Masaryk University
Faculty of Arts

Department of English
and American Studies

English Language and Literature

Tibor Varga

Language of YouTube Video Comments

Bachelor's Diploma Thesis

Supervisor: Mgr. Jan Chovanec, PhD.

2009

I declare that I have worked on this thesis independently,
using only the primary and secondary sources listed in the bibliography.

………………………………………………………
Author’s signature

Acknowledgment

I would like to thank Jan Chovanec for the guidance he provided me as a supervisor and for his patience with the delays which occurred during the writing process. Also, my special thanks go to my girlfriend and family. Without their support, the thesis would have never been completed.

Table of Contents

1. Introduction

YouTube: Broadcast Yourself ^{TM} ("YouTube") is a public video sharing Web site which was founded in February and officially launched in November 2005. The site consists of user-generated content and ranks among the most successful Web 2.0 projects. Registered users can upload their videos to the server, share them with the world, watch all other videos uploaded to the site,[1] and interact with the community. Unregistered users, too, can watch all videos, but they are not permitted to upload and share their own content and to interact with the community (Wikipedia 2009d).

This thesis deals with YouTube video comments. It analyzes their language with regard to its specific contexts of a type of CMC and focuses primarily on those of its formal features which are characteristic of Internet language in general. The selected features include acronyms, emoticons, laughter variants and spelling variants of personal pronouns (Tagliamonte & Denis 2008).

The thesis is divided in two parts. The theoretical part explores the contexts of the language found in the comments. It defines and classifies computer-mediated communication (CMC), determines the position of YouTube video comments within this classification system, introduces a communication model describing Web site mediated communication, and defines the nature of video comments as a concrete type of CMC.

The centre of the practical part is a linguistic analysis of the selected linguistic features, and their description. In its introductory section the methodology and the procedure of creating a corpus of comments are described. This section is followed by the actual analysis of the corpus, the results of which are summarized and discussed in the concluding section.

[1] Except for minors who cannot watch videos for adults. The same applies for unregistered users who cannot watch the content inappropriate for minors either (Wikipedia 2009d).

2. Theoretical Part

CMC is a concept which yields numerous terminological obfuscations and which is difficult to define without restricting one's unprejudiced approach to the object of analysis (Lange 2008). According to John December (1997), CMC is "a process of human communication via computers, involving people, situated in particular contexts, engaging in processes to shape media for a variety of purposes."

Webopedia (2009) offers a more concrete definition by claiming that CMC refers to human communication via computers and encompasses different forms of synchronous and asynchronous interaction that humans have with each other using computers as tools to exchange text, images, audio and video. These forms include e-mail, network communication, instant messaging, text messaging, hypertext, distance learning, Internet forums, USENET newsgroups, bulletin boards, online shopping, distribution lists and videoconferencing.

When combining these two definitions we are basically familiar with both an abstract notion of CMC and a list of concrete examples. However, neither of these definitions addresses specific qualities of a concrete kind of CMC, such as YouTube video comments. Therefore, we need to describe the particular nature of YouTube comments to understand their communicative and linguistic specifics.

To begin with, commenting videos is not only mediated by computer, but also by the Internet. In another article on CMC, December (1996) introduces and defines Internet-based CMC. A long time before Lange's article on obfuscating terms in online research (2008), December points to the fact that the term CMC as such is actually a misnomer when talking about the communication taking place on the Internet. For this reason, he defines a specific sub-type of CMC, an Internet-based CMC. Likewise, and explicitly this time, Lange (2008) points out that the term CMC "emphasizes a computer's rather than a network's role in facilitating remote communication.” Whenever I use CMC in this thesis concerning YouTube, an Internet phenomenon, I refer to it with regard to the importance of the Internet as a medium.

2.1 Classification of CMC

To arrive at an accurate description of the nature of video comments as a type of CMC, it is necessary to find their place within the system of general CMC classification. From various classifications available I selected those which take into account the synchronicity of communication and the functions the particular types of CMC perform.

As for the former criterion, it is generally accepted that CMC can be either synchronous or asynchronous (Šmahel 2003, Crystal 2006, Jordan 2008, Tagliamonte & Denis 2008, Marcus 2008). In the case of a synchronous mode, “users’ contributions are not buffered but transmitted directly” (Kimmerle & Cress 2008: 435) which means there is little or no time between posting the contribution and the moment in which the message is received by the recipient. In the case of asynchronous CMC, the message is posted at a given time and the recipient reads it at a later date (p. 435).

The asynchronous mode encompasses e-mails, threaded discussions and bulletin boards, whereas the synchronous mode includes UNIX based TALK that is used for one-to-one conferencing; MUDs (Multi-User Dimensions) and MOOs (MUD Object-Oriented) that refer to virtual text-based environments; and Chat systems, such as Internet Relay Chat (IRC) or instant-messaging.[2]

Scholars are generally aware of the fact that the above outlined distinction does not work for all communicative situations. Instant messaging clients, such as ICQ, can be easily

[2] Werry, C. Linguistic and interactional features of Internet Relay Chat. In Herring (Ed), *Computer-mediated communication: Linguistic: social ancultural perspectives* (pp. 47-64). Pragmatics and beyond new series 39. Amsterdam: John Benjamins, 1996. Cited in Yilmaz (2007: 30).

used in an asynchronous way as well – the recipient may read the message long time after it has been posted. People also utilize various types of away or status messages (Lange 2008: 439). Likewise, MUDs cannot only be classified as synchronous, as many of them contain features allowing people to send private messages or posts to the whole community and thus combine synchronous and asynchronous mode (p. 439). Therefore, it is more precise to claim that the synchronous forms of CMC are those which are *designed to be* instant. The instantaneousness of their actual use is conditioned by the physical presence of the recipient at the computer and his/her interest to read the message and thus become involved in the exchange.

When discussing the criterion of synchronicity in the case of YouTube video comments, it is evident that they are of asynchronous nature, similarly to discussion forums and bulletin boards which, too, represent many-to-many communication (Gong & Ooi 2008: 921). As regards the criterion of function, Gong & Ooi note that synchronous forms are usually classified as platforms for social interaction, while the asynchronous forms are regarded as more oriented towards information sharing and seeking (p. 921).

This is, however, not true of YouTube comments, the part of a social networking site, which are a platform for social interaction, too, in spite of being asynchronous. The comments serve from their very nature as a subjective way of evaluating videos, evaluating other users' comments (cf. Weder 2008) and as a way to post thoughts and opinions inspired by the video or the comments of other users. Some additional information about video or its meaning can be found in the comments, but it is at times really difficult to discern it from subjective, emotional and evaluative content.

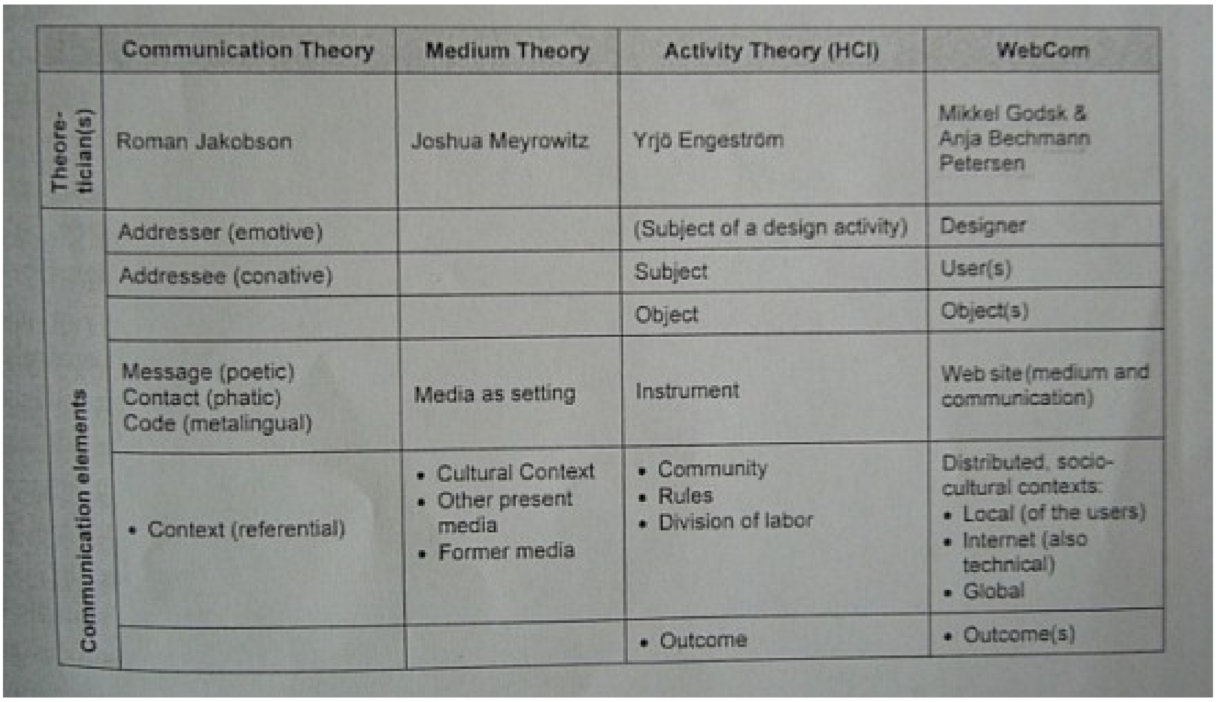
2.2 Communication Model for Web Site Communication

So far YouTube comments have been defined as an asynchronous form of CMC which is, more specifically, mediated by the Internet. When approaching YouTube comments, we can narrow the scope of CMC even further, labelling it as Web site mediated communication (WSMC). A theoretical model by Godsk & Petersen (2008), WebCom, is a useful concept to refer to when we are trying to understand, analyze and describe its nature.

WebCom is as a complex communication model which falls within the field of human-computer interaction theory (HCI). It does not address parts of concrete Web sites, such as video comments, but since it describes the Web specific aspects directly on the example of YouTube, it is possible to apply its principles deductively and describe in what way YouTube users communicate when using the video comments feature.

WebCom is based on three different theoretical approaches and tries to combine their best properties. The theories in question are communication, medium and activity theory (see Figure 1). Godsk & Petersen try to avoid being "too narrowly focused and limited to the communication elements of each individual field" and instead focus on bridging these fields (2008: 380).

