 2017. Reproducible Research: Strategies, Tools, and Workflows. In Turo Hiltunen, Joe McVeigh & Tanja Säily (eds.), *Helsinki: Research Unit for Variation, Contacts and Change in English* (VARIENG) (Studies in Variation, Contacts and Change in English 19). Helsinki: University of Helsinki.

Flesch, Marie. 2018. "That spelling tho": A Sociolinguistic Study of the Nonstandard Form of Though in a Corpus of Reddit Comments. In Reinhild Vandekerckhove, Darja Fišer & Lisa Hilte (eds.), *Proceedings of the 6th Conference on Computer-Mediated Communication (CMC) and Social Media Corpora (CMC-corpora 2018)*. 37–40.

Gawne, Lauren & Jill Vaughan. 2011. I can haz language play: The Construction of Language and Identity in LOLspeak. In Maïa Ponsonnet, Loan Dao & Margit Bowler (eds.), *Proceedings of the 42nd Australian Linguistics Society Conference*. 97–122.

Gerrig, Richard & Raymond Gibbs. 1988. Beyond the Lexicon: Creativity in Language Production. *Metaphor and Symbolic Activity* 3(3). 1–19.

Giles, Howard & Tania Ogay. 2007. Communication Accommodation Theory. In Brian Whaley & Wendy Samter (eds.), *Explaining Communication: Contemporary Theories and Exemplars*. 293–310.

Golbeck, Jennifer & Cody Buntain. 2018. This Paper is About Lexical Propagation on Twitter. H*ckin Smart. 12/10. Would Accept! In *International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 587–590.

Graham, Sage. 2019. A Wink and a Nod: The Role of Emojis in Forming Digital Communities. *Multilingua* 38(4). 377–400.

Grieve, Jack, Andrea Nini & Diansheng Guo. 2017. Analyzing Lexical Emergence in Modern American English Online. *English Language and Linguistics* 21(1). 99–127.

Henderson, Tristan, Luke Hutton & Sam McNeilly. 2012. Ethics and Online Social Network Research – Developing Best Practices. In *Proceedings of the 26th BCS Conference on Human Computer Interaction*. 1–4.

Herring, Susan. 1999. Interactional Coherence in CMC. In *Proceedings of the 32nd Annual Hawaii International Conference on Systems Sciences*. 1–13.

Herring, Susan. 2004. Computer-Mediated Discourse Analysis: An Approach to Researching Online Behavior. In Sasha Barab, Rob Kling & James Gray (eds.), *Designing for Virtual Communities in the Service of Learning*. New York: Cambridge University Press. 338–376.

Herring, Susan. 2007. A Faceted Classification Scheme for Computer-Mediated Discourse. *Language@Internet* (1). 1860–2029.

Herring, Susan. 2015. New Fronties in Interactive Multimodal Communication. In Alexandra Georgakopoulou & Tereza Spilioti (eds.), *The Routledge Handbook of Language and Digital Communication*. London: Routledge. 398–402.

Herring, Susan. 2020. Grammar and Electronic Communication. In Carol Chappelle (ed.), *The Concise Encyclopedia of Applied Linguistics*. Hoboken: Wiley Blackwell. 1–11.

Herring, Susan & Jannis Androutsopoulos. 2015. Computer-Mediated Discourse 2.0. In Deborah Tannen, Heidi Hamilton & Deborah Schiffrin (eds.), *The Handbook of Discourse Analysis* (Blackwell Handbooks in Linguistics). Wiley Blackwell. 127–151.

Herring, Susan & Ashley Dainas. 2017. "Nice Picture Comment!": Graphicons in Facebook Comment Threads. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2185–2194.

Hinrichs, Lars. 2016. Modular Repertoires in English-Using Social Networks: A Study of Language Choice in the Networks of Adult Facebook Users. In Lauren Squires (ed.), *English in Computer-Mediated Communication: Variation, Representation, and Change* (Topics in English Linguistics 93). Berlin: De Gruyter Mouton. 17-42.

Honkanen, Mirka. In press. *World Englishes on the Web*: *The Nigerian Diaspora in the USA* (Varieties of English around the World). Amsterdam, Philadelphia: John Benjamins.

Hutton, Luke & Tristan Henderson. 2015. "I Didn't Sign Up for This!": Informed Consent in Social Network Research. In *Proceedings of the Ninth International Conference on Web and Social Media*. 178-187.

Irvine, Judith 2006. Speech and Language Community. In Keith Brown (ed.), *Encyclopedia of Language and Linguistics*, 2nd edn. Amsterdam: Elsevier. 689–698.

