

# Catching the news: two key cases from today

Ruslana Margova,  
Irina Temnikova

# The news and the Internet

- The common understanding - the newest information is in the established news media.
- Nowadays the news could appear in comments under the news, in blogs or in social networks.

# Reader and/or editor

The Internet communication removes the distinction between readers and editors and the editor becomes a reader and the reader himself can generate the news.

# The emergence of the news

- Two key cases of the emergence of news in **Internet** analyzed here:
  - the emergence of news in comments
  - the emergence of news in information-sharing websites and their interpretation in the newspapers.

# Five levels of transtextuality

In 1982 Gerard Genette introduces:

- hypotext - the text basis
- hypertext - text's understanding
- paratext - title, subtitle, prefaces, footnotes
- metatext - commentaries, literary critique
- intertext - relationship between two or more texts: quotation; plagiarism; allusion

# ... and websites

All these levels can be seen in the Internet websites:

- hypotext – the essential information for the user
- hypertext - the user's previous knowledge
- paratext - all additional features: the Internet address, the images and surrounding items
- metatext - the comments, links to other sites and texts
- intertext - all quoted or misquoted pieces of text

# Why we need these levels?

- To understand the links between the news and to demonstrate the emergence of the news in all these different levels: what is metatext in one site is hypotext in another.

# Corpora



- The first corpus - the news published by the editors and the postings of bloggers under it. Time markers, bloggers names are preserved (10340 words).
- The second corpus – three different types of texts – the names of the videos in [www.youtube.com](http://www.youtube.com), a blog of Paulo Coelho, the results of the Google News Search engine (21659 words).

# What the news is?

In the theory of journalism the 'news' is a previously unknown information about a recent and important event. The problem is how to define what is previously known and what is not.

# Information Extraction

Information Extraction (IE) is the sub-area of NLP which deals with the extraction of news.

For the task of catching the news, the definition of 'previous knowledge' is very important.

# Event Extraction

Event Extraction (EE) can be defined as extracting all occurrences of a relationship between specific participants in an event from text. In order to capture this relationship patterns are being built.

Usually the attention is being focused on identifying and extracting Named Entities (NEs), such as names of persons, organisations and locations of time markers, people positions.

# IE and temporal labels

In our cases:

IE could be conceived as the technology of extracting the information and the temporal labels could be conceived as a guarantee for the novelty of information.

In our two corpora we keep the date and the hour from the paratext information.

# First case: the story

The owner of a newspaper publishes in his own newspaper “the news” that he wants an official apology from a television because of an insult.

Later, in the comments under this main news, like a blogger, he adds that he will receive the apology.

The metatext is more meaningful than hypotext and later becomes hypotext in another circumstances.



# The article



- Typically, the article is clearly structured, with clear simple phrases, with exact quotations and with enough information for the readers. The news is in the first phrase and the additional information follows.

# The comments

Grammatically the comments could be distinguished by:

- the faults: bad constructions, unfinished phrases, etc.
- posted by nicknamed people.
- emotional: defending two main opposite positions

# IE and comments

- Information extraction and Opinion mining could help to find more information about the main participants in the story.
- Ex.: The extraction of named entities provides some additional information – who is who, who is liked by whom, what somebody did or didn't do, etc. But that information is a previous knowledge.

# Proper names and co-references

- The list of proper names and co-references which are mentioned in the main news text and used by the bloggers:

<b>Стефан Гамизов;</b>	Гамизов; гамизов; енергиен експерт; Гамизовчето; келеме; провален собственик на медия; примат; агент на Черепа; клюкарка; председател на Гражданска Лига на България; Гемизо
<b>Иво Прокопиев;</b>	Прокопиев; премиер на България; келеш; ощипана мома; фамилията Прокопиеви; червенобузко
<b>Николай Барекон;</b>	Барекон; Барека; мъжка проститутка; боклук; водещ; бюреков; журналистическа проститутка; Дудука майна; дудук