Figure 1: WebCom (Godsk & Petersen 2008: 385)

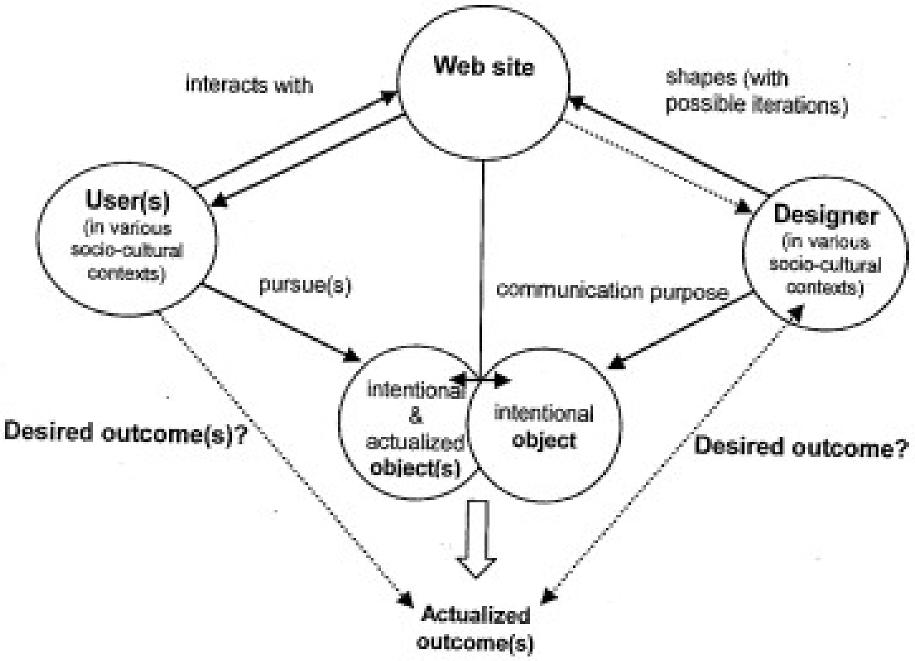


	Communication Theory	Medium Theory	Activity Theory (HCI)	WebCom
Theore-tician(s)	Roman Jakobson	Joshua Meyrowitz	Yrjö Engeström	Mikkel Godsk & Anja Bechmann Petersen
Communication elements	Addresser (emotive)		(Subject of a design activity)	Designer
	Addressee (conative)		Subject	User(s)
			Object	Object(s)
	Message (poetic) Contact (phatic) Code (metalingual)	Media as setting	Instrument	Web site (medium and communication)
	• Context (referential)	• Cultural Context • Other present media • Former media	• Community • Rules • Division of labor	Distributed, socio-cultural contexts: • Local (of the users) • Internet (also technical) • Global
			• Outcome	• Outcome(s)

As can be seen in the figure, the individual theories always lack some important elements. Joshua Meyrowitz's medium theory disregards the addresser and the addressee and focuses mainly on the media as a setting in a given cultural context (Godsk & Petersen 2008: 385). Yrjö Engeström's activity theory is quite complex, but it does not include the notion of what WebCom calls "the designer". Another problem with this theory is the fact that it cannot be applied to WSMC in the same way as in the case of interpersonal face-to-face communication (p. 385). Finally, the notorious Jakobson's communication theory is focused mainly on verbal interaction and does not cover the visual aspects unique to mass medial WSMC. Jakobson does not deal with the outcome of the communication act either, which means that his theory is not suitable for assessing the communication as successful and unsuccessful (Godsk & Petersen 2008: 382).

Godsk & Petersen claim that their WebCom combines the three above mentioned theoretical approaches and covers the wide spectrum of communication elements. Their scheme of the model as depicted in Figure 2 introduces the notions of designer, Web site, user(s), objects, contexts and outcomes.

Figure 2: WebCom (Godsk & Petersen 2008: 387)


Web site
interacts with
shapes (with possible iterations)
User(s) (in various socio-cultural contexts)
Designer (in various socio-cultural contexts)
pursue(s)
communication purpose
intentional & actualized object(s)
intentional object
Desired outcome(s)?
Desired outcome?
Actualized outcome(s)

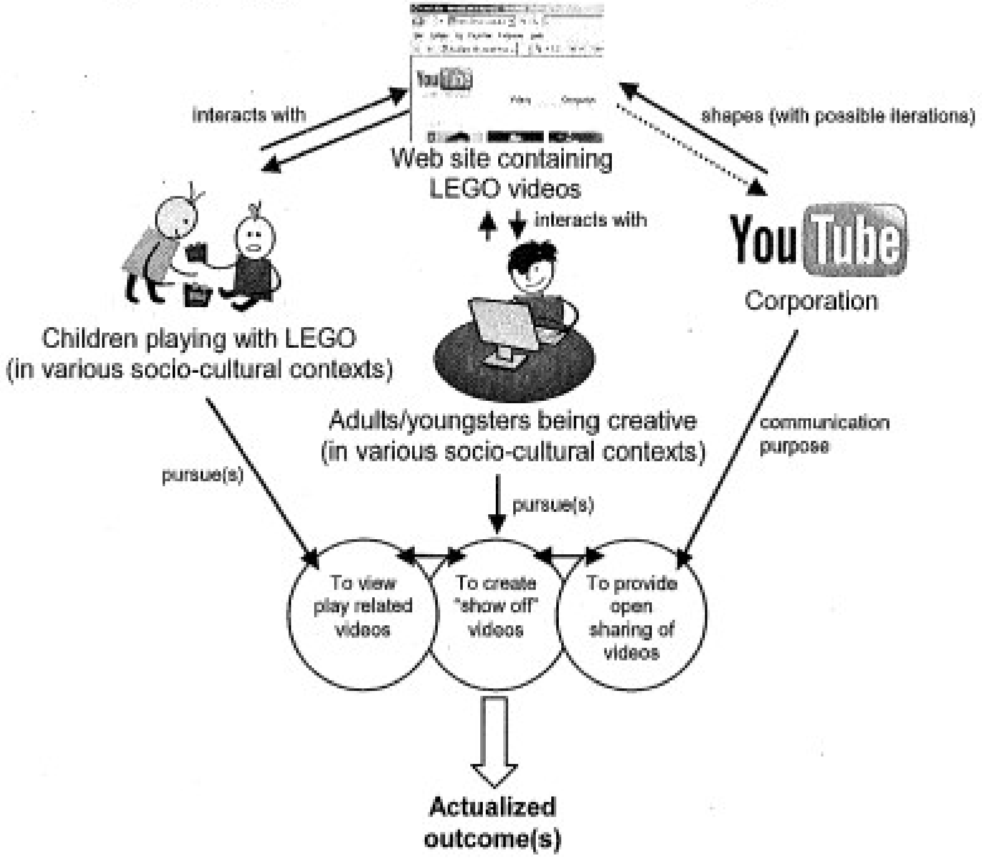
The first three elements constitute the imaginary communication triangle. The designer is defined as a creator of the Web site who considers all relevant factors and elements of communication during the site building process. Designer's site building activity is in many cases an iterative continuous process during which he often makes a use of user's feedback (Godsk & Petersen 2008: 388). The Web site is a "setting for communication and collaborative activities between users and between user and designer" and a place for content and form (p. 389). It is actually the medium which allows the communication to take place between the users and the designer. Sometimes (and YouTube and other Web 2.0 sites definitely are such cases) the communication technically only consists in the users collaborating with each other on a certain intentional object (see further) and the designer playing the role of a frame-setter (p. 388). Users (excluding designer) are "producing and consuming subjects, negotiating between the Web site and their objects through the rules and elements of communication" (Godsk & Petersen 2008: 390).

The other three elements are related to the wider context of WSMC. The objects are closely related to the intentions of the interactants. They are the material, the content of the Web site, towards which the activity of interactants is directed in order to achieve the use purposes (p. 390). As regards the contexts, they are of several types: the local context, which can be identified by investigating the life of the users and the context in which they use the site; the Internet context, which is defined by other Web sites and norms on the Internet; and the global context, which addresses all "socio-cultural elements that specify a common frame of reference" in Web site communication (p. 391). Finally, the outcome is a crucial element of WSMC, as it can indicate whether the communication act which took place was successful (p. 391).

The intentional, or desired, outcome is an ideal effect of communication for the sake of which the interactant took part in the communication act. The actualized outcome is the real effect the communication had on the interactants (the users and the designer). If intentional and actualized outcomes of the interactants coincide, both the parties have achieved their use purposes and the communication act was successful. If intentional and actualized outcomes differ, a problem has occurred with one or more elements of communication and the communication act was unsuccessful (Godsk & Petersen 2008: 391). Godsk & Petersen demonstrate their model on a concrete example of YouTube WSMC depicted in Figure 3.

Figure 3: Concrete example of WSMC on YouTube (Godsk & Petersen 2008: 393)

Figure 6. An example mapping of YouTube with regards to LEGO video clips


interacts with
Web site containing LEGO videos
shapes (with possible iterations)
interacts with
YouTube
Corporation
Children playing with LEGO (in various socio-cultural contexts)
Adults/youngsters being creative (in various socio-cultural contexts)
communication purpose
pursue(s)
pursue(s)
To view play related videos
To create "show off" videos
To provide open sharing of videos
Actualized outcome(s)

When applying this model specifically on YouTube video comments, the designer only interacts with the users through introducing various features and limits of the comments section of the page (cf. 2.3.), but, technically speaking, he only plays a role of the frame setter. The communication mainly takes place between users from various socio-cultural contexts within the text comments section of the video page which represents the medium. The user-generated content interacts with the users and depending on their use purposes it leads to the creation of actualized outcome, the crossing point of their desired outcome and the real effect of communication. The object is represented mainly by the text of the comments, but also by the comments rating (cf. 2.3.) and the nick names of the users which open further communication possibilities and provide the interactants with more information about each other.

2.3 Features Available to Comment Posters

YouTube video comments are specifically devised to fit their primary purpose – commenting on uploaded videos. They have certain extra features on the one hand and some limitations on the other hand which need to be addressed in order to better understand what is going on when users use the video comments platform.

