ЭЛЕКТРОННАЯ ПОЧТА VS. ЧАТ: ВЛИЯНИЕ КАНАЛА КОММУНИКАЦИИ НА ЯЗЫК

А. Бердичевский (alexander.berdichevsky@if.uib.no)

University of Bergen, Norway

Влияет ли смена канала коммуникации, не сопровождающаяся изменением других ситуационных параметров, на лингвистические характеристики коммуникации? Количественный анализ двух корпусов русских текстов, отличающихся исключительно каналом (электронная почта vs. чат) показывает, что влияет.

Ключевые слова: канал коммуникации, ситуационные параметры, характеристики коммуникации, лингвистические характеристики, анализ.

E-MAIL VS. CHAT: THE INFLUENCE OF THE COMMUNICATION CHANNEL ON THE LANGUAGE[1]

A. Berdichevskii (alexander.berdichevsky@if.uib.no)

University of Bergen, Norway

Does the mere change of the communication channel, unaccompanied by any other changes in situational characteristics, affect the language? Quantitative analysis of two corpora of Russian texts that differ solely by the communication channel from which they originate (e-mail vs. chat) proves that it does.

Key words: communication cannel, situational characteristics, communication characteristics, linguistic characteristics, analysis.

[1] This work was carried out as part of the project “The Future of Russian: Language Culture in the Era of New Technology”, supported by the Norwegian Research Council and the University of Bergen.

1. Introduction

Linguists are paying ever increasing attention to computer-mediated communication (CMC) and "the language of the Internet". At the "Dialogue" conference the Internet is usually viewed as a tool, not an object of linguistic research; however, even here one can find papers that focus on the linguistic properties of electronic communication (Макаров, Школовая 2006; Зализняк, Микаэлян 2006; Бурас, Кронгауз 2007; Богданов 2008; Anni 2008; Занегина 2009; Людовик 2010). In order to pursue a study of this kind, the scholar has to assume that the linguistic properties of CMC are somewhat different from those of other media (oral speech and written speech, for example) and thus worthy of separate research.

This assumption is often based on a more general one: the physical properties of the communication channel affect the linguistic properties of communication taking place in this channel, acting either as constraints or as enablements (see Hård af Segerstad 2002: 10–11 for the history of this term). This hypothesis has been well researched in the context of the differences between written and oral speech:e. g. see the classic works of Chafe (1982) and Biber (1988). Later, interest in this field was reinvigorated by the emergence and spread of a new channel, namely CMC. The constraints there seem to be heavier than in "traditional" channels, and the enablements wider, so that one might expect that their influence on the language would be clearly visible and detectable by quantitative methods.

Since the 1980s there have been quite a few studies that have used quantitative approaches to examine differences and similarities between CMC and other channels. It is important to keep in mind that CMC is not monolithic, and that we are in fact speaking about a set of different communication channels, united by the same physical medium: these channels have been compared to each other as well. See Collot and Belmore 1996, Yates 1996, Hård af Segerstad 2002 and review therein, Jensen 2007, Ling and Baron 2007, Tagliamonte and Denis 2008 and review therein. The results showed that CMC (or rather the specific channel studied — instant messaging, e-mail, computer conferencing and so on) is indeed a new linguistic register, neither oral speech nor written speech, and often looks like a hybrid of these two. Asynchronous communication channels with unlimited buffer size (e. g. e-mail) tend to be more similar to traditional written speech, whereas synchronous channels, especially with limited buffer size (instant messaging), are more similar to oral speech. However, Ko (1996) showed that, in certain parameters, CMC is even more "spoken" than speech and more "written" than writing.

2. Aim of this study

My intention is to compare two communication channels within CMC: e-mail and a certain type of instant messaging. A principal novelty of this study is that the

registers compared differ just by one parameter, namely the communication channel, whereas all other parameters (communicators and their relation to each other, subject matter, time of the discussion etc.) are controlled for as much as possible.

The studies mentioned above are often criticized precisely because of the lack of a control for additional parameters. Critics claim that the differences ascribed to the influence of the communication channel might in reality depend on other factors, e. g. the subject of discussion. Androutsopoulos (2006) takes this criticism even further: he states that the focus of attention should be the social context of a discourse and not its channel-specific properties. He even raises doubts about the existence of any linguistic features which might be ascribed to a communication channel: "It is empirically questionable whether in fact anything like a 'language of e-mails' exists, simply because the vast diversity of settings and purposes of e-mail use outweigh any common linguistic features" (Androutsopoulos 2006: 420).

