

# Processing Web-derived Text

---

Working with very  
messy data

Sebastian Hoffmann  
University of Zurich  
[sebhoff@es.unizh.ch](mailto:sebhoff@es.unizh.ch)

# General approach

---

Until we have both the tools and the methodology to work with the web as corpus, using web-derived data can complement the available corpus resources.

- *The Statesman* news archive
- Usenet - newsgroups

# http://www.thestatesman.org

The Statesman Afternoon

# The Statesman

The first completely customisable news site on the web  
130 years in print

Quest Online Booking Classified Ads

Advanced Search

Home | Classified | Jobs | Matrimonials | Archives | Advertise | Feedback | About Us

Sunday, July 3 2005

News  
Page one  
India  
World  
Editorial  
Perspective  
Business  
Sport  
Bengal  
Magazine  
Sports & Leisure  
Career & Campus  
Science & Technology  
Voices  
Lifestyle  
Kolkata Plus  
Bengal Plus  
Viewpoint  
North East Page  
Orissa Plus  
Note Book  
N.B & Sikkim Plus  
Entertainment  
NB Extra  
World Focus

8TH DAY

Crystal Ball

Kolkata Plus

SUBSCRIPTION

Make the promise of a lifetime  
Register FREE Bengali Matrimony.com No. 1 Bengali Matrimony

 HIGH FLYERS: Canadian Air Force aircraft fly in formation over the Peace Tower during Canada Day celebrations in Ottawa on Friday.- PTI

[Page One]

- Nothing fails like Failure!
- Column Ones
- Ash winds up in Rajasthan
- The road not taken for Salt Lake

(more...)

[India]

- Will Reangs get back home?
- Epidemic fears grip Gujarat
- Don't mess with Lalu, Congress tells Paswan
- Fardeen: Casting cocaine crook

News Flash

Kissinger regrets Nixon's diction

Ajit Jogi's son held for NCP leader's murder

Ball By Ball Coverage

Login

Username

Password

Login

New user? Sign-in

Poll

Peace with Pakistan is not a realistic desire given current conditions.

Agree 66%

Disagree 33%

Can't Say 1%

Vote | archive

Stock

Index	Closing
BSE	7210.77
S & P CNX Nifty	2211.90

Exchange

# A typical archive entry

The screenshot shows the website 'The Statesman Afternoon' with the tagline 'The first completely customisable news site on the web' and '130 years in print'. The date is Thursday, June 30, 2005. The user is logged in as 'Sebastian'. The main article is titled 'EC wants warrants executed in Bihar' and discusses the Election Commission's directives to the Bihar government regarding non-bailable warrants and special search operations. The article is part of an archive for the period 2005-06-24.

**The Statesman Afternoon**  
Tomorrow's

**The Statesman**  
The first completely customisable news site on the web  
130 years in print

Search  in  The Statesman  Web Search

Thursday, June 30 2005

Hi! Sebastian [Home](#) [Classified](#) [Jobs](#) [Matrimonials](#) [Archives](#) [Advertise](#) [Feedback](#) [About Us](#)

**News**

- Page one
- India
- World
- Editorial
- Perspective
- Business
- Sport
- Bengal

**Magazine**

- Sports & Leisure
- Career & Campus
- Science & Technology
- Voices
- Lifestyle
- Kolkata Plus
- Bengal Plus
- Viewpoint
- North East Page
- Orissa Plus
- Note Book
- N.B & Sikkim Plus
- Entertainment
- NB Extra
- World Focus

**[Archives]**

[Back to archives for 2005-06-24](#)

**[Page One]**

**EC wants warrants executed in Bihar**

Statesman News Service  
NEW DELHI, June 23. — The Election Commission today issued a strict directive to the Bihar government to execute all non-bailable warrants (NBWs), conduct special search operations to unearth illegal arms and send district-wise weekly consolidated reports on the compliance of the directions issued by the EC.

This is the first time that the EC has issued such instructions to a state government. The Assembly polls are due in four months' time in Bihar.