First of all, the video comments are not the only way for users to interact via text on YouTube. There are several other options, each of them fitting for specific communication purposes. These options, if enabled, include channel comments, private messages, stream chat and group posting (YouTube 2009e). Put simply, a channel is a sophisticated user profile. Therefore, channel comments are the best way to comment on specific user's videos, playlists, activities and his/her personal description. For sending a private message to a specific user, there exists a Send Message form which makes it possible to contact the addressee without revealing the contact details to the sender. Stream chat is a tested feature as of 1st April 2009 and it allows users to interact in real time while watching different videos – for the individual user the stream is a room consisting from the user-selected video and chat (YouTube 2009d). Therefore, the stream chat is the only feature of a synchronous nature in the text-based CMC on YouTube. The group posting is closely related to YouTube groups and it enables the users to discuss the common theme publicly within a group of people (YouTube 2009f).

Second, the text comments on YouTube are subject to partial moderation, or self-moderation. Provided that the video owner enables posting comments and their rating, YouTube only shows comments which obtain -5 points or better implicitly. The point system is simple: any registered user can click the thumbs-up or thumbs-down icon which equals +1 or -1 point, respectively, depending on the user-assigned quality of the comment. The comments with less than -5 points are not displayed but can be viewed by changing the options in the Text Comments bar below the video window. The comments which have been marked as spam can also be viewed by clicking the Show button next to the spam message. Still, such a way of moderation cannot eliminate all abusive content, personal insults and undesirable off-topic posts. It is also worth a mention that with an increasing dissatisfaction of users with the nature of YouTube comments (Berens 2006), the Hide Comments option has been enabled and the videos can now be viewed with no text comments displayed at all.

Third, the text comments on YouTube have fixed rules for ordering. The order in which they appear cannot be forced by users, as it is governed by the overall number of comments. If all the comments to a video fit on one page, they appear in reversed chronological order (i.e. from the most recent on top to the oldest one at the bottom). If, however, there is more than one page of comments, the reply structure is not adhered to and the original tree structure can only be viewed by clicking the View All Comments button. Only then the user can view his/her post as a reply to the post he/she replied to (YouTube 2009b). The order of the comments is therefore not uniform and can be misleading at times, especially for new or inexperienced users.

Fourth, YouTube has some other specific features for monitoring, deleting and posting comments. The user can choose to display a module with his/her five most recent comments on other videos on his/her profile page. Also, the user can trace all comments on the videos uploaded by him/her. If the user decides to remove an already posted comment, it is possible, but YouTube stores the content of the comment for their purposes. The deletion of the comment is indicated by “Comment removed by user” message. Similarly, the author of the video has the right to remove any comment to his video at his discretion. The message in such cases reads “Comment removed by author”. Finally, when posting the comment, user can choose to incorporate a deep link into the comment which makes it possible to highlight specific moments in the video. The time code such as 1:45 is transformed into the clickable hyperlink and when other users click it, the video starts playing from the given time (YouTube 2009c).

Fifth, the comment posters may choose to play back their comments before posting them. On clicking the Audio Preview button, the text-to-speech engine reads their comment for them. There has been some debate over usefulness of such a feature (Shankland 2008), but in any case, this tool as if pushes this kind of text-based CMC yet further towards oral communication on oral/written scale (cf. Rivens Mompean 2003).

Quite naturally, there also exist certain limitations for comment posters. Apart from the above described self-moderation executed by registered users and the moderation on the part of video owners, there is also a maximum character count to be adhered to which is set to 500 character as of 1st April 2009 (YouTube 2009a). YouTube does not limit the number of comments, but, in order to prevent potential spammers, it prompts a verification code box when the comments from one user are too frequent over a certain period of time (YouTube 2009a). The video owner can also choose to entirely disable comments or to always manually approve the individual posts before they appear on the site. Likewise, the comments rating can be turned off if the video owner so wishes (YouTube 2009g).

2.4 Description of Video Comments as a Type of CMC

It has been demonstrated that commenting videos on YouTube is mediated through computer, the Internet and the actual Web site. We may therefore refer to it as a kind of CMC when being aware of a descendent hierarchy of these three terms; CMC as an umbrella term and Web site mediated communication as a specific term.

As regards the criterion of synchronicity, YouTube comments are of asynchronous nature. What happens, though, is that the most viewed videos are responded to so frequently that the posts quickly disappear from the video viewers' sight, similarly as in chats. Although the older comments are archived and they are linked to at the bottom of the page, there is neither search engine nor another function to find (or filter to find) the specific ones. Therefore, the asynchronous mode which usually implies "append-only class of persistance" (Leetaru 2008: 876), is not precisely append-only in YouTube's case, given that the number of comments is at times really huge.[3]

Mentioning sight in the last paragraph was quite intentional, as "the visual" plays a critical role in the genre of video comments. They are, unlike discussion forums and other asynchronous kinds of CMC, inseparable from the audiovisual objects of discussion, i.e. videos. In reality, this interconnectedness of the text and the video has further consequences. All users are normally acquainted at least with a part of the video and this common experience shapes their opinions on it. In the case of discussion forums, the interactants are familiar with the topic as well, but each of them has a unique experience of it and there is no common shaping experience as in YouTube's case. The situation, however, works vice versa as well. When users read the comments while watching the videos, the comments may alter

[3] The most discussed video as of 15th April is "Macedonia is Greece" with the total of 637,001 comments. See: 'Macedonia is Greece'. *YouTube: Broadcast Yourself.* 15 April 2009 <http://www.youtube.com/watch?v=-oHivXjiX_w>

and shape their viewing experience, as they can learn more details about the video or become biased towards it by the evaluative content of some comments.

By addressing the shaping of user's experience, I have touched upon the function the comments perform as a kind of CMC. Taking these from mass communication theory perspective, there are four general functions of CMC as described by Wenner[4] within his uses-and-gratifications approach: surveillance, entertainment / diversion, interpersonal utility and parasocial interaction. When applied to YouTube video comments specifically, surveillance includes obtaining various pieces of information by being passively or actively involved in commenting the videos. This includes for instance the above-mentioned shaping of user experience of viewing the video. The dimension of entertainment and diversion is fulfilled whenever commenting serves to escape casual life and entertain the user. By interpersonal utility we understand obtaining and sharing information on the video and also mediating new experience and awareness of the plurality of possible views on the video (cf. Lády 2007: 11). And finally, if the users feel a sense of (YouTube) community, the comments also function as way of parasocial interaction. Quite naturally, these functions are not represented equally; some of them may not be present at all in some cases, in others they all may co-exist with different intensity.

When analyzing video comments, at least two specific functions need to be added, one of which was outlined by Berens (2006): "The difference between watching movie in the theater and watching it alone in your house is profound because once you add other people that viewing experience can become eventful." Thus, he introduces another function of YouTube comments – that of adding "eventness" to the user's experience of the video. This function may be partly covered by the four primary functions listed above, but still it is worth mentioning it in isolation, as its presence in the basic functions is not self-evident.

[4] Wenner, L. A. (1986). 'Model specification and theoretical development in gratifications sought and obtained research: A comparison of discrepancy and transactional approaches'. Communication Monographs, Vol. 53, 160-179. Cited in December (1996).

Berens, inspired by Mikhail Bakhtin, points out another co-related term in this context: unrepeatability. “A movie is endlessly repeatable, but the audience will probably never coalesce again. The presence of other people makes us experience a movie differently: we’re more likely to laugh out loud, for example. Watching a performance collectively, we ourselves perform our role as audience” (Berens 2006).

However, Berens himself strictly condemned YouTube comments as “inane”. I do not entirely agree to this point. The comments may be information- or thought-wise empty, but their sole existence and flourishing speaks for itself. The comments perform similar role as the audience in the theatre, or rather, as the audience at the pop concert. Hardly any information or deep thoughts can be found in their utterances, but still these manifestations, such as shouting and screaming, certainly belong to these events and are meaningful for the people present in the crowd.

The other specific function of YouTube comments is that of stimulating linguistic awareness. Even though many popular sources (such as Berens 2006, O’Connor 2005 and Axtman 2002) deprecate the comments linguistically, many scholars (Tagliamonte & Denis 2008, Crystal 2006, Thurlow 2006) argue the opposite. It is clear that the issue can never be seen black-and-white and it should always be judged in relation to the individual person’s linguistic competence.

As far as the people can switch from the casual language of comments to other stylistic levels of language, the act of writing “super-brief comments like, ‘crap!‘, ‚brilliant‘, ‚this guy sucks‘, and ‚OMG so funny‘” (Berens 2006) not only cannot harm their linguistic skills, but can paradoxically stimulate their linguistic awareness. These people need to be aware of how to behave linguistically so that their comment is accepted as appropriate by the community. Otherwise, with their comments too formal, people would be laughed at and made fun of, like in the popular article titled Ten YouTube Comments Translated into Standard English (authors' surnames unknown, Andy & Dave 2009).

The theoretical part has demonstrated in what ways YouTube video comments are specific and what contexts need to be taken into consideration when analyzing their language. These include above all the asynchronous nature of the comments, different levels of persistance depending on the number of comments on the page, the interconnectedness of the video and its comments, and the functions the comments perform. With these broad theoretical contexts in mind the linguistic analysis of authentic video comments will be approached.

3. Practical Part

Throughout the years of CMC research many terms have been coined to denote and define the kind of language found on the Internet. Annick Rivens Mompean (2003) uses the term "electronic language"[5] to refer to the language of e-mails and chat rooms. David Crystal (2006: 19), in his *Language and the Internet,* uses Netspeak as [a] "succint, and functional enough [name], as long as we remember that 'speak' here involves writing as well as talking, and that any 'speak' suffix has a receptive element, including 'listening and reading'."

Some other terms for Internet language are also listed in Crystal's book: Netlish, Weblish, Internet language, Cyberspeak, electronic discourse, interactive written discourse and even CMC (which, as he points out, refers to the medium rather than to the language). In Wikipedia, for instance, the entry dealing with the issue of language on the Internet is called Internet slang (Wikipedia 2009b). As Crystal points out, all these "terms have different implications" (Crystal 2006: 19).

[5] Strictly speaking, she only uses "English" instead of "language" in her study.