The question I am addressing is the following: *does the communication channel per se have any influence on the linguistic properties of communication?*

Another novelty of my study is that I am analyzing Russian: it seems important to take CMC studies beyond the Anglophone world.

3. Materials

As a data source, I am using the contents of my own Gmail mailbox. Gmail provides not only the usual e-mail communication, but also a chat system (called Gmail chat). Since the chat is integrated into the same window (Fig. 1) and is easy to use, it is becoming increasingly popular.

Hence, it is common for the same two people to communicate both via e-mail and via chat. I collect my chat and e-mail conversations with three persons from my contact list. In order to avoid the observer's paradox, I am only using conversations which took place after June 2007 (after I graduated) and before March 2009 (before I submitted a proposal for my current PhD position), that is, when I was neither studying nor working as a linguist and did not have an idea of the current study (or anything similar) in mind.

That allows me to control for all the parameters except the communication channel itself. Indeed, the interlocutors are always the same, the setting is always the same, the subject matter may, of course, vary, but in general it might be quite clearly seen that the same things are discussed in both chat and in e-mail messages. There is no distribution of topics (such as chat for personal matters, e-mail for business). Conversation topics include mostly personal, business, scholarly and educational matters, and none of these four classes is restricted to a particular channel.

There are four subjects in my corpora: all male, native speakers of Russian, at the moment of communication aged 18 to 32, one university student, two journalists and one researcher. The e-mail corpus consists of 12 260 words and the chat corpus of 17 671 words, giving 29 931 words in total. The communication is always one-to-one.

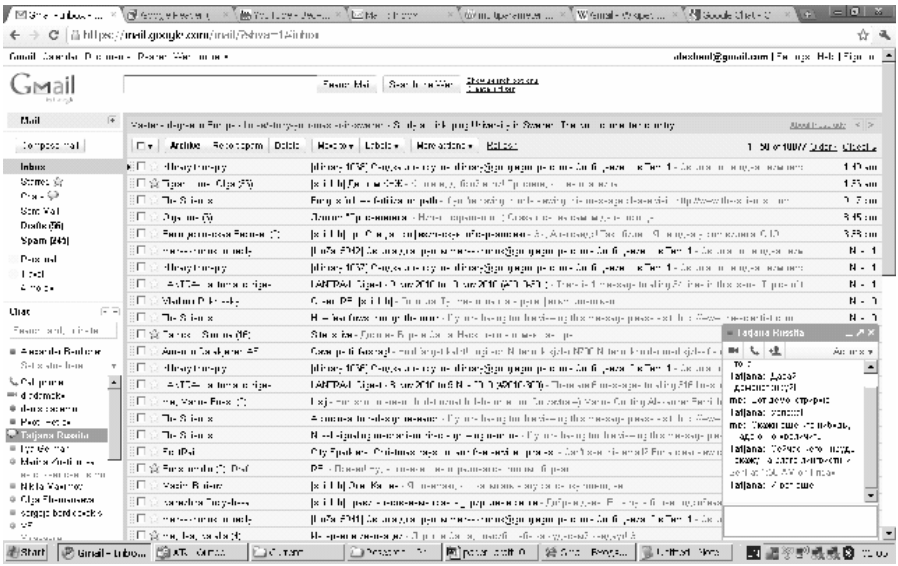

Fig. 1. Gmail chat

The Chat window is in the bottom right-hand corner. The contacts list can be seen in the bottom left-hand corner, and the name of the person who has sent a new message is highlighted.

4. Methods

Biber (1994) outlines a framework for the comparison of two registers. The framework consists of three components: analysis of the situational characteristics of the registers, analysis of the linguistic characteristics of the registers, and analysis of the functional and conventional associations between situational and linguistic characteristics. This section includes the situational analysis and lists the parameters for linguistic analysis . The “Results” section provides the results of the comparison of these parameters . The “Conclusions” section discusses the associations between situational and linguistic characteristics . This approach might be viewed as behavioural reductionism: I try to look at the influence of simple situational parameters on linguistic behaviour.

4.1. Differences between the situational characteristics of e-mail and Gmail chat

First, chat messages are delivered instantly. E-mails are also delivered quickly, but it might take a few seconds (or even minutes) for a letter to come.

Second, when you type an e-mail, your text is being auto-saved on a regular basis, so you do not have to worry about losing it should your browser crash, your

Internet connection be lost, or your computer stop working. When you type a message in a chat window, it is not saved anywhere until you send it.