The directions, however, did not mention transfers and postings of government employees and implementation of the model code of conduct, which the BJP had demanded alleging that Mr Lalu Prasad would misuse the official machinery to influence voters.

In its nine-point instruction order to the Bihar government, the EC has said that all NBWs pending against criminals must be executed, special search operations must be conducted to unearth illegal arms and preventive action taken against habitual offenders and anti-social elements across the state. The EC told the state government to provide proper communication facilities — mainly telephones, wireless sets and vehicles — to all police stations and directed the state government to maintain arterial and important roads in "good condition".

Keeping in mind the approaching monsoon and possible floods, the EC called upon the state government to repair damaged roads and related infrastructure immediately. It also urged the Bihar government to provide necessary funds to the Chief Electoral Officer for issuing voters' photo identity cards.

The EC has also asked the state government to implement the Prevention of Defacement of Property Act "in letter and spirit". It asked all district collectors to conduct special comprehensive training programmes on the functioning of the electronic voting machines for both the representatives of political parties and state government officials.

A two-member team of the EC, comprising the Deputy Election Commissioner, Mr Anand Kumar, and the Election Commission Adviser, Mr KJ Rao, had visited Patna early this week to oversee the poll preparations in the state.

**News Flash**

- New-look Rajdhani from this October
- India, USA ink defence pact
- Brazil wins

**User**

Welcome sebhoff2002  
[Change profile](#) | [Password](#)  
[Logout](#)

# The Statesman Archive

---

- January 2002 - March 2005
- 81,150 news articles
- Approx. 31,5 million words

# Verb-complementation in Indian English

---

- (1) The first obligation of these media institutions will be to *inform the country the facts and events* without prejudice or conscious partisanship.  
(2003-11-07)

# Usenet

---

- Dates back to 1979
- Thousands of newsgroups
- Wide range of topics discussed
- “Interactive” -> users can reply to previous posts (and quote text)

# Sample from alt.usage.english



[alt.usage.english](#)

[Ohne Frame](#) | [Nach Da](#)

## contact with someone

- | [1 walker](#) 3 Jul.
- | [2 ray o'hara](#) 3 Jul.
- ▶ [3 Don Phillipson](#) 3 Jul.
- | [4 Steve Hayes](#) 4 Jul.
- | [5 CDB](#) 4 Jul.
- | [6 walker](#) 4 Jul.
- | [7 Bill Bonde \('by a cc](#)
- | [8 walker](#) 4 Jul.

## ★ contact with someone or contact someone?

« Themenbeginn « Ältere Nachrichten 1 - 8 von 8 Neuere » [Themenende](#) »

**1. walker** 3 Jul. 22:28 [Optionen anzeigen](#)

Hi guys,

I googled and found both. Which one is used more widely?

Thanks in advance.

Kong

▶ [Antworten](#)

# Sample from alt.usage.english

**3. Don Phillipson**

3 Jul. 23:55

[Optionen anzeigen](#)

"walker" <a...@xyz.com> wrote in message  
[news:da9hpk\\$jcl\\$1@domitilla.aioe.org](mailto:news:da9hpk$jcl$1@domitilla.aioe.org)

...

> I googled and found both. Which one is used more widely?

Googling for one or two words helps little because Google cannot differentiate the verb contact from the noun contact, e.g.

1. I am in contact (noun) with XYZ.
  2. I want to contact (verb) XYZ.
- Both forms are currently used.

--

Don Phillipson  
Carlsbad Springs  
(Ottawa, Canada)

# Sample from alt.usage.english

---

**4. Steve Hayes** 4 Jul. 01:36 [Optionen anzeigen](#)

On Mon, 4 Jul 2005 00:58:22 +0430, "walker" <a...@xyz.com> wrote:  
>Hi guys,

>I googled and found both. Which one is used more widely?

The one whose meaning is needed more frequently.

--

Steve Hayes from Tshwane, South Africa

<http://www.geocities.com/Athena/7734/stevesig.htm>

E-mail - see web page, or parse: shayes at dunelm full stop org full stop uk

# Newsgroups corpus

---

Possible research questions:

- Discourse strategies - written/spoken?
- Cohesion - e.g. anaphoric reference
- Coherence
- ....