I, therefore, use the term “Internet language” in this thesis because I regard it as a neutral expression for the kind of language one can experience when using the Internet. What I understand by Internet language covers not only the marked elements, such as acronyms and emoticons, but also the nature of the language itself, which comes somewhere between oral and written (Rivens Mompean 2003).

3.1. Corpus and Methodology

I decided to analyze formal aspects of the comments. Since no study focusing directly on the language of YouTube video comments has been conducted so far, it is proper to start from the basic level, i.e. that of form. Moreover, formal features can indicate a lot of discourse tendencies and stimulate future research into more advanced levels of communication.

To obtain a representative corpus in which the identification of the most frequent formal features would be possible, I chose five videos from different video categories (YouTube 2009h) and abstracted approximately 250 comment-long stretches from each of them.

The criteria for the selection were carefully set. However, it must be emphasized that this task alone would need an expert in-depth research to return the most relevant videos for the analysis because YouTube is so diverse a platform that basically any filtering fundamentally diminishes the diversity of the corpus. Furthermore, as a linguist I was not trying to encompass all the content diversity, i.e. different sorts of videos as they come in the categories, but primarily the linguistic diversity, i.e. different sorts of registers as I expected them to emerge depending on different categories.

Figure 4: YouTube Categories (YouTube 2009h)

Cars & Vehicles
Comedy
Education
Entertainment
Film & Animation
Gaming
Howto & Style
Music
News & Politics
Non-profits & Activism
People & Blogs
Pets & Animals
Science & Technology
Sport
Travel & Events

I grouped the above listed categories into two domains which contain the material of a similar nature. The domains and the respective categories are listed in Figure 5.

Figure 5: Domains

Domain	**Categories**
1. Leisure (L)	Cars & Vehicles Comedy Entertainment Film & Animation Gaming Music People & Blogs Pets & Animals Sport Travel & Events
2. Information & Knowledge (IK)	Education Howto & Style Science & Technology News & Politics Non-profits & Activism

The overall similarity of the material within these two domains is useful for choosing relevant videos for the analysis. Although these domains can appeal to the same audience (anyone can be interested in both of them), the ways in which the audience discuss the videos are likely to develop in different directions, depending on the content. Likewise, the formality and informality in these domains is likely to differ because the stylistic properties of the language present in videos themselves are different, and so is the audience. These hypotheses will be confirmed or refuted by the analysis and the discussion will be made in 3.6.

Obviously, both domains, regardless of their actual sizes, have to be represented in the corpus provided that I want to test the difference between them. With regard to the size of L domain and the fact that YouTube is above all a place to spare leisure time at (Wesch 2008), there will be three videos from L domain and two videos from IK domain. This disproportion will not harm the analysis provided that the linguistic features in individual domains will be weighed to the overall number of words in the domains (see further).

The second criterion I decided to set up for the selection of videos was that of relevance. From my observation, it is evident that there are many videos which do not really suit the category they are in. I therefore tried to choose videos I found typical and representative for the given category, although this was only done on the basis of a very subjective personal experience.

The third criterion appeals to the selection of five categories within individual domains. In this particular case I drew on the study of most popular YouTube categories published by professor Wesch (2008) from Kansas State University and on another source providing statistics of popular YouTube searches (Meunier 2008). The list of selected categories and concrete videos is provided in Figure 6.

Figure 6: Selected Videos (see Primary Sources)

Domain	**Category**	**Video title**
Leisure	Music	Britney Spears - Circus
Leisure	People & Blogs	First Blog / Dorkines Prevails[6]
Leisure	Film & Animation	Wolverine Movie Trailer 2009
Info & Know	Education	A Vision of Students Today
Info & Know	News & Politics	Gaza Tunnels – Israel/Palestine

[6] Note that lonelygirl15, the author of the video, is not an amateur girl publishing her video blog, but a professional actress who stylized in the role of an unknown home-schooled girl (Woletz 2008: 595).

The final step was to choose approximately 250 comment-long stretches to be analyzed. All the stretches are ordered from the oldest comments, i.e. from the very first comment on the video up to 250th one.

Thus, a corpus of comments was created in which the source domains are clearly indicated and thus make the analysis of differences between the two domains comfortable and technically possible.

The methodology used was inspired by the study by Tagliamonte & Denis (2008) who carried out a thorough analysis of an instant messaging corpus. I chose four groups of basic formal features of Internet language present in their study: acronyms, emoticons, laughter variants and spellings variants of personal pronouns *I* and *you*.

Subsequently, I studied these features using mainly quantitative statistical method to identify the frequency of the analyzed phenomena in the corpus as a whole and in its two domains (Hastrdlová 2006: 48). Further, I provided the proportional occurrence of the phenomenon in the two domains to compensate the disproportion in size between them. Next, I compared the proportional occurrences of the phenomena in the two domains by counting their rate. I also used the analytical procedure of making comparisons (p. 48). Where relevant, the qualitative description of the phenomena was provided, too, and the phenomena were classified according to their meaning and syntactic functions.

3.2 Acronyms

This section analyzes the frequency of common acronyms in the corpus, describes their usage and presents their internal classification. Also, it looks into the distribution of acronyms in the two domains under investigation and looks for the tendency towards the difference between them.

Figure 7 lists the most frequent acronyms as they were found in the corpus. Overall frequency means the total number of occurrences in the corpus, regardless of the domains. The percentage given in the columns Leisure and Info & Know is the proportional occurrence of a particular acronym in each domain and Ratio is the rate of proportional occurrences of the phenomenon between the two domains. The same applies for all other figures, unless stated otherwise.

Figure 7: Acronyms

Acronym[7]	**Overall Frequency**	**Leisure**	**Info & Know**	**Meaning**	**Ratio**
LOL	67	0,507%	0,073%	Laugh(ing) out loud	87:13
OMG	14	0,131%	0,000%	Oh my God	100:0
BTW	9	0,066%	0,011%	By the way	85:15
WTF	6	0,028%	0,017%	What the f*ck	63:37
ROFL	4	0,038%	0,000%	Rolling on the floor laughing	100:0
LMAO	5	0,038%	0,006%	Laughing my *ss off	87:13
JK	2	0,019%	0,000%	Just kidding	100:0
TOTAL	**107**	**88**	**19**		**89:11**

[7] During my analysis, all variants of acronyms, emoticons and laughter variants were considered and advanced queries were produced to find them. However, only the basic variants are listed in the tables.

LOL is by far the most frequent acronym in the corpus. As Crystal points out (2006: 37), no one can actually tell whether the speaker really laughs when using the acronym, since no study has been conducted so far on whether people do "laugh out loud" when they post LOL (p. 37). However, we can identify a number of instances in which the use of LOL corresponding with its literal meaning is highly probable:

> (1) [8] donjag: u dunt have to be an emmalina-dancing-clone to make it on YouTube! do ur own THANG! (that's me, a brown guy tryin to b black LoLs....)
>
> lonelygirl15: thanks, i am gonna do my own THANG lol :p

I will now focus on the second occurrence of LOL in this exchange. The user Donjag has made a pun which lonelygirl15 found funny and adopted it for her own reply. The fact that LOL is followed by an emoticon yet strengthens the possibility of literal meaning of LOL, similarly as in:

> (2) yulaw: did you get ILM to do your special effects in your video?? no one can do that stuff with there face it had to be camera tricks...lol j/k pretty cool tho :)

Here the emoticon is divided from LOL by a sentence. But again, building on general face-to-face communication experience, people do smile when making jokes to show the utterance is not meant literally. It is therefore plausible to regard this use of LOL as a more or less literal meaning of "laughing out loud". As mentioned above, the actual speaker may or may not laugh out loud, but the listener is very likely to interpret the utterance as if he did.

This issue is related to the absence of kinesic and proxemic features in Internet language. These natural limitations, as far as paralinguistics is concerned, result in using the

[8] The examples are numbered throughout the thesis. See Primary Sources for a complete corpus of comments and for the lists of comments in their original context.

devices which produce analogical effects (Leetaru 2008), of which some acronyms and all emoticons are a part.

When returning once again to the exchange held in (1), there is another occurrence of LOL in the conversation. This one is interesting not only as another instance of joking (and therefore literally interpretable meaning of LOL), but also formally. The plural "s" morpheme is present which is added to make a plural form of LOL. The fact that LOL is nominalized means that it is taken as a (visual) "object" and functions in a similar way as an emoticon. Some other instances of nominalization are listed below:

(3) tzaqriah: damn brittany.....this ghurl got it all. she got da moves, da music, and more importantly, the bang..........lols

(4) FromDaBay: yup i would say you lost your mind lolz

(5} MrsEManning10: Awsome song! finally not a perverted video! lolz <333

(6) AYEMAQ: now my question is how old was ur mom when she visited the Rabbi
oh how old was ur little sister when u took her 2 the great rabbi lolz

Yet, there is even more to LOL from a morphological point of view. Not only do speakers nominalize, but they also verbalize:

(7) FlyingSquirellInFace: 1:58 i lol'd

(8) xmandiex: i lol'd.

The question now arises, what part of speech LOL is anyhow? Formally, it is held that it stands for an interjection (Pullum 2005). However, the nominalization and the visual character of LOL also points to its iconicity, pushing it as if beyond the traditional morphological categories. I think it is not absolutely necessary to draw a distinction between LOL as an acronym expressing conventional interjection, and LOL as an "iconic" paralinguistic feature similar in its nature to emoticons. It should however be noted that this difference exists and that it is manifested in the language of YouTube video comments.

When comparing the use of LOL to Tagliamonte & Denis study (2008), it is evident that not all possible uses of LOL are represented in my corpus. The authors, for instance, point out that LOL is sometimes "used by participants in the flow of conversation as a signal of interlocutor involvement, just as one might say mm-hm in the course of a conversation" (p. 11). It is quite logical that I have not encountered this use of LOL, as there is actually no conversation in synchronous sense in the video comments.

In examples (9)-(11) we can see another productive acronym from the corpus: the interjection OMG (Oh my God!). As I have shown, LOL can sometimes be interpreted as an emoticon or as a phatic filler (Baron 2004: 23). In contrast, I have found no instance of OMG functioning in any other way than as an ordinary interjection abbreviated for the sake of saving keystrokes. Indeed, in all the utterances listed below, OMG works as an emotional element added to the sentence (or as a holophrastic expression). There is no indication of it being nominalized or verbalized either.