# Newsmakers or bloggers

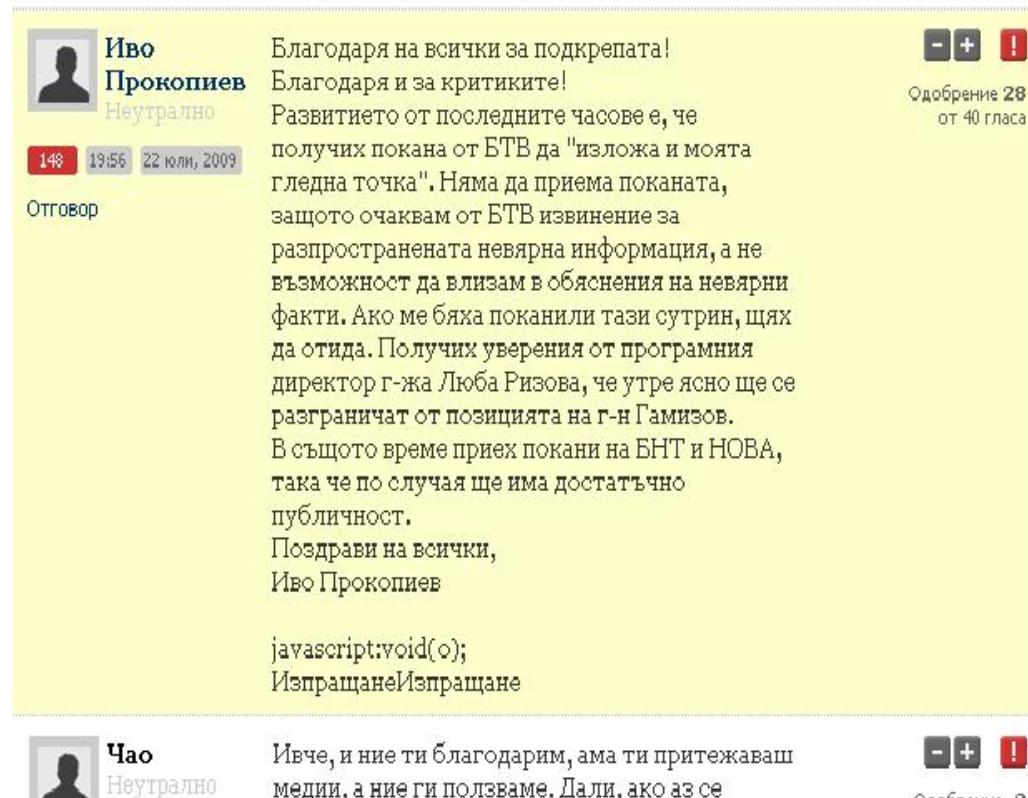
- In one of the postings, a blogger identifies himself as the guest and puts an answer.
- The linguistic analysis shows that the posting is bad written:
  - an inversion in the noun phrase,
  - missing information (to whom/where),
  - ambiguity,
  - repetition,
  - misquotation.

# Newsmakers or bloggers 2

- Unfortunately the intervention of the guest in the comments is still not a news – it is only a comment. The bad style in the first phrase makes unclear where/to whom the guest sent this letter, and how this blogger knows about that – so whether it is a truth or not.

# Newsmakers or bloggers 3

A little bit later, as the temporal labels show, another blogger identifies himself as the owner of the newspaper and adds a news in the comments – he announces that there will be an apology.



**Иво Прокопиев**  
Неутрално  
148 19:56 22 юли, 2009

Отговор

Благодаря на всички за подкрепата!  
Благодаря и за критиките!  
Развитието от последните часове е, че получих поканата от БТВ да "изложа и моята гледна точка". Няма да приема поканата, защото очаквам от БТВ извинение за разпространената невярна информация, а не възможност да влизам в обяснения на невярни факти. Ако ме бяха поканили тази сутрин, щях да отида. Получих уверения от програмния директор г-жа Люба Ризова, че утре ясно ще се разграничат от позицията на г-н Гамизов. В същото време приех покани на БНТ и НОВА, така че по случая ще има достатъчно публичност.  
Поздрави на всички,  
Иво Прокопиев

javascript:void(0);  
ИзпращанеИзпращане

Одобрение 28 от 40 гласа

**ЧАО**  
Неутрално  
19:56 23 юли, 2009

Ивче, и ние ти благодарим, ама ти притежаваш медии, а ние ги ползваме. Дали, ако аз се