-> Data requires post-processing

# Step 1: Download

---

- Google archive is not an option
- Existing newsreaders: not flexible enough
- Perl (with module `Net::NNTP`) -> creation of primitive newsreader
- MySQL database to keep track of downloaded data

# Download (continued)

---

- Advantage of commercial news server (e.g. Giganews): retention rates
- Download speed: 1 message per second (Giganews: 10 concurrent connections) -> 850,000/day
- Which newsgroups should be downloaded? -> representativeness

# Downloaded groups

Table 1: Newsgroups downloaded

Name	N messages	N words
alt.alien.research	58,189	13,978,358
alt.coffee	68,056	12,076,503
alt.fan.noam-chomsky	62,703	25,683,845
alt.games.warcraft	23,849	3,896,371
alt.music.oasis	21,883	1,934,548
alt.support.marriage	76,183	20,757,171
news.software.nntp	5,033	1,177,785
news.software.readers	31,874	4,790,531
rec.audio.tech	24,645	5,247,679
rec.gambling.sports	34,848	4,955,744
rec.music.classical.recordings	147,866	28,188,853
rec.photo.digital	279,846	46,358,859
rec.sport.swimming	15,980	3,360,974
Total:	849,955	172,407,221

## Step 2: Conversion

---

- Convert files into format that can be used with standard concordancing software
- Take care of quoted text -> assign correct author

# A message header

---

From: Dan Swartzendruber <dswartz@druber.com>  
Newsgroups: alt.fan.noam-chomsky, soc.history, soc.history.ancient, soc.history.what-if, talk.politics.misc  
Subject: Re: What is the most dangerous false belief in the world today ?  
Date: Tue, 1 Jul 2003 21:17:20 -0400  
Organization: Posted via Supernews, <http://www.supernews.com>  
Message-ID: <MPG.196bfcab94cf0a8989718@news.supernews.net>  
References: <a1333567.0307010632.744e81cd@posting.google.com>  
<bdt8ab\$sojs\$1@newsg4.svr.pol.co.uk>  
<MPG.196bf1e0fe42036a989717@news.supernews.net>  
<bdtbh7\$jhi\$1@news6.svr.pol.co.uk>  
X-Newsreader: MicroPlanet Gravity v2.50  
X-Complaints-To: [abuse@supernews.com](mailto:abuse@supernews.com)  
Lines: 26