Johnson, Christopher. 2001. A Survey of Current Research on Online Communities of Practice. *Internet and Higher Education* (4). 45–60.

Kershaw, Daniel. 2018. *Language Change and Evolution in Online Social Networks*. Lancaster University Dissertation.

Kozbelt, Aaron. 2014. Creativity. In Salvatore Attardo (ed.), *Encyclopedia of Humor Studies*. Los Angeles: Sage References. 181–185.

Lange, Patricia 2007. Publicly Private and Privately Public: Social Networking on YouTube. *Journal of Computer-Mediated Communication* 13(1). 361–380.

Lee, Carmen & David Barton. 2011. Constructing Glocal Identities Through Multilingual Writing Practices on Flickr.com®. *International Multilingual Research Journal* 5(1). 39–59.

Leppänen, Sirpa. 2015. Dog Blogs as Ventriloquism: Authentication of the Human Voice. *Discourse, Context & Media* 8. 63–73.

Levshina, Natalia. 2015. *How to Do Linguistics with R*. Amsterdam: John Benjamins.

Liimatta, Aatu. 2016. *Exploring Register Variation on Reddit*: *A Multi-Dimensional Study*. Helsinki: University of Helsinki Master Thesis.

Malmkjær, Kirsten. 2010. *The Routledge Linguistics Encyclopedia*, 3rd edn. New York: Routledge.

Marcoccia, Michel. 2004. On-line Polylogues: Conversation Structure and Participation Framework in Internet Newsgroups. *Journal of Pragmatics* 36(1). 115–145.

Marwick, Alice E. & danah boyd. 2011. I Tweet Honestly, I Tweet Passionately: Twitter Users, Context Collapse, and the Imagined Audience. *New Media & Society* 13(1). 114–133.

Massanari, Adrienne. 2015. *Participatory Culture, Community, and Play*: *Learning from Reddit.* New York: Peter Lang.

Mattiello, Elisa. 2005. The Pervasiveness of Slang in Standard and Non-Standard English. *Mots Palabras Words* (6). 7–41.

Merriam Webster. *Words We're Watching: Doggo: They're Good Dogs, Webster*. https://www.merriam-webster.com/words-at-play/words-were-watching-doggo. (last accessed 07.06.2020).

Meyerhoff, Miriam. 2004. Communities of Practice. In Jack Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change* (Blackwell Handbooks in Linguistics). Malden: Blackwell. 526-548.

Milroy, Lesley. 2004. Social Networks. In Jack Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change* (Blackwell Handbooks in Linguistics). Malden: Blackwell. 549–572.

Miltner, Kate. 2014. "There's no place for lulz on LOLCats": The Role of Genre, Gender, and Group Identity in the Interpretation and Enjoyment of an Internet Meme. *First Monday* 19(8).

Miniwatts Marketing Group. 2020. *Internet World Stats: Usage and Population Statistics.* https://www.internetworldstats.com/stats.htm. (last accessed 07.06.2020).

Moore, Carrie & Lisa Chuang. 2017. Redditors Revealed: Motivational Factors of the Reddit Community. In *Proceedings of the 50th Hawaii International Conference on System Sciences*. 2313–2322.

North, Sarah. 2007. 'The Voices, the Voices': Creativity in Online Conversation. *Applied Linguistics* 28(4). 538–555.

O'Grady, William & Sook Cho. 2001. First Language Acquisition. In John Archibald & William O'Grady (eds.), *Contemporary Linguistics*: *An Introduction*. St. Martin's Press. 326–362.

O'Halloran, Kay. 2011. Multimodal Discourse Analysis. In Ken Hyland & Brian Paltridge (eds.), *Bloomsbury Companion to Discourse Analysis*. London, New York: Continuum. 120–137.

Paolillo, John 2001. Language Variation on Internet Relay Chat: A Social Network Approach. *Journal of Sociolinguistics* 5(2). 180–213.

Pavalanathan, Umashanthi & Jacob Eisenstein. 2015. Audience-Modulated Variation in Online Social Media. *American Speech* 90(2). 187–213.

Preece, Jenny, Blair Nonnecke & Dorine Andrews. 2004. The Top Five Reasons for Lurking: Improving Community Experiences for Everyone. *Computers in Human Behavior* 20(2). 201–223.

Punske, Jeffrey & Elizabeth Butler. 2019. Do me a Syntax: Doggo Memes, Language Games and the Internal Structure of English. *Ampersand* 6. 1–9.

Python Core Team. 2015. *Python*: *A Dynamic, Open-Source Programming Language.* Python Software Foundation.

R Core Team. 2019. *R*: *A Language and Environment for Statistical Computing.* Vienna: R Foundation for Statistical Computing.

