

# HG2052

## Language, Technology and the Internet

### Text and Meta-Text

Francis Bond

**Division of Linguistics and Multilingual Studies**

<http://www3.ntu.edu.sg/home/fcbond/>

bond@ieee.org

Lecture 11

# Text and Meta-text

---

- Explicit Meta-data
  - Keywords and Categories
  - Rankings
  - Structural Markup
- Implicit Meta-data
  - Links and Citations
  - Tags
  - Tables
  - File Names
  - Translations

# Explicit Metadata

---

- You can get information from metadata within documents
  - When they are accurate they are very good
  - They are often inaccurate
    - \* Sometimes deliberately deceitful
    - \* More often incomplete or out-of-date

*Never attribute to malice that which is adequately explained by stupidity.*

Hanlon's Razor

*You have attributed conditions to villainy that simply result from stupidity*

Robert A. Heinlein (1941) *Logic of Empire*

# HTML Metadata

---

- Most document types contain metadata of some description:

```
<head>
  <title>CSLI LinGO Lab</title>
  <META HTTP-EQUIV="Content-Type" CONTENT="text/html; charset=iso-8859-1">
  <meta http-equiv="Content-Style-Type" content="text/css">
  <meta name="keywords" content="linguistic grammars online,
  LinGO, computational linguistics,
  head-driven phrase structure grammar, hpsg, natural language processing,
  parsing, generation, augmentative and alternative communication, aac,
  LinGO Redwoods, multiword expressions, MWE, grammar matrix">
  <meta name="description" content="This page provides information about
  the CSLI Linguistic Grammars Online (LinGO) Lab at Stanford
  University.">
```

- Should we also extract out this data, or is metadata too unreliable to consider using?

# PDF Metadata

---

➤ Checkout this file (now look at earlier weeks)

InfoKey: Creator

InfoValue: xetex(k) 5.98 Copyright 2009 Radical Eye Software

InfoKey: Title

InfoValue: Lecture 11:Text and Meta-Text

InfoKey: Author

InfoValue: Francis Bond

InfoKey: Producer

InfoValue: GPL Ghostscript 8.71

InfoKey: Keywords

InfoValue: Language, Technology, Internet

InfoKey: Subject

InfoValue: HG2052/HG252: Language, Technology and the Internet

InfoKey: ModDate

InfoValue: D:20120315121040+08'00'

InfoKey: CreationDate

InfoValue: D:20120315121040+08'00'

PdfID0: 39bb293fa576c18e1ae64480cb8974

PdfID1: 39bb293fa576c18e1ae64480cb8974

---

NumberOfPages: 23

# HTML Metadata

---

- HTML Metadata is generally considered unreliable
  - Authors don't see it, so they don't update it
  - As it is unseen, it is easy to lie in the MetaData

It wasn't long before webmasters with no scruples saw an opportunity to gain favour with the search engines by adding in keywords that did not pertain to the content of their pages. Various tactics were thought up to get ranked higher for certain keywords, and an entire industry sprang up to optimise search engine positioning. This was, in effect, cheating, and “keyword spamming” became a serious problem for search engines, who vainly attempted to add filters that would notice when a webmaster was loading up on the wrong keywords.

# Keywords and Categories

---

- Sites with **visible tags** are more trustworthy/reliable
- Tags within blogs/photo cites
- Keywords in journals and conferences



## Example tags from Science Professor

---

# academia (109)  
# academic novels (9)  
# accounting nightmares (10)  
# administrative assistants (7)  
# adviser-student (69)  
# attempt at humor (18)  
# awards (7)  
# bizarre (56)  
# blogging (22)  
# books (23)  
# broader impacts (8)  
# career issues (27)  
# cats (19)  
# citations and citation index (19)

# Rankings

---

- Another good source of meta-data is rankings/forums

**HG251**

**Q A** Solved

- **Sentiment Analysis** tries to judge whether text is favorable or unfavorable
  - Link text to rankings for data
  - Link posts to tags for usefulness in QA

## hungry go where

---

Overall: 7 Recommend.

I spent about S\$10 Per Person

Food/Beverage: 6

Ambience: 5

Value: 9

Service: 5

Cheap but not very cheerful

10 June, 2010

Absolutely love this neighbourhood eatery, mainly because I have been eating here since I was a child, so it brings back many happy memories. Granted, service is kind of lacking but a cheap and yummy home-style meal can always be had. Must-haves for me are the Honey Pork (love the 3 or 4 little green peas they garnish it with), Ayam Buah Keluak, Bakwan Kepeting (meatball soup) and Sayur Lodeh. The Otak and Ngor Hiang are not bad too.

# Links and Citations

---

- Citation frequency can be used to measure the **impact** of an article.
  - Simplest measure: Each article gets one vote – not very accurate.
- On the web: citation frequency = **inlink count**
  - A high inlink count does not necessarily mean high quality ...
  - ...mainly because of link spam.
- Better measure: **weighted** citation frequency or citation rank
  - An article's vote is weighted according to its citation impact.
  - This can be formalized in a well-defined way and calculated.