(9) burv: OMG!! You sound the same as Marizpan from homestarrunner.com!

(10) XxrunawaytimmyxX: OMG YES! THE BEST CHARACTER IN THE XMEN SERIES IS IN THIS MOVIE!!

(11) joeyfalcone: omg its a moive..wait till it comes out before you judge it mr comicbook nerd...

By contrasting the two most frequent acronyms in my corpus both formally and notionally, I wanted to point out the fact that the group of acronyms is not homogenous. The acronyms, such as LOL, ROFL, LMAO are among "laughter variants" (Tagliamonte & Denis 2008: 11, cf. 3.4) and they differ from other acronyms not only in their meanings, but also in their morphological and syntactic qualities. Also, many acronyms come as swearwords, the most frequent being WTF, which can also be identified in the corpus. Syntactically, they can be either a holophrastic expressions or parts of a sentence. The meaning can vary depending on syntactic positions:

(12) snake289: wtf

In (12) WTF means something to the effect of „What on Earth is this?“. It has a strong evaluative meaning and indicates speaker's negative or at least not-entirely-positive attitude to the video. In contrast, WTF within a sentence or clause stands for "What (on Earth)":

(13) semi801: wtf is wrong with you... lol

(14) gosciu555: wtf are you talking about u bias cunt. terrorism is state sponsored. you're regime supports it read victor ostrovski

However, sometimes WTF syntactically stands within a sentence but has an evaluative meaning (similar to the one identified in (12)):

(15) vippieroxs: wtf weres sound?

The fact that vippieroxs's comment can be interpreted as "What the f*ck! Where is sound?" can be attributed to the absence of punctuation. Thus I have shown that not only do acronyms vary as a group, but also their meanings can vary depending on their syntactic positions.

There certainly are other acronyms which can be found on YouTube and do not appear in my corpus. However, my aim was to list the most frequent ones and attempt at their classification:

Figure 8: Classification of the Most Frequent Acronyms in the Corpus

Laughter Variants	LOL ROFL LMAO
Exclamations	OMG
Neutral acronyms	BTW, JK
Swearwords	WTF

Since the number of acronyms other than LOL is too low for a representative comparison between L and IK domains, I only compared the use of LOL between the domains. As stated in Figure 9, 87% of all LOL occurrences can be found in L domain. This observation is in correspondence with the hypothesis which supposes different stylistic properties of comments depending on the category (or, more generally, the domain) of the videos.

The fact that LOL appears more often in the L domain signalizes a potentially lower degree of formality of L domain compared to IK domain. However, judging the level of formality by only one informal feature would result in significant simplification. Therefore, the question of formality and informality will be left open until the final section of this part where the summary of all findings and their discussion will be made.

Another interesting issue concerning LOL is its overall frequency in the corpus, which may seem relatively low: there are 67 occurrences in the whole corpus which only equals 0,234% of total word count. Nevertheless, when searching for the words with similar frequency, the program returns the following words: *would, should, why, only, really, think* and *know*. We can see that although the percentage is (in numbers) quite low, LOL is as frequent as basic modal verbs, intensifiers and full verbs. I hold that Tagliamonte & Dennis

(2008) did not emphasize this fact enough, probably because they were arguing against the overblown role of LOL as a “linguistic ruin” of the English language. In my corpus 1 in 18 comments on average contains LOL, in the L domain it is 1 in 13 comments, which means that LOL is a fairly productive feature.

The observations made in this part show that the acronyms are among important formal aspects of the language of YouTube video comments, and that the most frequent acronym by far is LOL, followed by OMG and BTW. It has also been demonstrated that the acronyms serve various purposes (expressing emotions and laughter, saving keystrokes, swearing) and that they can play multiple roles in syntax, with LOL being nominalized and verbalized. The use of acronyms in L domain evidently tops their use in IK domain, with the overall frequency similar to that of basic modal verbs, intensifiers and full verbs.

3.3 Emoticons

Another specific formal aspect of Internet language, including YouTube video comments, is emoticons. Their origin on the Internet dates back to 1982 when Scott Fahlman introduced the sequence of :-) to mean joke. The entire Bboard thread has been retrieved and can be read by anyone on the Internet. The post in which Fahlman proposes the use of :-) reads as follows (Fahlman 1982):

> I propose that the following character sequence for joke markers:
>
> :-)
>
> Read it sideways. Actually, it is probably more economical to mark
>
> things that are NOT jokes, given current trends. For this, use
>
> :-(

It is necessary to add that the idea of marking “things that are not jokes” by :-(had been originally proposed only for the very discussion thread in which :-) originated. The majority of messages were not meant seriously and so it would be “more economical to mark things that are not jokes“. The later development and adoption of emoticons for the widespread use adopted both smiley and frowny face as an expression of writer’s mood or facial expression. In other words, while the original idea was to have only :-) for expressing a marked utterance (leaving :-(for a neutral one), the contemporary use of both the basic emoticons always indicates the speaker’s mood or emotion (DPF 2006).

There are many types of emoticons, but all of them are considered to be a simple form of ASCII art. ASCII stands for American Standard Code for Information Interchange and it consists of 95 printable characters which are combined to create pictures (Wikipedia 2009a). The simplest classification of emoticons is probably the one based on horizontal/vertical dichotomy. The horizontal emoticons have to be read with the head tilted to one side (usually to the left), the typical example being :-) and :-(. These are widely used in the Western culture and are used worldwide. The Eastern cultures have their own style of emoticons which can be read without tilting one’s head to the side and because of this fact they are called “verticons” (Wikipedia 2009c). A typical example is (*_*). Outside ASCII there are countless graphic forms. As can be seen in the following examples, both horizontal and vertical emoticons are represented in my corpus:

(16) matty123986754: straight away i have a feeling your going to be very big in YouTube :)

(17) NoYouAre: Love that song :-)

(18) xin0: and you believe Wikipedia:/?

(19) tittenfisch: 08:50 How right you were :(

(20) Crater300: oooh yeah, you're so 'Crazy'.......... o.O

(21) spiderman1289: Oh my! Gambit....and i guess the white-fuzzy-turning lady could be Mrs Emma "i'm-sexy" Frost...awsome....MARVEL i luv thee ^^

In (16) there is a perfect example of what Crystal (2006: 39) emphasizes when talking about the use of emoticons: there are some instances of unmarked utterances where the absence of an emoticon could be misinterpreted as marked. The example (16) is exactly the case where the absence of an emoticon might lead to ambiguity, i.e. to the addressee interpreting matty123986754's utterance as ironic. It is true that the emoticon here primarily represents a paralinguistic feature, that of a positive, encouraging tone and smile. However, as the wording of the utterance as such is neutral, the emoticon seems to be added to prevent an ambiguous meaning.

As regards the frequency of emoticons, there are no studies describing their distribution on YouTube. I have therefore studied this phenomenon in detail, as it can tell more about the style of the comments and about formality and informality of the genre. If we slightly oversimplify, the more emoticons there is in the text, the less formal the text is.[9]

I mentioned two basic types of emoticons above, namely horizontal (or Western) type, and vertical (or Eastern) type. Both of these are represented in my corpus, with Eastern type being negligible, as Figure 9 clearly documents:

Figure 9: Types of Emoticons

	Horizontal Type	**Vertical Type**
Overall frequency	132	10
Proportion of total	93%	7%

[9] Huffaker, D. A., S. L. Calvert (2005). 'Gender, Identity, and Language Use in Teenage Blogs'. *Journal of Computer-Mediated Communication* 10, 2, article 1.Cited in Leetaru (2008: 873).

The most "frequent" verticon is ^^, which is an equivalent to horizontal :-). Its full textual representation should be (^_^), but there is no such occurrence in the corpus. It has probably something to do with the economy of writing, i.e. with saving keystrokes, and with the adoption of Eastern type when discussing Western videos on Western Web site. The other two examples of verticons' occurrence are ^- (representing ;-)) and o.O (representing :-O)[10].

Horizontal emoticons are frequent enough to be classified and listed in tables. I studied both positive and negative emoticons, as I was analyzing their mutual rate in the two domains as well. The first overview presents horizontal emoticons which express positive emotions and facial expressions. The basic forms in the table express smiling, grinning, joking and winking, respectively, with all frequent variants considered. The second overview presents horizontal emoticons which express negative or non-positive emotions. The basic forms in the table express the state of being sad, bored/annoyed, confused/embarrassed and shocked/surprised, respectively. Again, including all frequent variants:

Figure 10: Positive Emoticons

Emoticon	**Overall Frequency**	**Leisure**	**Info & Know**	**Ratio**
:-)	61	0,563%	0,006%	99:1
:-D	37	0,338%	0,006%	98:2
:-P	16	0,131%	0,011%	92:8
;-)	4	0,038%	0,000%	100:0
TOTAL	**118**	**114**	**4**	**98:2**

[10] See Wikipedia 2009e for more details.

Figure 11: Negative Emoticons

Emoticon	**Overall Frequency**	**Leisure**	**Info & Know**	**Ratio**
:-(	10	0,075%	0,011%	87:13
:-/	2	0,019%	0,000%	100:0
:-S	1	0,009%	0,000%	100:0
:-O	1	0,009%	0,000%	100:0
TOTAL	**14**	**12**	**2**	**91:9**

These tables are quite disproportional in that L domain contains almost all the emoticons present in the corpus (with the Ratio of 97:3), which is quite enormous a difference. This fact can be partly attributed to the nature of videos which are “funny” in L domain and “serious” in the IK domain. However, the disproportion itself is probably caused by higher degree of formality in IK domain.

Another interesting aspect of emoticons in the corpus is the disproportion of positive and negative emoticons (89% of all comments are positive). This definitely does not mean that YouTube users are always positive when discussing videos. It just evidences that the they do not frequently *use* emoticons when being critical or angry with the video or other users’ comments.

In this section, the emoticons in the corpus were analyzed. Their historical background was presented and they were classified according to their position on the line as horizontal and vertical. The special emphasis was placed on the horizontal ones which are prevalent in the corpus. It was found out that emoticons appear almost exclusively in L domain and that they are predominantly positive.