RStudio Team. 2019. *RStudio: Integrated Development for R.* RStudio, Boston. http://www.rstudio.com/.

Reddit Inc. 2018. *Reddit Privacy Policy*. https://www.redditinc.com/policies/privacy-policy. (last accessed 07.06.2020).

Reddit Inc. 2018. *Reddit User Agreement*. https://www.redditinc.com/policies/user-agreement. (last accessed 07.06.2020).

Reddit Inc. 2020. *Reddiquette*. https://www.reddithelp.com/en/categories/reddit-101/reddit-basics/reddiquette. (last accessed 07.06.2020).

Reddit Inc. 2020. *Reddit - Homepage*. https://www.redditinc.com/. (last accessed 07.06.2020).

Rheingold, Howard. 1993. *The Virtual Community*: *Homesteading on the Electronic Frontier*. Reading: Addison-Wesley.

Rickford, John & Penelope Eckert. 2001. Introduction. In Penelope Eckert & John Rickford (eds.), *Style and Sociolinguistic Variation*. Cambridge: Cambridge University Press. 1–20.

Schilling-Estes, Natalie. 2004. Investigating Stylistic Variation. In Jack Chambers, Peter Trudgill & Natalie Schilling-Estes (eds.), *The Handbook of Language Variation and Change* (Blackwell Handbooks in Linguistics). Malden: Blackwell. 375–401.

Seargeant, Philip & Caroline Tagg. 2014. Introduction: The Language of Social Media. In Philip Seargeant & Caroline Tagg (eds.), *The Language of Social Media*. London: Palgrave Macmillan UK. 1–22.

Seargeant, Philip & Caroline Tagg (eds.). 2014. *The Language of Social Media.* London: Palgrave Macmillan UK.

Sebba, Mark. 2007. *Spelling and Society*: *The Culture and Politics of Orthography around the World.* Cambridge: Cambridge University Press.

Sebba, Mark. 2012. Orthography as Social Action: Scripts, Spelling, Identity and Power. In Alexandra Jaffe, Jannis Androutsopoulos, Mark Sebba & Sally Johnson (eds.), *Orthography as Social Action*: *Scripts, Spelling, Identity and Power* (Language and Social Processes 3). Boston, Berlin: De Gruyter Mouton.

Serpell, James. 2004. Factors Influencing Human Attitudes to Animals and Their Welfare. *Animal Welfare* (13). 145–151.

Sherzer, Joel. 2014. Speech Play. In Salvatore Attardo (ed.), *Encyclopedia of Humor Studies*. Los Angeles: Sage References. 727–730.

Shifman, Limor. 2014. Internet Humor. In Salvatore Attardo (ed.), *Encyclopedia of Humor Studies*. Los Angeles: Sage References. 389–393.

Shoemark, Philippa, James Kirby & Sharon Goldwater. 2017. Topic and Audience Effects on Distinctively Scottish Vocabulary Usage in Twitter Data. In *Proceedings of the Workshop on Stylistic Variation*. 59–68.

Singer, Philipp, Fabian Flöck, Clemens Meinhart, Elias Zeitfogel & Markus Strohmaier. 2014. Evolution of Reddit: From the Front Page of the Internet to a Self-Referential Community? In *International World Wide Web Conference Committee (IW3C2).* 517–522.

Squires, Lauren (ed.). 2016. *English in Computer-Mediated Communication: Variation, Representation, and Change* (Topics in English Linguistics 93). Berlin: De Gruyter Mouton.

Stommel, Wyke. 2008. Conversation Analysis and Community of Practice as Approaches to Studying Online Community. *Language@Internet* 5(5). 1–22.

Tagg, Caroline & Philip Seargeant. 2014. Audience Design and Language Choice in the Construction and Maintenance of Translocal Communities on Social Network Sites. In Philip Seargeant & Caroline Tagg (eds.), *The Language of Social Media*. London: Palgrave Macmillan UK. 161–187.

Tannen, Deborah, Heidi Hamilton & Deborah Schiffrin (eds.). 2015. *The Handbook of Discourse Analysis* (Blackwell Handbooks in Linguistics). Malden: Wiley Blackwell.

Thurlow, Crispin, Christa Dürscheid & Federica Diémoz (eds.). 2020. *Visualizing Digital Discourse*: *Interactional, Institutional and Ideological Perspectives* (Language and Social Life 21). Boston: De Gruyter Mouton.