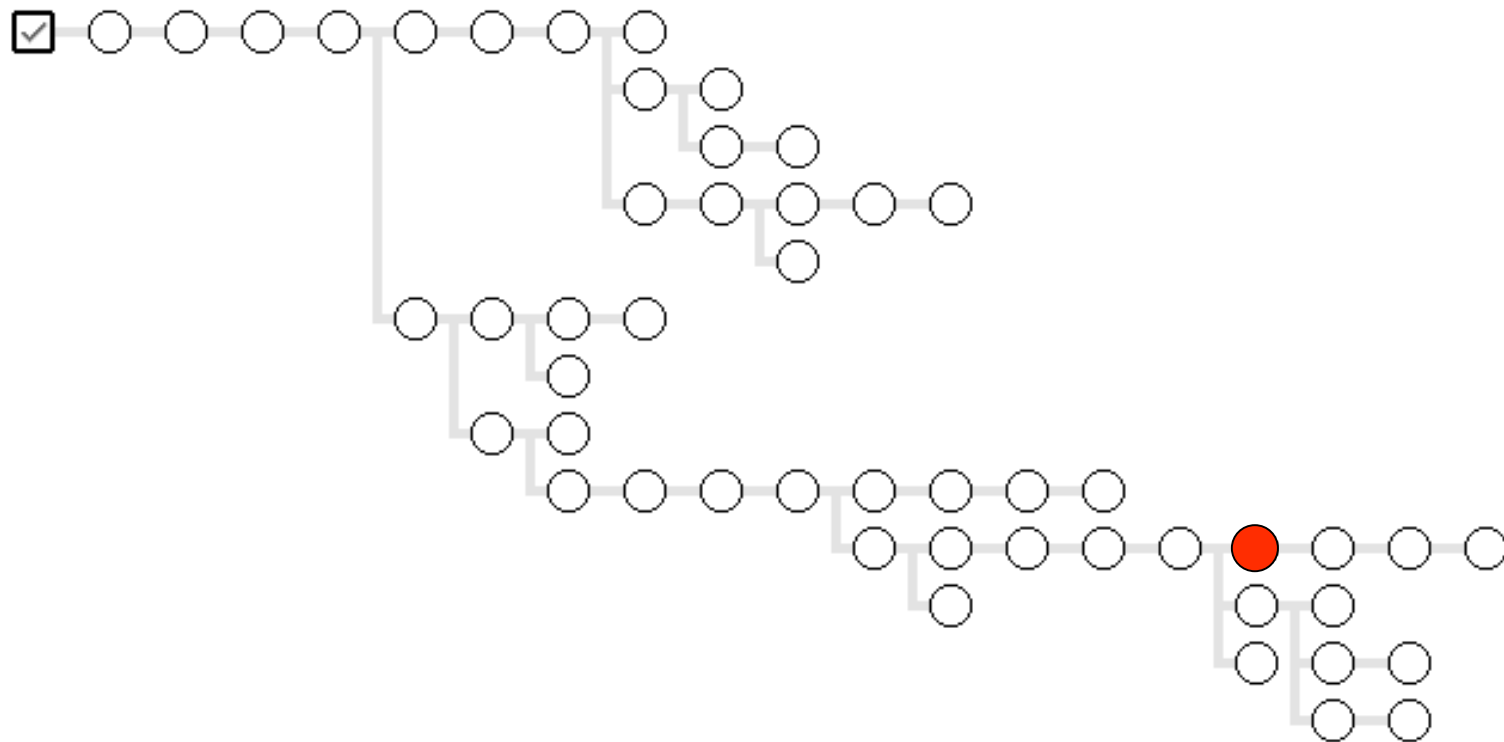
# A message header

---

From: Dan Swartzendruber <dswartz@druber.com>  
Newsgroups: alt.fan.noam-chomsky, soc.history, soc.history.ancient, soc.history.what-if, talk.politics.misc  
Subject: Re: What is the most dangerous false belief in the world today ?  
Date: Tue, 1 Jul 2003 21:17:20 -0400  
Organization: Posted via Supernews, <http://www.supernews.com>  
Message-ID: <MPG.196bfcab94cf0a8989718@news.supernews.net>  
References: <a1333567.0307010632.744e81cd@posting.google.com>  
<bdt8ab\$oj\$1@newsg4.svr.pol.co.uk>  
<MPG.196bf1e0fe42036a989717@news.supernews.net>  
<bdtbh7\$jhi\$1@news6.svr.pol.co.uk>  
X-Newsreader: MicroPlanet Gravity v2.50  
X-Complaints-To: [abuse@supernews.com](mailto:abuse@supernews.com)  
Lines: 26



# A typical thread



# References

---

From: Dan Swartzendruber <dswartz@druber.com>

Newsgroups: alt.fan.noam-chomsky, soc.history, soc.history.ancient, soc.history.what-if, talk.politics.misc

Subject: Re: What is the most dangerous false belief in the world today ?

Date: Tue, 1 Jul 2003 21:17:20 -0400

Organization: Posted via Supernews, <http://www.supernews.com>

Message-ID: <MPG.196bfcab94cf0a8989718@news.supernews.net>

References: <[a1333567.0307010632.744e81cd@posting.google.com](mailto:a1333567.0307010632.744e81cd@posting.google.com)>

<[bdt8ab\\$sojs\\$1@newsg4.svr.pol.co.uk](mailto:bdt8ab$sojs$1@newsg4.svr.pol.co.uk)>

<[MPG.196bf1e0fe42036a989717@news.supernews.net](mailto:MPG.196bf1e0fe42036a989717@news.supernews.net)>

<[bdtbh7\\$jhi\\$1@news6.svr.pol.co.uk](mailto:bdtbh7$jhi$1@news6.svr.pol.co.uk)>

X-Newsreader: MicroPlanet Gravity v2.50

X-Complaints-To: [abuse@supernews.com](mailto:abuse@supernews.com)

Lines: 26

# Message body

---

In article <bdtbh7\$jhi\$1@news6.svr.pol.co.uk>, agamemnon@hello.to.NO\_SPAM says...