3.3 Laughter Variants

The third group of widely used Internet language features has been to a great extent discussed already because textual representation of laughter sounds can have many forms which partly overlap with acronyms and emoticons. Tagliamonte & Denis (2008: 12) call them laughter variants. These include acronyms, such as LOL, ROFL and LMAO (cf. 3.2), various forms of interjections *haha* and *hehe* and some emoticons (:-D, :-P, possibly even :-)). Because the distribution of emoticons and acronyms has already been treated in detail in previous sections, I will only focus on the conventional interjections in the present one.

Tagliamonte & Denis (2008) found out that *haha* is the most productive laughter variant in their IM corpus (p. 11) with thousands of instances in their million word corpus. In my corpus, there are only 34 *haha* and 7 *hehe* variants (see Figure 12) which means this feature is not the most productive of all laughter variants. *Haha* and *hehe* naturally have various forms – they are often mistyped, repeated and also accompanied by emoticons:

> (22) secondlyquixotic231: Britney is back. You can't doubt it. You like it. hahaha : P
>
> (23) kathyta54: me too, I just listen rock and metal, gothinc and rock but I like britney and this son (but is a secret) hahahahahha....... XD......
>
> (24) EssenceOfTruth: Its it cool that his name is Remy and Gambit's name is "remy"? Heheh! Totally awesome!

As shown in the figure below, the distribution of these interjections in individual domains is again disproportional, true, but still the difference in the rate of their proportional occurrences is the lowest from all three groups analyzed so far. The fact that we have 97:3 rate with emoticons, 89:11 with acronyms and only 84:16 with laughter interjections can, but need not, be a coincidence.

This tendency would make perfect sense because the use of both emoticons and acronyms is still almost exclusively associated with Internet language, whereas conventional laughter interjections can be found elsewhere, including fiction. They can, therefore, be perceived as the most conventional form of expressing laughter from the three groups in question, and, because of that, as the most formal from the three (which would again signal that IK domain tends to be more formal in style). Also, their deciphering does not need any special knowledge and all users are familiar with their meaning, regardless of their computer literacy.

Figure 12: Laughter Interjections

Emoticon	**Overall Frequency**	**Leisure**	**Info & Know**	**Ratio**
haha	34	0,235%	0,050%	82:18
hehe	7	0,056%	0,006%	91:9
TOTAL	**41**	**34**	**7**	**84:16**

In this section, the ways of expressing laughter in the video comments were treated with reference to the previous sections which also contained some of the laughter variants. The observations I made include the fact that, unlike in IM corpus by Tagliamonte & Denis (2008), conventional laughter interjections (*haha*, *hehe*) are not the most productive variant in the case of YouTube. In fact, they have the lowest representation of all the three investigated ways of expressing laughter. It has also been observed that with regard to the rate of proportional occurrences, the conventional laughter interjections form the closest gap between L and IK domain from all the three variants. This fact coincides with the interjections being the most conventional of them.

3.4 Spelling Variants of Personal Pronouns

The last formal linguistic phenomenon to deal with forms a part of a fairly wide and complex domain of orthography and spelling. There are countless spelling innovations and deviations in Internet language, such as capitalization, absence of punctuation, letter and number homophones (Davis & Mason 2008: 635, Gong & Ooi 2008: 925). Number of these alternative spellings emerges from users' needs and creativity. The traditional orthographic norms are routinely violated at the expense of saving keystrokes, but also of shifting the written words closer to their phonetic representations (Gong & Ooi 2008: 928). The physical constraints thus paradoxically initiate the creative processes in the users' minds and allow for experimenting with the language (p. 928). In a sense, spelling as such encompasses all previous sections of my thesis.

In order to keep the size of this text in reasonable limits, this section only deals with the distribution of personal pronouns variants in the corpus. This choice was made because personal pronouns are among the most used words in speech. They are universal and appear in all kinds of texts, which makes them ideal for a linguistic analysis.

Before the analysis of actual spelling variants, the proportional distribution of grammatical persons in the corpus has been studied and compared to Tagliamonte & Denis's (2008) data. Although this comparison is not directly related to the spelling deviations, it is relevant, as it discloses the frequency of personal pronouns *I* and *you* in the corpus. These pronouns will subsequently be the objects of analysis and it is relevant to ask not only whether they have fewer or more variants, but also whether they are represented equally in other types of communication media.

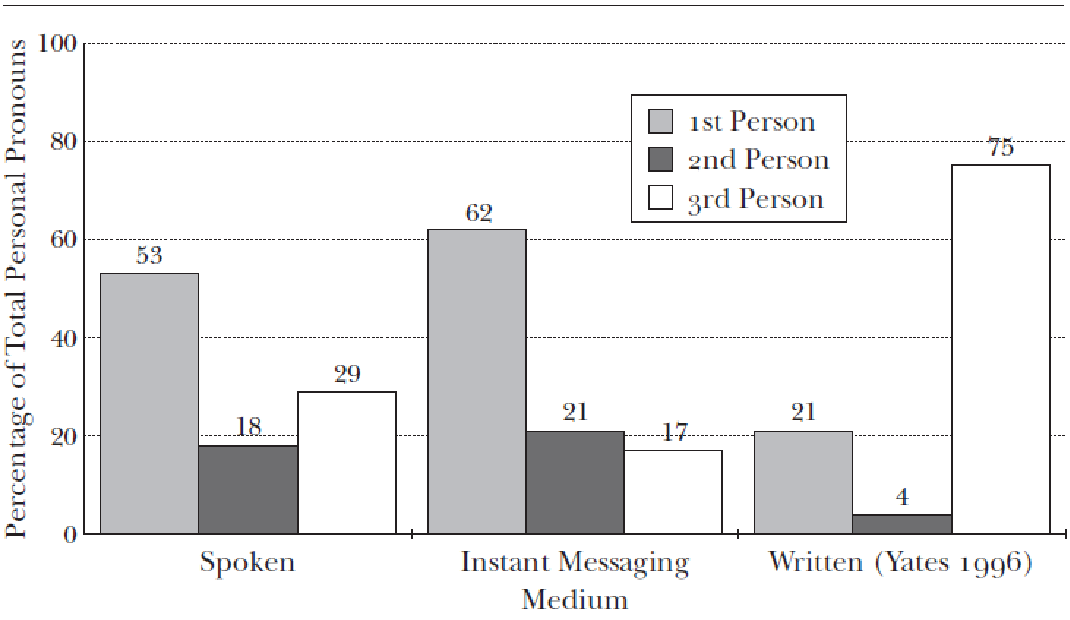
Figure 13: The Proportional Distribution of Personal Pronouns in YouTube Video Comments

	1st person	2nd person	3rd person
Proportion of total	36%	30%	34%

Figure 14: Tagliamonte & Denis (2008: 16)

FIGURE 4

Distribution of Personal Pronouns by Medium



We can see that my corpus is characterized by a distributional pattern dissimilar to any of the three mediums studied and compared by Tagliamonte & Denis. Technically speaking, YouTube video comments seem to be the most similar to speech, as this medium has relatively the most equal proportion of grammatical persons. Still, the difference is really significant and taking this rough similarity as granted would lead to the dead ends of simplification.

The answer to the question of a surprisingly unique pattern can be found in the content of the discourse (see Tagliamonte & Denis 2008: 15) which corresponds with (and is determined by) the content of the videos. It is the selected videos that determines what is discussed and referred to, and how.

The most frequent discussion topics include, depending on the individual videos, Britney Spears and her new song; lonelygirl15 and her video blog (including interaction with her and the users); the upcoming movie Wolverine and its main protagonists; vision of contemporary students and the clash between old and new ways of education; and children digging tunnels from Egypt to Gaza. In the case of lonelygirl15, Wolverine and the vision of students, the video itself is widely discussed in terms of its artistic quality.

It was observed that the video comments are unique in terms of grammatical persons' representation. The analysis of spelling variants of *I* and *you* which is now to be carried out will cast more light at the question of formality brought up several times in this thesis. The figure below is an overview of the research into the basic variants of personal pronouns I and you proportionally within each domain.

Figure 15: Variants of *I* and *you*

Variants	**Leisure**	**Info & Know**	**TOTAL**
I v. i	59:41	65:35	**62:38**
you v. u	85:15	80:20	**82:18**

From the data, it is evident that it is again L domain which is more likely to contain marked variants. Especially in the case of *I* and *i*, the rate is quite balanced, being slightly lower than 6:4. Quite unusually, the IK domain has a very similar tendency concerning 1st person pronoun *I*. The ratio differs from L domain only by 6% on each side, which is really a rare observation when compared to the previous data.

In the case of *you* and *u,* it is the first (and the last, too) time when IK domain carries more marked features than L domain. This is really a surprising situation given that throughout the whole analysis, the IK domain was always far behind the other one in the number of the analyzed features. There can be several reasons for this anomaly.

First, the frequency of personal pronouns in the course of interaction is so high that even in the IK domain, which was always more resistant towards CMC specific features, the users feel the need to save keystrokes in order to communicate promptly. Second, the use of *i* and *u* has already been "codified" in Internet language and it is considered unmarked by the significant group of users. Third, the average comment length, which is 15 words in L domain and 36 words in IK domain, plays the important role, pressing the IK comment posters to save time more than those in L domain. Fourth, the frequent use of personal pronouns in other forms of CMC, and also in SMS, for instance, makes the use of *you* and *I* highly automatic, unlike the use of acronyms and emoticons.

Most probably, the answer lies in between the four possibilities, varying individually. In fact, the specific user's usage is especially important this time, as the scholars have proved that the use of personal pronouns variants is consistent and that little variation is common in individual users (Tagliamonte & Denis 2008: 14).

3.5 Discussion

The four major groups of formal linguistic features were analyzed in the practical part: acronyms, emoticons, laughter variants and variants of personal pronouns *I* and *you*. In this section, the observations will be summarized, compared to the hypotheses and related to the theoretical part. Also, some general properties of the corpus needed for its full description will be provided, which have not been fitting for any of the previous sections so far.

As the figure below demonstrates, the research into the corpus of YouTube comments has shown that all of the above mentioned linguistic phenomena are its substantial components. These typical groups of CMC forms are ordered by their overall frequency in the corpus. Their proportion of total word count is provided, and so is their relative proportion in individual domains and the rate of proportional occurrences between them. Several tendencies can be identified in these features which tell us more about how YouTube users behave linguistically.