Third, the chat window is narrow and small (see Fig. 1), while e-mail can occupy almost the whole screen. It is possible to open the chat in a separate window (and make it as large as one wants), or to install additional software in order to make chatting more convenient, but my subjects typically use the basic small window.

Fourth, when your interlocutor is typing a chat message to you, you can see an info message "XXX is typing..." (or "XXX has entered text") in the chat window.

Fifth, chat is more prone to technical failures: messages are more likely to get lost.

These are the real and primary differences between the two channels. They lead to the emergence of numerous secondary differences. For instance, in theory you may use chats to write long complex texts, but that would also be awkward : first, you always risk losing everything you have typed, second, it is inconvenient to read (and type, and edit) a large text in a small window. Some of these secondary differences are not, in fact, driven by physical reality, they are conventional. Strictly speaking, you do not have to answer to chat messages immediately, but that is what you are expected to do and what you usually do (and info messages contribute to users staying online and waiting for a reply to come).

Thus, chatting is usually a more synchronous, faster form of communication, implying immediate responses and rapid changes of turn. It is also somewhat less reliable and more volatile.

According to Biber, one of the principal oppositions in register comparison is informational versus involved production: "discourse with interactional, affective, involved purposes, associated with strict real-time production and comprehension constraints, versus discourse with highly informational purposes, which is carefully crafted and highly edited" (1988: 115). Oral speech is usually located closer to the "involved" pole of this dimension, while written speech — closer to the "informational" pole. It seems natural to expect that the same would be true forchat and e-mail respectively. Thus, many of the linguistic parameters discussed below are those that allow one to estimate the position of a register on this scale.

4.2. Quantitative parameters for discovering linguistic characteristics of e-mail and chat[2]

1. Mean length of an utterance (MLU)

Utterance here means 'sentence', with one exception: in chat, each turn is considered a separate utterance, i. e. a turn[3] might consist of several utterances (=sentences), but not vice versa. If a user chooses to split one sentence into nine turns (this is known to happen, although in my corpus they are rare), they are counted as nine utterances.

---

[2] Qualitative differences are not analyzed in this study.

[3] A turn is one chat message. In terms of Baron (2004: 408): 'composition (i. e., by typing) and transmission of an instant message'.

Otherwise, periods, exclamation, interrogation and ellipsis marks as well as emoticons were considered as marks to end an utterance. MLU is measured in symbols. High MLU is typical of informational speech production.

2. Mean length of a word (MLW)

High MLW is typical of informational speech production. Ko (1996) found MLW to be equal in speech and in instant messaging, but different from that in writing.

3. Lexical density (LD)

The ratio of lexical items (nouns, adjectives, verbs, adverbs, pronouns, numerals, as opposed to conjunctions, interjections, particles and prepositions) to the total number of words in a text. High LD is typical of informational speech production. Yates (1996: 35–39) showed that the LD of computer conferencing is close to that of writing, although still significantly different.

4. Type/token ratio (TTR)

The ratio of *different* words (types) in the text to the total number of words (tokens) in a text. Different word forms of the same lexeme were considered the same type, but different tokens. This measure depends on the text length, so it was calculated using two sub-corpora of equal size: 4 000 words.

High TTR implies a rich vocabulary and is typical of informational speech production. Yates (1996: 33–35) showed that the TTR of computer conferencing is close to that of writing, although still significantly different.

5. Sentence end marks

The percentage of sentences with any visible end marks: period, exclamation, interrogation or ellipsis marks. Sentences ending with an emoticon were also considered to have an end mark: sentence end is the most typical position for emoticons, and the period is usually omitted before them, so they can be viewed as an explicit signal of sentence end.

6. Capitals

The percentage of sentences beginning with a capital letter, as required by the rules of Russian punctuation/orthography.

7. Personal pronouns (first person, singular)

The ratio of the number of occurrences of the pronoun *я* ('I, me') (in all its forms) to the total number of words in a text. A high ratio is typical of "involved" speech production. Yates (1996: 40–41) found that the proportion of first-person pronouns in total pronoun use in CMC is higher than in speech, and in speech higher than in writing. Tagliamonte (2008: 16) confirmed the first part of this finding for instant messaging.

8. Brackets

The ratio of the number of brackets to the total number of words in the text. Brackets, too, serve as an indicator of informational production: a complex embedded

structure (both semantic and syntactic) is difficult to create (and perceive) when text is produced (and read) "on the fly".

9. Emoticons

The ratio of the number of emoticons to the total number of words in the text. The functions of emoticons are quite broad, but it is possible to state that, in general, speakers use them to compensate for the lack of non-verbal cues. Thus, high emoticon ratio would imply higher involvement.