Одобрение 2

# Newsmakers or bloggers 4

- The news is introduced by the phrase: “*Развитието от последните часове е*” (*the breaking news of the last hours is*).
- The information is presented in first person, from somebody who pretended to be the real owner.
- It's not an edited text – there are two repetitions (невярна информация, невярни факти) and one misspelled word (невярни).
- But this comment is a news – the new developments of the event, quoted by many media and finally happened.

# How many news in the comments

There are only two other postings, conceived as news from the bloggers themselves, but they are not. All other texts are interpretations, analyses and opinions.

**до Щастливеца**  
Неутрално  
129 18:31 22 юли, 2009  
Отговор

Те ти булке Спасов ден:  
Десислава Танева няма да бъде министър на земеделието. Номинацията ѝ е оттеглена от бъдещия премиер Бойко Борисов ден след като той лично обяви, че тя ще заеме поста. Името на бизнес дамата от Сливен се свързва с кръга "Капитал", тъй като тя участва в управлението на Фонд за земеделска земя „Мел инвест“. 26,1% от него се държат от „Алфа финанс“ на Иво Прокопиев.

Одобрение 0 от 8 гласа

---

**Новината**  
Неутрално  
161 21:59 22 юли, 2009  
Отговор

Тази вечер министър председателя Иво Прокопиев ще говори по новините на БНТ 1. Това е новината а другото е между другото!

Одобрение -5 от 9 гласа

# Some conclusions

- The real news is not very often seen in the comments. In all 186 comments there is only one, which is introduced by a concrete clear phrase: “*Развинуемо om последните часове e*” (*the breaking news of the last hours is*).
- The intervention of the owner was easy to identify, thanks to the paratext markers (time, nickname).
- **Still many problems:** whether everyone, who presents himself as a particular person, is really the same person; how the reader could be sure about that; and furthermore - how the automatic extraction of information has to be made and how authentic it could be.

# Second case: the story

After the presidential elections in Iran in mid-June, a girl (Neda) is killed during the protests in the street and her death is filmed by bystanders and broadcasted over the Internet – in Youtube ([www.youtube.com](http://www.youtube.com)) and social networks like FaceBook ([www.facebook.com](http://www.facebook.com)) and Twitter ([www.twitter.com](http://www.twitter.com)). The news is later interpreted by all daily online newspapers and the social networks are quoted as the official sources.





## Event reconstruction and new channels

- temporal anchor: June 21 – the day when the video of Neda's death appeared in Youtube, in Facebook and in Twitter
- Video: appeared first in a non-news website
- Problem: how to catch the news from the Internet

# Associated Press, 22 June 2009:

CAIRO (AP) [Amateur video of a young Iranian woman lying in the street – blood streaming from her nose and mouth](#) — has quickly become an iconic image of the country's opposition movement and unleashed a flood of outrage at the regime's crackdown.

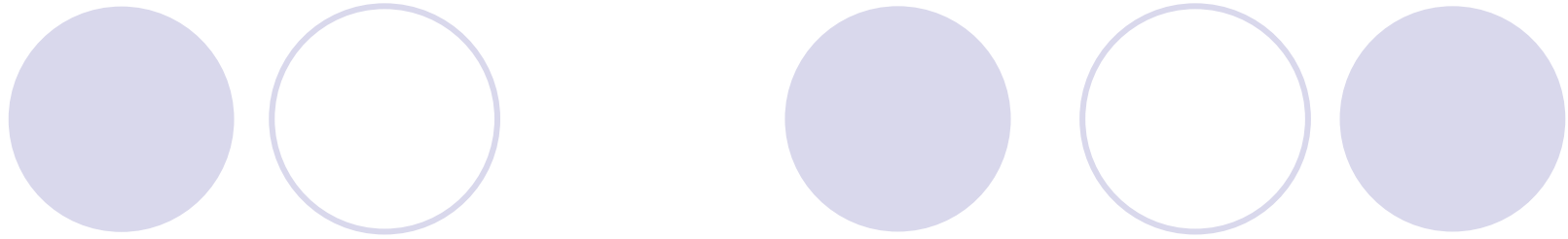
The footage, less than a minute long, appears to capture the woman's death moments after she was shot at a protest — a powerful example of citizens' ability to document events inside Iran despite government restrictions on foreign media and Internet and phone lines.