- Structural Markup gives useful cues
  - Words in headers are often good keywords
  - TableOfContents!
    - 1 Review
      - 1.1 Language Identification
      - 1.2 Normalization
    - 2 Text and Meta-text
      - 2.1 Implicit Tags
      - 2.1 Explicit Tags

# Implicit Metadata

---

- You can get clues from metadata within documents
  - as they are non-intended, they tend to be noisy
  - but they are rarely deceitful

# Tags

---

- Hypertext Anchors and other formatting gives phrase boundaries

whereas McCain is secure on the topic, Obama  
<a>[VP worries about winning the pro-Israel vote]</a>

[NP [NP Libyan ruler]  
<a>[NP Mu 'ammar al-Qaddafi]</a>] referred to

Mainly NPs

- This can be very useful in restricting parser possibilities

Valentin I. Spitkovsky, Daniel Jurafsky, Hiyan Alshawi (2010) *Profiting from Mark-Up: Hyper-Text Annotations for Guided Parsing* ACL

# Tables

---

➤ You can learn many things from Tables

➤ for example, categories

Vehicle	Price	Manufacturer	Type	Rating
Raum	XXX	Toyota	Hatch-back	Solid
icw30	XXX	Hyundai	Station Wagon	Exciting
Corolla	XXX	Toyota	Station Wagon	Solid
Camry	XXX	Toyota	Sedan	Bland

Toyota  $\subset$  manufacturer



# File Names

---

➤ How to find definitions?

➤ Look for files called [glossary](#), [dictionary](#), ...

➤ Is there an English version of this?

➤ <http://nlpwww.nict.go.jp/wn-ja/index.ja.html>

➤ <http://nlpwww.nict.go.jp/wn-ja/index.en.html>

# Translations

---

- A translation into another language can be seen as markup

## Bracketed Glosses

---

GPS

通用 回解决者

2 8 附录：（通用 回解决者》（G P S）计算机程序解决・・河内塔。

{2069:corpus0.txt}

SomeThoughtsConcerningEducation

教育漫话

笔者在认真阅读洛克的教育著作《教育漫话》（S o m e T h o u g h t s C o n  
《关于理解的指导》以及《贫穷儿童劳动学校计划》（P l a n o f W o r k i n

{25094:corpus0.txt}

# Cross-lingual Disambiguation

---

(1)  $I_1$  saw<sub>2</sub> the kid<sub>3</sub> with a telescope<sub>4</sub>

(2)  $\phi_1$  望遠鏡<sub>4</sub> で 子供<sub>1</sub> を 見た<sub>2</sub>  
bouenkyou de kodomo wo mita  
NULL telescope with child ACC see-past  
With the telescope, I saw the kid.

- We can disambiguate the PP attachment: *de* only modifies verbs
- We can disambiguate the verb *see/saw*: *mita* is only “see”
- We can resolve the zero pronoun: It must be the speaker.

# Query Data

---

AOL user 2708:

- revenge tactics
- the woman's book of revenge
- dirty tricks for chicks
- ...
- locatecell.com
- what can i do to an old lover for revenge
- mean revenge tactics
- death records in hampstead

# Wikipedia Redirections

---

- Alternative names (*Edison Arantes do Nascimento* → *Pelé*).
- Abbreviations (*DSM-IV* → *Diagnostic and Statistical Manual of Mental Disorders*).
- Alternative spellings or punctuation. (*Colour* → *Color*; *Al-Jazeera* → *Al Jazeera*).
- Likely misspellings (*Condoleeza Rice* → *Condoleezza Rice*).
- Plurals (*Greenhouse gases* → *Greenhouse gas*).
- Related words (*Symbiont* → *Symbiosis*).
- Representations using ASCII characters (*Kurt Goedel* and *Kurt Godel* → *Kurt Gödel*).

## Cross Wikipedia Links

---

**en** Forensic linguistics

**ca** Lingüística forense

**cs** Forezní lingvistika

**de** Forensische Linguistik

**es** Lingüística forense

**nl** Forensische taalkunde

**no** Forensisk lingvistikk

**tr** Adli dil bilimi

**zh** 司法语言学

# Deliberately Deceit: Phishing

---

- **Phishing** is a way of attempting to acquire information such as usernames, passwords, and credit card details by masquerading as a trustworthy entity in an electronic communication.
- Communications typically pretend to be from social web sites, auction sites, online payment processors or IT administrators.
- Phishing is typically carried out by e-mail spoofing, linking users to a fake website whose look and feel are almost identical to the legitimate one.
- Phishing is an example of social engineering.
- A phishing technique was described in detail in 1987, and (according to its creator) the first recorded use of the term *phishing* was made in 1995.



## From my own spam box

---

System Administrator s28407548@tuks.co.za via srs.ieee.org  
to undisclosed recipients

You have exceeded the storage limit on your mailbox.

You will not be able to send or receive new mail until  
you upgrade your email quota.

Click the below link and fill the form to upgrade your  
account.

<http://millerofficetrailers.com/forms/use/hepldesk/form1.html>

System Administrator  
192.168.0.1

# The fake form

File Edit View History Bookmarks Tools Help

me, 2.... phishi... What ... Phishi... APWG APWG... ht...l

Please fill in all fields marked with a \*

	Email	<input type="text"/>
	User Id	<input type="text"/>
	Password	<input type="password"/>
	Confirm password	<input type="password"/>

Submit Form Reset Form

POWERED BY  
php  
FormGenerator

✕ Z Keep ⭐⭐⭐⭐⭐

# Some Distinguishing Features

---

- Surprisingly many grammatical mistakes
- Spoofed URLs
- Ultimatums
- Weird misspellings: [NTU.edu.org](http://NTU.edu.org)

# Why is this important?

---

- The 1990s started a revolution in empirical linguistics
  - New insights come from **Data Mining** large text collections
    - \* Corpus Linguistics
    - \* You can do with a computer what you can't do with paper
  - New tools come from supervised **Machine Learning**
- Annotation is expensive and tedious to do
- We want to get annotation for free
- People appreciate clever ideas
- In Singapore multiple-languages can serve as annotation, ...