10. Complex sentences

The ratio of complex sentences (i. e. sentences containing more than one clause) to the total number of sentences. Complex sentences are typical of informational production. This measure could not be calculated automatically, so it was calculated manually using the same sub-corpora that were compiled for measuring TTR.

Results

The results are summarized in Table 1. All the parameters were computed for each person and each pair separately, but only the results for the whole corpus are reported, since patterns were nearly the same in all cases.

The results of significance testing are reported, as well as effect sizes[4]. Parameters which are manifestly different for the two channels (difference is both significant and important) are highlighted in bold.

Table 1. Results

					% (1 per 100)				‰ (1 per 1000)	
	Utterance length	**Word length**	**LD**	**TTR**	**Sentence end marks**	**Capitals**	**1sg pronouns**	**Complex sentences**	**Brackets**	**Emoticons**
Chat	**33.8**	4.88	74.1	31.4	**54.7**	**78.3**	3.1	**22.9**	4.4	**22.9**
E-mail	**56.5**	5.03	75.1	30.2	**98.0**	**97.3**	3.0	**42.9**	9.3	**7.3**
Δ	**22.7**	0.15	0.9	1.2	**43.3**	**19,0**	0.1	**20.0**	4.9	**15.6**
Significant	**yes***	yes	no	no	**yes**	**yes**	no	**yes**	yes	**yes**
Effect size	**medium****	none	none	none	**large**	**medium**	none	**small**	none	**small**

[4] Significance testing shows how likely it is that the observed effect is random. It does not show how large and important it is. Since large samples can make very small effects visible, it is becoming increasingly common to report not only traditional significance, but also effect size (APA 2010: 33, Perry 2005: 224).

**yes* means p≤0.05 (in fact p is smaller than 0.001 in all the cases), *no* — p>0.05
***large* means h>0.80, *medium* — h>0.50, *small* — h>0.10, *none* — h≤0.10
Welch two-sample t-test (two-sided) applied for MLU and MLW; two-sample proportion test, for all the other cases. Effect size calculated as Cohen's d for MLU and MLW and as Cohen's h (arcsine transformation) in all the other cases.

Conclusions

Since five parameters appeared to be truly different for e-mail and chat, we can give a positive answer to the main research question: *yes, the communication channel does influence the language*.

The sentences are shorter in chat, due to the higher speed of communication: since an immediate answer is expected, people try to be quick rather than elaborate, and do not waste much time on editing and improving their texts (especially given that chat is not the best place to do that). Interestingly, that does not affect word lengths: the pressure is probably not strong enough to make that happen.

The lack of sentence end marks and capital letters occurs for two reasons. First of all, the need for speed leads to a weakening of the norm. Second, the norm actually turns out to be unnecessary: if a turn contains only one sentence (and that is usually the case), then even without capitals and periods it is clear where the sentence begins and where it ends. It would be different in a letter or in a turn containing several sentences, but in these cases the norm is usually not ignored.

It might also be supposed that chat is considered to be a less formal channel where norm violations are more appropriate, but this claim is hard to prove or disprove using my data.

Emoticons are more numerous in chat, since in a synchronous mode it is more important to show a "polite smile" to an interlocutor. They also have a phatic function: you are showing that you are interested in what your partner is saying, and you might reply to a message with a single smiling emoticon if you do not have anything else to say. As one of the subjects of this study put it, when questioned, "...I also want to be polite, so in chat I actually use a smiley instead of a period :)".

It is interesting to compare my results to those reported in Baron 2004 for English instant messaging (IM). Baron has found 49 instances of emoticons in her corpus of 11 718 words (Baron 2004: 413), the ratio being 0.004. My ratios are 0.023 (405/17671) for chat, 0.007 (90/11260) for e-mail, and 0.017 counted together — that is, much higher. This seems unusual, since the participants in my study are older and more educated than in Baron's. Besides, Baron's sample includes female subjects, and women tend to use more emoticons than men (Baron 2004: 416). It is unlikely that Russian IM is so much richer in emoticons than English IM. One explanation might be the observer's paradox: Baron's subjects knew they were being recorded while chatting, and this might easily have influenced their speech production (they might have tried to avoid "informal" traits like emoticons). Alternatively, it is possible that emoticons were less popular when Baron's study was conducted[5].

[5] This possibility was suggested to me by Alexander Piperski.