The limits imposed amid the unrest over the disputed June 12 election make details of the woman's life and events immediately preceding her apparent death difficult to confirm. But clips of the woman being called Neda are among the most viewed items on YouTube — with untold numbers of people passing along the amateur videos through social networks and watching them on television.

[The images entered wide circulation Saturday when two distinct videos purporting to show her death appeared separately on YouTube and Facebook.](#)[...]

Thousands of people inside and outside Iran have written online tributes to the woman, many condemning the government and praising her as a martyr.

Some posted photos of a gently smiling woman they said was Neda, some calling her "Iran's Joan of Arc."



After the acceptance of these channels as news information media, the next step is to answer how to use these new channels.

# Summarization of the information

The search in [www.youtube.com](http://www.youtube.com) of the keyword “neda” returns a set of titled videos. The collection of all these titles could be conceived as a kind of summarization of the real story about the death of Neda Agha-Soltan.

Neda's Death Becomes Iranian Symbol  
Neda Agha Soltan, killed 20.06.2009, Presidential Election Protest, Tehran  
Her name was Neda  
Neda before she gets shot  
IRAN PROTEST IMAGES IN TRIBUTE TO NEDA  
RIP Neda Soltani, Neda! Don't Be Scared, Neda!  
CNN: "Death Of Neda" Video Becomes Symbol Of Iranian Protests  
Fiance tells of Neda's last moments - 23 Jun 09  
Twitter Revolution – Iran  
CIA KILLED NEDA  
United for Neda  
I Am Neda  
For Neda  
John McCain Addresses Killed Iranian 'Neda' on Senate  
SONG FOR NEDA original music by Greg V. In honor of Neda

*Neda Agha Soltan is killed on 20.06.2009, in a presidential election protest at Tehran. Her fiance tells of Neda's last moments. She becomes an Iranian symbol. There are some allegations that Neda is killed by CIA. There is a Twitter Revolution – Iran. There is a tribute to Neda – songs, addresses in the Senate.*

# What can be done:

- Keywords extraction → automatic summary of the story
- Thematically related titles → co-reference chains → anaphora resolution in news topic tracking

# Information from blogs

23 June 2009

My best friend in Iran, a doctor who showed me its beautiful culture when I visited Teheran in 2000, who fought a war in the name of the Islamic Republic (against Iraq), who took care of wounded soldiers in the frontline, who always stood by real human values, is seen here trying to resuscitate Neda - hit in her heart.

26 June 2009

Here is how I discovered that my friend Arash Hejazi was the doctor trying to resuscitate Neda.

- The videos of Neda's death are posted on many private websites and blogs: for example - of the Brazilian writer Paulo Coelho.
- In the post from 23 June he announced the news and in the post from 26 June he announced the name of doctor: [Arash Hejazi](#). Five days later the news agencies announced: [Arash Hejazi](#) is wanted by intelligence ministry and Interpol.

# Conclusions



Starting from understanding intertextuality, we presented the development of the news in two key cases.

- The main conclusion is that the shared-information websites, social network sites, blogs and comments under a particular article can be without a doubt conceived as **new news channels**.
- The **Information Extraction technology** can help as a method of catching a suitable information from these new sources. The introduction of **time labels** as a guarantee of the novelty of information is helpful to determine which are the real news.

# Future work



Future work will start with **in-depth linguistic analysis** of the news comments and blog posts to show their differences from the classical news articles texts

Considering the **news articles texts as a form of a controlled language** will help to make this distinction clearer.

The collection of a **more representative corpus of new text types** and their analysis should be considered as the next step.

# Any questions?



*Thank  
You!*