Tolins, Jackson & Patrawat Samermit. 2016. GIFs as Embodied Enactments in Text-Mediated Conversation. *Research on Language and Social Interaction* 49(2). 75–91.

van Leeuwen, Theo. 2015. Multimodality. In Deborah Tannen, Heidi Hamilton & Deborah Schiffrin (eds.), *The Handbook of Discourse Analysis* (Blackwell Handbooks in Linguistics). Malden: Wiley Blackwell. 447–465.

Vasilev, Evgenij. 2018. *Inferring Gender of Reddit Users*. Koblenz: Universität Koblenz-Landau Master Thesis.

Wahlster, Wolfgang. 1994. User and Discourse Models for Multimodal Communication. In Joseph Sullivan & Sherman Tyler (eds.), *Intelligent User Interfaces*. New York: ACM Press. 45–67.

Wenger, Etienne. 1998. Communities of Practice: Learning as a Social System. *Systems Thinker*. 1–10.

Winter, Bodo. 2020. *Statistics for Linguists*: *An Introduction Using R.* New York: Routledge.

Appendix 1: Transcripts of pilot study

[redacted due to ethical concerns]

Appendix 2: Coding scheme

Counted	**Not counted**
aoow, good slep, diggy dig dig, more enjoy, oh heck, the L E N G T H of this lad, snow pup, sky water, He looooonnnngggg, snootboye, barkour, protec, he snac, he loves bac, frikk, him loves bred, mlem mlem mlem, she construc, longboye does a dilemma, pittie, boop, she nom, distract with cute, longboye lookie-likies, spoopy doggo, l o n g cowpoke, I can has cheezburger, mission impawssible, henlo, fren, druggos, chonky little hands, pick up more stuffs, I have ball too, countryside S N I F F S, seal pupper, splootin', She's doing a shrink, smol curli boi, pawp, anteater boye, treatos, wolfdoggo, 10/10 good g0rl, zoomies, she a good girl, snoot, after soooooo l o n g, gras dog, doggie gets a smol doggo, bestest boyyy, betterest boy, Assassin's breed, Caninessin's breed, bouy, deserbs, dunt, one runny boye, velvet hippo, visible shook, very scare, spoopy boi, much startle, many scared, doby, left pawed, blacklab, woof, Bostie, Boston, Luke Sky water, longegoof, doggy, tuckies, shelti, grampa, hewwo, pibble, goodestest, ear-ectile dysfunction, whipper schnauzer, so m a n y, salivation, pit, Han Solong, dutchie, grey, tailcopter, thatttt, foods, meerdogs, meerpups, tippytaps, to tippy tap, dawg, cuteee, foods, helpful blop, she building a house, he s a c r i f y, tuff, shhhhh, souper, sm0rt, ginormous, teef, such cute, once in a whale, tootsies, pyrs, kitters, situation at paw, everyfing, moar, fuzzies, glow-pup, lomg, fanks, noodle poodles, noodle horse, noice, koalifications, through stick and thin, be in safe paws, longgirl, longboy, dobergirl goldens, BC ('border collie'), *longfriend, long boy, staffy, monch, floof*	*pic* ('picture'), *OP* ('original poster'), *gif, lmao* ('laugh my ass off'), *lol* ('laughing out loud'), *IG* ('Instagram'), *slo-mo* ('slow motion'), *on the reg* ('regular'), *omg* ('oh my god'), *sub* ('subreddit'), *pls* ('please'), *xtra* ('extra'), *fave* ('favourite'), *rep* ('reputation'), *bc (*'because'*), congratz* ('congratulations'), *lil* ('little'), *legit* ('legitimate'), *He's ADORABLE, ThAt BiG mEaN dOg, peeps* ('people'), *Free fallin', wanna, gotta, tho, bud, sup, yo, djyou get that thing I sentchya?, fella, mom, mumma, them felines, show me them puppies, in da planet earth, cuz, he'z, yikes, daddio, n, woah, yelp, oof, teeheeheeehee, hehehe, ahem, oh lawd he comin, cutie, go potty, relaxxin, gator* ('alligator'), *no biggie, I love me some ..., kiddo, boo!, fo sho, nyoom, nuf said, corgi, boxer, collie, snuggle, wiggle, fluffy, swag, bark, lurcher, slobbery, beefy, honk, whimper, pooch, smooch, poop, splash, dufus, pat pat, You sicko!, spooktober, hospital dog, cuddlebugs, glomper, amp* ('amputation'), *good boy, good girl, pet giraffe, sharkdog, paw-to-paw-ratio, I sheet myself, goofball, fluffballs, snow-five, atta boy, derpy, poopy, bonks, sproing, beep, boing, moonmoon, goldings, goober, decon* ('decontamination'), *dang, noot noot, long lady, I cano hear you, smooshy, thicc*