For brackets, the difference is significant, but the effect size is too small. Most likely this means that there actually is a difference, but the sample is too small to show it.

As for the other parameters, we might be quite sure that there are no differences, or that they are really tiny. This means that the influence of the communication channel should not be overestimated.

Further development of this study might include analysis of more complex parameters and data from the other social groups: less educated, less language-aware and not including myself, the researcher. It would also be useful to compare another set of channels, but it would be difficult, if not impossible, to reduce the distinction between registers to this single parameter.

References

1. *American* Psychological Association. Publication manual of the American Psychological Association, 6th edition. 2010.
2. *Androutsopoulos J.* 2006. Introduction: Sociolinguistics and Computer-mediated Communication. Journal of Sociolinguistics, 10 (4): 419–438.
3. *Anni O.* 2008. Choosing Language in Internet Conversations Between Russians and Estonians. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14): 602–605.
4. *Baron N.* 2004. See You Online: Gender Issues in College Student Use of Instant Messaging. Journal of Language and Social Psychology, 23 : 397–423.
5. *Biber D.* 1994. An Analytical Framework for Register Studies. Sociolinguistic Perspectives on Register : 44–56.
6. *Biber D.* 1988. Variation Across Speech and Writing.
7. *Bogdanova A. V.* 2008. Spelling in Internet: Analysis of one Misspelling Case [Orfografiia v Internete: Analiz odnoi orfograficheskoi oshibki]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2008" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2008"), 7 (14): 50–56.
8. *Buras M. M., Krongauz M. A.* 2007. The Language of Corporation Websites: Game, Parody, Provocation [Iazyk Korporativnyh Saitov: Igra, Parodiia, Prokatsiia]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2007" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2007") : 109–114.
9. *Chafe W.* 1982. Integration and Involvement in Speaking, Writing, and Oral Literature. Spoken and Written Language: Exploring Orality and Literacy : 35–53.
10. *Collot M., Belmore N.* 1996. Electronic language: A new variety of English. Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives : 13–28.
11. *Hård af Segerstad Y.* 2002. Use and Adaptation of Written Language to the Conditions of Computer-Mediated Communication.

12. *Jensen B. U.* Syntactic Variables in Pupils' Writing: a Comparison between Handwritten and PC-written Texts, available at http://privat.hihm.no/buj/dokumenter/2007-06-04 %20Bergen%20artikkel%20BJ.doc.
13. *Ko K.-K.* 1996. Structural Characteristics of Computer-Mediated Language: A Comparative Analysis of InterChange Discourse. Electronic Journal of Communication, 6 (3).
14. *Ling R.,Baron N.* 2007. Text Messaging and IM: Linguistic Comparison of American College Data. Journal of Language and Social Psychology, 26 (291): 291–299.
15. *Liudovyk T. V.* 2010. SMS Analysis with the Imrovement of Quality Target [Analiz tekstov SMS-soobshchenii c tsel'iu povysheniia kachestva]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2010" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2010"), 9 (16): 313–317.
16. *Makarov M. L., Shkolovaia M. S.* 2006. Linguistic and Semiotic Aspects of the Identity Construction in Electronic Communication [Lingvisticheskie I Semioticheskie Aspekty Konstruirovaniia Identichnosti v Elektronnoi Kommunikatsii]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006") : 364–369.
17. *Perry F. L.* 2005. Research in Applied Linguistics: Becoming a Discerning Consumer.
18. *Tagliamonte S., Denis D.* 2008. Linguistic Ruin? Lol! Instant Messaging and Teen Language. American Speech, 83 (1): 3–34.
19. *Yates S. J.* 1996. Oral and Written Linguistic Aspects of Computer Conferencing: A Corpus Based Study. Computer-Mediated Communication: Linguistic, Social and Cross-Cultural Perspectives : 29–46.
20. *Zalizniak A. A., Mikaelian I. L.* 2006. Emaling as a Linguistic Object [Perepiska po Elektronnoi Pochte kak Lingvisticheskii Ob"ekt]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2006" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2006") : 157–162.
21. *Zanegina N. N.* 2009. I Didn't Say it: on Lituratives, Strikeout and Quasi-texts [Ia etogo ne govoril: O Liturativakh, Zacherkivaniiakh ili Mnimukh Tekstakh]. Komp'iuternaia Lingvistika i Intellektual'nye Tekhnologii: Trudy Mezhdunarodnoi Konferentsii "Dialog 2009" (Computational Linguistics and Intelligent Technologies: Proceedings of the International Conference "Dialog 2009"), 8 (15):112–115.