Figure 16: Summary

Feature	**Overall Frequency**	**Proportion of total word count**	**Leisure**	**Info & Know**	**Ratio**
i and u variants	398	1,392%	1,877%	1,104%	63:37
Emoticons	142	0,497%	1,276%	0,033%	97:3
Acronyms	107	0,374%	0,826%	0,106%	89:11
Laughter variants[11]	41	0,143%	0,291%	0,056%	84:16
TOTAL	**688**	**2,407%**	**4,269%**	**1,299%**	**77:23**

First, the linguistic behaviour of the users conforms to the fact that commenting and watching the videos is mediated by computer. All the forms represented in the table are frequent enough to be considered typical of the video comments language. *i* and *u* variants, even when treated separately, are 12th and 38th most frequent words, respectively, in the corpus, thus outnumbering the use of words, such as: *get, good, song, love, can, movie, think*, and the like.

The above mentioned comparison of the frequency of personal pronouns variants to the frequency of lexical words must be taken very cautiously, as grammatical and lexical words are difficult to compare, if at all. However, given that *i* and *u* are both marked variants, their high frequency in the corpus signalizes that the nature of video comments corresponds with the general assumption that all kinds of CMC contain, to a certain extent, similar deviations from the norm, typical of Internet language.

[11] Only *haha* and *hehe* are mentioned, LOL, ROFL, LMAO and emoticons are not included.

Emoticons as a whole appear 142 times in the corpus, which means that, were they treated regardless of their concrete representation, they would be similarly frequent as *just, more, will,* and as *u* variant of *you*. Therefore, emoticons are also among the productive features of video comments.

Similarly, the acronyms are productive enough to be regarded as typical of the comments. However, the use of one acronym, LOL, is dominant and other instances are negligible, which is similar to :-) and :-D in emoticons. We must therefore be careful when judging the productivity of the group as a whole and the productivity of its individual forms, as there are only some of them which are really dominant. This fact is taken into account in Figure 17 which lists the eight most frequent forms.

Figure 17: The Eight Most Frequent Forms

Form	**Overall frequency**	**Proportion of total word count**
i variant	275	0,962%
u variant	123	0,430%
LOL	67	0,234%
:-)	61	0,213%
:-D	37	0,129%
haha	34	0,119%
:-P	16	0,056%
OMG	14	0,049%

The laughter variants do also come roughly within 100 most frequent words, but they are more specific, as they overlap with acronyms and emoticons and their occurrence in Figure 16 only includes conventional *haha* and *hehe* interjections. When the respective emoticons and acronyms, as described in 3.2 and 3.3, were added (including :-)), the overall frequency of laughter variants would rise to 231 occurrences and the proportion of total word count would be 0,808%. Therefore, even the laughter variants are a prevalent feature of YouTube video comments. By providing this brief summary, I have demonstrated that the choice of typical CMC features was fortunate in that all groups of the selected features are productive in the corpus.

Further, the results of my analysis will be compared to my hypotheses. I argued that the ways in which the audience discusses the videos are likely to develop in different directions, depending on the content of the videos, i.e. on the L and IK domains. Likewise, the formality and informality in these domains was expected to differ because the stylistic properties of the language present in videos themselves are different, and I assumed that this fact would be reflected in the language of the comments.

The observations relevant for the first hypothesis include the average length of the comments, the diversity of the language (total number of unique words) and the use of investigated formal features. Figure 18 shows that the ways in which the users comment on the videos do substantially differ in the two domains.

Figure 18: Domains Comparison

	Leisure	**Info & Know**
Total number of comments	728	498
Total number of words	10 658	17 931
Average comment length	15 words	36 words
Unique words	2339	3635
Typical features distribution		
Acronyms	0,826%	0,106%
Emoticons	1,276%	0,033%
Laughter variants	0,291%	0,056%
i and u variants	1,877%	1,104%
TOTAL	**4,269%**	**1,299%**

Although the total number of comments in L domain is higher than in IK domain (see section 3.1. for details), the total number of words is approximately 1.7 times higher in IK than in L domain, and there are more than twice as long comments in IK than in L domain. Also, the diversity of language used for commenting the videos is different, as there are only 2339 unique words in L domain, but 3635 unique words in IK domain. The distribution of the surveyed features was mentioned already in this section. It is evident that the difference between the two domains is really vast. Therefore, this part of hypothesis concerning the difference in the way of discussing videos is confirmed, at least as regards the linguistic features measurable by statistical quantitative analysis.

The other hypothesis concerns formality and informality of the domains. When deciding about the level of formality, I draw on Leetaru's (2008) claim that the more tones (*haha, hehe*) and senses (emoticons, acronyms) conveyed through CMC, the more informal its language is. Quite naturally, I also include deviations from orthographic norms (*i* and *u* variants) which represent another marker of an informal style.

When looking again into the typical features distribution in Figure 18, we can see that the level of formality, indeed, differs substantially between the domains, with all groups of formal features being much more frequent in L domain. Therefore, L domain can be regarded as more informal and IK domain as more formal of the two. This observation points to the interdependence of video content and linguistic properties of the comments mentioned in 2.4.

However, no universally valid generalization as regards the formality of the comments as a whole is possible, except for claiming that the language of YouTube video comments is fairly diverse and that it yields a number of surprising internal differences which are related to its various contexts.

Further, the issue of synchronicity and asynchronicity was discussed in the theoretical part and some basic classification was made, placing video comments among asynchronous types of CMC. However, it should be noted that the video comments also bear some chat-like, i.e. synchronous-like, features, such as the comments disappearing from user's sight once their number does not fit one page and the comments being really short, with the average of only 23 words per comment (15 in L and 36 in IK domain).

In the very end of this section, it is necessary to draw the reader's attention to the limitations of my research and to make some suggestions as far as the future research on YouTube video comments is concerned. My corpus has been devised to contain a relatively small number of comments because I aimed at a basic insight into the issue, not at a complex quantitative analysis with hundred thousands of words. Also, the two domains I defined contain much more diverse material than is analyzable in a limited space of this thesis (see the list of video categories in 3.1). Therefore, the limiting factors include the small size of the corpus and the reduction of the diversity of comments. And finally, as my research draws exclusively on formal features and disregards more complex problems, such as turn-taking and complex discourse analysis, it would be highly beneficial to study these in the future, as they can help open numerous ways to approach and understand the language of video comments.

4. Conclusion

YouTube video comments are an asynchronous type of computer-mediated communication which allows the users of the Web site to respond to the viewed videos.

Their asynchronous nature consists in the delay between the time of posting the comment and the time of it being read by other users. It was, however, found out that they also come with synchronous-like characteristics which include the quick disappearance of the comments to the frequently commented videos from the first page, and the low average number of 23 words per comment in the corpus under investigation.

Communication through video comments is mediated by the YouTube Web site and can be best described by using HCI theoretical communication model, WebCom. The designer of the page is a frame setter who does not participate in the communication, except for introducing new features of video comments platform. The communication thus takes place solely among the users who come from various socio-cultural contexts and who create a user-generated content, i.e. the text of the comments. It, as an object of communication, interacts with other users who enter the communication to pursue their intentional outcomes, and leads to the creation of the actualized outcome, or effect, of the communication.

The comments can serve numerous functions. Apart from surveillance, entertainment, interpersonal utility and parasocial interaction, which are common for mass media in general, they can add eventness to the video watching by creating the virtual audience. Also, their peculiar style can stimulate users' linguistic awareness.

The users can utilize some extra features of the comments, including their rating, editing and deleting. By rating the comments the users execute a self-moderating activity, as the site implicitly displays only the comments of a certain minimal value. The users also need to observe a 500-character limit per comment and respect YouTube's anti-spam policy.

All these theoretical aspects of video comments are relevant for their linguistic analysis, as they reflect in their language. When being mediated by computer, the Internet and a concrete Web site with various features available, the language changes in a unique way to compensate the shortcomings of the given type of communication.

My analysis was directed at the selected formal features of the language of comments which included acronyms, emoticons, laughter variants and spelling variants of personal pronouns *I* and *you*.

The analysis showed that the distribution of these features is dependent on the videos to which the comments were posted. The comments to the videos from the Leisure domain are shorter, contain fewer unique words and more formal features under investigation. The higher frequency of these forms in the Leisure domain also proved that the domains differ in the level of formality, with the Leisure domain more informal than the Information & Knowledge domain.

All formal features in question were found productive in the corpus, with spelling variants of *i* and *u* being the most frequent. The second most productive group of the features was emoticons, followed by acronyms. The laughter variants *haha* and *hehe* were the least frequent from the four when treated separately. However, since laughter variants also include emoticons and some acronyms, even this group was observed to be sufficiently represented in the corpus.

The most frequent unique forms in the corpus include *i* and *u* variants, the acronym LOL, emoticons :-) and :-D and a laughter interjection *haha*. Except for *i* and *u* variants, all these forms express positive emotions. The notable frequency of these in the corpus is another key tendency observed during the analysis.

The four groups of analyzed formal features ordered by the rate of their proportional occurrences between the individual domains reveal that the greatest disproportion between the domains can be seen in emoticons and acronyms. The conventional laughter interjections are less disproportionate than the two and the spelling variants *i* and *u* are by far the most proportionate, with the rate of 63:37.

Many of these observations demonstrate that YouTube is diverse both in its audiovisual and linguistic content and that not only in different types of CMC, but also within its specific type, the Internet language is not uniform and needs to be studied in relation with all its contexts.

Primary sources

‘Britney Spears - Circus (Official Music Video With Lyrics)’. 2009. *YouTube: Broadcast Yourself.* 1 April 2009 <http://www.youtube.com/watch?v=nGnmRUVf-QI>

‘First Blog / Dorkiness Prevails’. 2006. *YouTube: Broadcast Yourself.* 1 April 2009 <http://www.youtube.com/watch?v=-goXKtd6cPo>

‘Wolverine Movie Trailer 2009 (X-Men Origins) in HD’. 2009. *YouTube: Broadcast Yourself.* 1 April 2009 <http://www.youtube.com/watch?v=Hiemc14iifw>

‘A Vision of Students Today’. 2007. *YouTube: Broadcast Yourself.* 1 April 2009 <http://www.youtube.com/watch?v=dGCJ46vyR9o>

‘Gaza Tunnels - Israel/Palestine’. 2008. *YouTube: Broadcast Yourself.* 1 April 2009 <http://www.youtube.com/watch?v=f9IL86T6Nc8>

Note: The comments included in the corpus are archived in their original context at the pages with the above listed videos (they represent approximately 250 oldest comments to the videos). Also, they are archived at the Language of YouTube Video Comments Homepage <http://www.linguistics.euweb.cz> which was created for the purpose of this thesis. The corpus of comments and the software used for the analysis is attached as a CD-ROM to the cover of the thesis. It can also be found at <http://www.linguistics.euweb.cz>.