>

> "Dan Swartzendruber" <dswartz@druber.com> wrote in message

> news:MPG.196bf1e0fe42036a989717@news.supernews.net...

>> In article <bdt8ab\$ojis\$1@newsg4.svr.pol.co.uk>,

>> agamemnon@hello.to.NO\_SPAM says...

>>>

>>> "Zardoz" <zardoz07@myfastmail.com> wrote in message

>>> news:a1333567.0307010632.744e81cd@posting.google.com...

>>>> What is, in your opinion, the most influential and dangerous false

>>>> belief in today's world?

>>>>

>>>> By false belief, I mean something that had been refuted by the experts

>>>> beyoun a reasonable doubt, but is still held by the general public (or

>>>> a part thereof) as true.

>>>>

>>>> Neo-Conservatism, Zionism, and Islam.

>>

>> I guess Marxism is not a valid choice, since the "still held by the

>> general public" is no longer true :)

>

> Marx was a Zionist.

Which is irrelevant, as far as I can tell. Certainly 99.999% of the people who purported to follow Marxism weren't.

# Message body

---

In article <bdtbh7\$jhi\$1@news6.svr.pol.co.uk>, agamemnon@hello.to.NO\_SPAM says...

>

> "Dan Swartzendruber" <dswartz@druber.com> wrote in message

> news:MPG.196bf1e0fe42036a989717@news.supernews.net...

>> In article <bdt8ab\$oj\$1@newsg4.svr.pol.co.uk>,

>> agamemnon@hello.to.NO\_SPAM says...

>>>

>>> "Zardoz" <zardoz07@myfastmail.com> wrote in message

>>> news:a1333567.0307010632.744e81cd@posting.google.com...

>>>> What is, in your opinion, the most influential and dangerous false

>>>> belief in today's world?

>>>>

>>>> By false belief, I mean something that had been refuted by the experts

>>>> beyond a reasonable doubt, but is still held by the general public (or

>>>> a part thereof) as true.

>>>

>>> Neo-Conservatism, Zionism, and Islam.

>>

>> I guess Marxism is not a valid choice, since the "still held by the

>> general public" is no longer true :)

>

> Marx was a Zionist.

Which is irrelevant, as far as I can tell. Certainly 99.999% of the people who purported to follow Marxism weren't.

# Problem 1: Quote markers

---

- Angle brackets are not the only way to mark quoted text:
  - #
  - !
  - >!
  - :

Decision: Angle brackets only



# Conversion process

---

- Isolate “root” messages
- Then match quotes of “level 1” messages to text in root messages
- Work through all levels successively, matching quotes to converted text of previous level

# Individual steps

---

- First try matching whole lines
  - If match: annotate with Msg-ID of “parent”
- Allow optional line break after each word
- Allow last character of line to be cut off
- Allow elisions of text (<SNIP>, [...], etc.)
- ... ..

# Caveat

---

- Quote assignment can never be 100%
  - “spurious” quoting
  - correction of spelling mistakes
  - deliberate changes in quoted text

Optimization possible but time-consuming  
-> may also introduce new imprecisions

# Statistics

Table 2: Success rates for quote assignment

Name	messages	with quotes	% un-assigned	% unassigned per level
rec.gambling.sports	34,848	17,628	26.9%	17.2%
alt.fan.noam-chomsky	62,703	52,801	10.9%	6.1%
alt.alien.research	58,189	48,189	18.3%	6%
rec.music.classical.recordings	147,866	119,269	8.4%	5.1%
rec.sport.swimming	15,980	12,799	6.9%	4.4%
alt.support.marriage	76,183	69,955	7.1%	3.8%
rec.photo.digital	279,846	234,497	6.1%	3.6%
news.software.readers	31,874	26,480	5.5%	3.5%
rec.audio.tech	24,645	18,868	5%	3.2%
alt.music.oasis	21,883	14,989	4.7%	3.1%
alt.coffee	68,056	50,243	4.4%	3%
news.software.nntp	5,033	3,709	4.6%	2.9%
alt.games.warcraft	23,849	20,135	2.6%	1.4%
Total:	849,955	689,562		