References

Andy and Dave [surnames unknown] (2009). ‘Ten YouTube Comments Translated into Standard English’. *DelSquacho*. 15 March 2009 <http://www.delsquacho.com/articles/ten-youtube-translations.php>

Axtman, Kriss (2002). ‘”r u online?”: The Evolving Lexicon of Wired Teens’. *The Christian Science Monitor*. 3 March 2009 <http://www.csmonitor.com/2002/1212/p01s01-ussc.html>

Baron, Naomi S. (2004). ‘See You Online: Gender Issues in College Student Use of Instant Messaging’. College of Arts & Sciences, American University. 17 January 2009 <http://www1.american.edu/tesol/Baron-SeeYouOnlineCorrected64.pdf>

Berens, Brad (2006). ‘What the Comments on YouTube Really Mean’. *Mediavorous*. 2 April 2009 <http://mediavorous.com/archives/what-the-comments-on-youtube-really- mean>

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

December, John (1997). ‘Notes on Defining of Computer-Mediated Communication’. *CMC Magazine* Vol. 4, No. 1, January, 1997. 24 January 2009 <http://www.december.com/cmc/mag/1997/jan/december.html>

December, John (1996). ’Units of Analysis for Internet Communication’. *Journal of Computer-Mediated Communication* Vol. 1, No. 4, March, 1996. 13 December 2008 <http://jcmc.indiana.edu/vol1/issue4/december.html>

DPF = Knowles, Elizabeth (ed.) (2006). ‘Emoticon’. *A Dictionary of Phrase and Fable..* Oxford: Oxford UP. *Oxford Reference Online*. Masaryk University. 5 April 2009 <http://www.oxfordreference.com/views/ENTRY.html?subview=Main&entry=t214.e2412>

Fahlman, Scott (1982). ‘Original Bboard Thread in which :-) was proposed’. Language Technologies Institute and Computer Science Department, Carnegie Mellon University. 27 November 2008 <http://www.cs.cmu.edu/~sef/Orig-Smiley.htm>

Godsk, Mikkel, Anja Bechmann Petersen (2008). ‘WebCom as a Model for Understanding Website Communication’. In: Kelsey, Sigrid, Kirk St. Amant (eds.). *Handbook of Research on Computer Mediated Communication* (Vol. 1). Hershey: Information Science Reference, 379-400.

Gong, Wengao, Vincent B. Y. Ooi (2008). ‘Innovations and Motivations in Online Chat’. In: Kelsey, Sigrid, Kirk St. Amant (eds.). *Handbook of Research on Computer Mediated Communication* (Vol. 2). Hershey: Information Science Reference, 917-933.

Hastrdlová, Šárka (2006). ‘Language of Internet Relay Chat’. Dissertation. Masaryk University.

Jordan, Robert (2008). ‘Preparing Participants for Computer Mediated Communication’. In: Kelsey, Sigrid, Kirk St. Amant (eds.). *Handbook of Research on Computer Mediated Communication* (Vol. 1). Hershey: Information Science Reference, 25-33.

Kimmerle, Joachim, Ulrike Cress (2008). ‘Knowledge Communication with Shared Databases’. In: Kelsey, Sigrid, Kirk St. Amant (eds.). *Handbook of Research on Computer Mediated Communication* (Vol. 1). Hershey: Information Science Reference, 424-435.

Lády, Tomáš (2007). ‘České diskusní servery’. BA Thesis. Masaryk University.

Lange, Patricia G. (2008). ‘Terminological Obfuscation in Online Research’. In: Kelsey, Sigrid, Kirk St. Amant (eds.). *Handbook of Research on Computer Mediated Communication* (Vol. 1). Hershey: Information Science Reference, 436-450.

Leetaru, Kalev (2008). 'Instant Messaging as a Hypermedium in the Making'. In: Kelsey, Sigrid, Kirk St. Amant (eds.). *Handbook of Research on Computer Mediated Communication* (Vol. 2). Hershey: Information Science Reference, 868-882.

Marcus, Sara Rofofsky (2008). 'IM's Growth, Benefits, and Impact on Communication'. In: Kelsey, Sigrid, Kirk St. Amant (eds.). *Handbook of Research on Computer Mediated Communication* (Vol. 2). Hershey: Information Science Reference, 804-814.

Meunier, Bryson (2008). 'YouTube Video Keyword Research and Characteristics of Popular YouTube Queries'. *Natural Search & SEO Blog*. 14 February 2009 <http://www.brysonmeunier.com/youtube-video-keyword-research-and-characteristics-of-popular-youtube-queries>

O'Connor, Amanda (2005). 'Instant Messaging: Friend or Foe of Student Writing?' *New Horizons for Learning*. 21 March 2009 <http://www.newhorizons.org/strategies/literacy/oconnor.htm>

Pullum, Geoffrey K. (2005). 'English in Deep Trouble?' *Language Log*. 17 November 2008 <http://itre.cis.upenn.edu/~myl/languagelog/archives/001829.html>

Rivens Mompean, Annick (2003). 'Electronic English, Oral or Written English?' In: Posteguillo, Santiago, et. al (eds.). *Internet in Linguistics, Translation and Literary Studies*. Castello de la Plana: Publicacions de la Universitat Jaume I, 273-289.

Shankland, Stephen (2008). 'Feature or Google's sense of humor? Audio tool speaks your YouTube comments'. *Digital Media – CNET News*. 1 March 2009 <http://news.cnet.com/8301-1023_3-10062213-93.html>

Šmahel, David (2003). *Psychologie a internet: děti dospělými, dospělí dětmi*. Praha: Triton.

Tagliamonte, Sali A., Derek Denis (2008). 'Linguistic Ruin? LOL! Instant Messaging and Teen Language'. *American Speech*, Vol. 83, No. 1, Spring 2008 [viewed online, original pagination not provided]. American Dialect Society.

Thurlow, Crispin (2006). 'From Statistical Panic to Moral Panic: The Metadiscursive Construction and Popular Exaggeration of New Media Language in the Print Media'. *Journal of Computer-Mediated Communication*, Vol. 11, No. 3, April 2006. 12 February 2009 <http://jcmc.indiana.edu/vol11/issue3/thurlow.html>

Webopedia (2009). 'CMC'. *Webopedia Computer Dictionary*. 17 December 2008 <http://www.webopedia.com/TERM/C/CMC.html>

Weder, Mirjam (2008). 'Form and Function of Metacommunication in CMC'. In: Kelsey, Sigrid, Kirk St. Amant (eds.). *Handbook of Research on Computer Mediated Communication* (Vol. 2). Hershey: Information Science Reference, 570-586.

Wesch, Michael (2008). 'YouTube Statistics'. *Mediated Cultures*. 26 January 2009 <http://mediatedcultures.net/ksudigg/?p=163>

Wikipedia (2009a). 'ASCII Art'. *Wikipedia: The Free Encyclopedia*. 7 January 2009 <http://en.wikipedia.org/wiki/ASCII_art>

Wikipedia (2009b). 'Internet Slang'. *Wikipedia: The Free Encyclopedia.* 7 January 2009 <http://en.wikipedia.org/wiki/Internet_slang>

Wikipedia (2009c). 'Emoticon'. *Wikipedia: The Free Encyclopedia.* 7 January 2009 <http://en.wikipedia.org/wiki/Emoticon>

Wikipedia (2009d). 'YouTube'. *Wikipedia: The Free Encyclopedia.* 7 January 2009 <http://en.wikipedia.org/wiki/YouTube>

Woletz, Julie D. (2008). 'Digital Storytelling from Artificial Intelligence to YouTube'. In: Kelsey, Sigrid, Kirk St. Amant (eds.). *Handbook of Research on Computer Mediated Communication* (Vol. 2). Hershey: Information Science Reference, 587-601.

Yilmaz, Yucel (2007). 'Collaborative Dialogue During Tasks in Synchronous Computer-Mediated Communication'. Dissertation. Florida State University. 16 December 2008.

<http://etd.lib.fsu.edu/theses/available/etd-05072008-104352/unrestricted/YilmazYDissertation.pdf>

YouTube (2009a). 'Getting Started: Comment Limits'. *YouTube: Broadcast Yourself.* 29 January 2009 <http://help.youtube.com/support/youtube/bin/answer.py?hl=en-uk&answer=66813>

YouTube (2009b). 'Learn More: Changing the Order of Comments'. *YouTube: Broadcast Yourself.* 29 January 2009 <http://help.youtube.com/support/youtube/bin/answer.py?answer=77433&topic=17181>

YouTube (2009c). 'Learn More: Deep Links'. *YouTube: Broadcast Yourself.* 29 January 2009 <http://help.youtube.com/support/youtube/bin/answer.py?hl=en-uk&answer=116618>

YouTube (2009d). 'Stream Definition'. *YouTube: Broadcast Yourself.* 29 January 2009 <http://help.youtube.com/support/youtube/bin/answer.py?answer=57949>

YouTube (2009e). 'Video Comments'. *YouTube: Broadcast Yourself.* 31 January 2009 <http://help.youtube.com/support/youtube/bin/topic.py?hl=en-uk&topic=17179>

YouTube (2009f). 'YouTube Glossary: Groups'. *YouTube: Broadcast Yourself.* 29 January 2009 <http://help.youtube.com/support/youtube/bin/answer.py?answer=95443&topic=13660>

YouTube (2009g). 'Getting Started: Comments on my videos'. *YouTube: Broadcast Yourself.* 29 January 2009. <http://help.youtube.com/support/youtube/bin/answer.py?hl=en-uk&answer=58123>

YouTube (2009h). 'Browse: Categories'. *YouTube: Broadcast Yourself.* 8 March 2009 <http://www.youtube.com/browse>