# Statistics

Table 2: Success rates for quote assignment

Name	messages	with quotes	% un-assigned	% unassigned per level
rec.gambling.sports	34,848	17,628	26.9%	17.2%
alt.fan.noam-chomsky	62,703	52,801	10.9%	6.1%
alt.alien.research	58,189	48,189	18.3%	6%
rec.music.classical.recordings	147,866	119,269	8.4%	5.1%
rec.sport.swimming	15,980	12,799	6.9%	4.4%
alt.support.marriage	76,183	69,955	7.1%	3.8%
rec.photo.digital	279,846	234,497	6.1%	3.6%
news.software.readers	31,874	26,480	5.5%	3.5%
rec.audio.tech	24,645	18,868	5%	3.2%
alt.music.oasis	21,883	14,989	4.7%	3.1%
alt.coffee	68,056	50,243	4.4%	3%
news.software.nntp	5,033	3,709	4.6%	2.9%
alt.games.warcraft	23,849	20,135	2.6%	1.4%
Total:	849,955	689,562		

# Statistics

Table 2: Success rates for quote assignment

Name	messages	with quotes	% un-assigned	% unassigned per level
rec.gambling.sports	34,848	17,628	26.9%	17.2%
alt.fan.noam-chomsky	62,703	52,801	10.9%	6.1%
alt.alien.research	58,189	48,189	18.3%	6%
rec.music.classical.recordings	147,866	119,269	8.4%	5.1%
rec.sport.swimming	15,980	12,799	6.9%	4.4%
alt.support.marriage	76,183	69,955	7.1%	3.8%
rec.photo.digital	279,846	234,497	6.1%	3.6%
news.software.readers	31,874	26,480	5.5%	3.5%
rec.audio.tech	24,645	18,868	5%	3.2%
alt.music.oasis	21,883	14,989	4.7%	3.1%
alt.coffee	68,056	50,243	4.4%	3%
news.software.nntp	5,033	3,709	4.6%	2.9%
alt.games.warcraft	23,849	20,135	2.6%	1.4%
Total:	849,955	689,562		

# Statistics

Table 2: Success rates for quote assignment

Name	messages	with quotes	% un-assigned	% unassigned per level
rec.gambling.sports	34,848	17,628	26.9%	17.2%
alt.fan.noam-chomsky	62,703	52,801	10.9%	6.1%
alt.alien.research	58,189	48,189	18.3%	6%
rec.music.classical.recordings	147,866	119,269	8.4%	5.1%
rec.sport.swimming	15,980	12,799	6.9%	4.4%
alt.support.marriage	76,183	69,955	7.1%	3.8%
rec.photo.digital	279,846	234,497	6.1%	3.6%
news.software.readers	31,874	26,480	5.5%	3.5%
rec.audio.tech	24,645	18,868	5%	3.2%
alt.music.oasis	21,883	14,989	4.7%	3.1%
alt.coffee	68,056	50,243	4.4%	3%
news.software.nntp	5,033	3,709	4.6%	2.9%
alt.games.warcraft	23,849	20,135	2.6%	1.4%
Total:	849,955	689,562		

# Web-derived data

---

- is clearly useful data...

but...

- clean-up procedure is time-consuming

# Messiness of data

---

- Clean-up process does not result in a well-balanced corpus
- Next to no metatextual data is available
- Authorship is very difficult to determine
- Influence of spam and troll posts

# Conclusion

---

- Web-derived data can complement existing corpora
- But: This should only be an intermediary step!
- Forcing “new data” into old formats
- “Cleaning up data” -> problematic
- We need new ways of dealing with data!

---

**Thank you!**