Social Digital Discourse: New Challenges for Corpus- and Sociolinguistics

Josef Schmied
English Language & Linguistics
Chemnitz University of Technology
www.tu-chemnitz.de/phil/english/Schmied
josef.schmied@phil.tu-chemnitz.de
Introduction: SDD= social digital discourse everywhere

“The Falles (in Valencian) or Fallas (in Spanish) are a Valencian traditional celebration in praise of Saint Joseph in Valencia, Spain. The term Falles refers to both the celebration and the monuments created during the celebration. ...

The ninots and their falles are developed according to an agreed upon theme that was, and continues to be a satirical jab at anything or anyone unlucky enough to draw the attention of the critical eyes of the fallers — the celebrants themselves.”

http://en.wikipedia.org/wiki/Fallas (20/03/11)
El retor fa l'omilìa
amb la parella que es casa.
Sense adonar-se del que passa:
Tot és fruit de la mentida.

Els sogres volen casar-los
da la fí, sigua com sigua.
Perquè a veure estos dos "cardos"
una altra volta qui els lliga.
La Fada de l'internet
És la bruixa de huí en dia.
La xarxa farcida és
de fraus i pornografia.

Aquesta Fada tan guapa
Vol que naveguem per Internet,
Per a que caiguem a la xarxa
perduts pels seus indrets.
Examples: multiple identities
Social Graph Platform Wars

Google / Open Social
("The Empire"?, "Coalition of The Willing")

Facebook
("Rebel Alliance"?)

MySpace
("Clone Army"?)

Photobucket

= announced Platform / API
☆ = unannounced Platform / API
Media concepts and SDD

- social networks: “in the Paston letters”? - social network space (sns) is digital/virtual
- digital discourse: Skype?, telephone?
- computer-mediated communication? forums digital networks involve several participants

→ digital discourse in social networking services?

“Discourse is a term used to describe networks of ideas about reality that have been developed in specific social contexts, in line with the interests of the social actors in those contexts.”

http://newmedia.wikia.com/wiki/Discourse (20/03/11)
Media concepts: sns

“A social networking service is an online service, platform, or site that focuses on building and reflecting of social networks or social relations among people, e.g., who share interests and/or activities. A social network service essentially consists of a representation of each user (often a profile), his/her social links, and a variety of additional services. Most social network services are web based and provide means for users to interact over the internet, such as e-mail and instant messaging. Although online community services are sometimes considered as a social network service. In a broader sense, social network service usually means an individual-centered service whereas online community services are group-centered. Social networking sites allow users to share ideas, activities, events, and interests within their individual networks.”

http://newmedia.wikia.com/wiki/Social_networking (20/03/11)
A user profile (userprofile, or simply profile when used in-context) is a collection of personal data associated to a specific user. A profile refers therefore to the explicit digital representation of a person's identity. A user profile can also be considered as the computer representation of a user model.

A profile can be used to store the description of the characteristics of person. This information can be exploited by systems taking into account the persons' characteristics and preferences. For instance profiles can be used by adaptive hypermedia systems that personalise the human computer interaction.

http://en.wikipedia.org/wiki/User_profile (20/03/11)
Linguistic concepts and SDD

- Can we maintain old concepts? Social network
- Better data for old questions? Null subjects (Dana)
- Better data collection? Twitter corpus (Sascha)
  - Quicker since digital?
  - More directly accessible and still, users are unaware
    (no observers’ paradox, but an ethical problem?)
  - More current, topical?
  - More oral?

→ Quick, big but dirty?
  - Less stratified since less control of variables?
Linguistic concepts: genre/text-type

genre, text-type discussion:
■ Is “digital” one variable or several?
■ Is there one variety “(social) digital English”?
→ depends on the language users, the community of practice, the discourse community
Linguistic concepts: disc. community

A discourse community, Swales’ defining characteristics:

- “has mechanisms of intercommunication among its members.
- uses its participatory mechanisms primarily to provide information and feedback.
- utilizes and hence possesses one or more genres in the communicative furtherance of its aims.
- in addition to owning genres, it has acquired some specific lexis.
- has a threshold level of members with a suitable degree of relevant content and discoursal expertise.
- has a broadly agreed set of common public goals.”

http://en.wikipedia.org/wiki/Discourse_community (20/03/11)
“The blogosphere is made up of all blogs and their interconnections. The term implies that blogs exist together as a connected community (or as a collection of connected communities) or as a social network in which everyday authors can publish their opinions. Since the term has been coined, it has been referenced in a number of media and is also used to refer to the Internet.”

http://en.wikipedia.org/wiki/Blogosphere (20/03/11)
Forum participants as a community of practice
Linguistic concepts: SDD

clarification of concept: social digital discourse

- special case of digital discourses
  - social, social functions + default is the user-group (not only the option, like email)

- prototype concept (with fuzziness!):
  - central: multi-nodal (Wikipedia)
    - many active “editors” + many passive “consumers”
  - less central: focussed (Twitter)
    - few active producers + many passive “followers”
  - peripheral: peer2peer (email)
    - few active producers + few passive recipients
Linguistic concepts: differences

linguistic differences in digital discourse networks

- text seize: Wikipedia > Facebook > Twitter
- text number: Wikipedia < Facebook < Twitter
- text type: informative - persuasive?? - narrative?
- in-group/subculture: Twitter specific > Wikipedia opening Facebook both: personal hybrid, company opening

- text function focus
  - on affiliation: Twitter
  - on community building/maintenance: Facebook

hybrids and “multi-channels”,
Facebook from Wiki-like to email-like → Will Facebook break apart?
Challenges for sociolinguistics

Socio-biographical data are constructed ("persona")
private - public spaces,
personal - company profiles
emphasises the constructivist view of language

"assumed" identities (multiple identities):
How real, how consistent is a persona profile?
What is the role of conscious/unconscious
writer identities for language analysis / change?

→ a logical expansion of sociolinguistic trends:
from the language by males to male language ...
Challenges for corpus-linguistics

easy collection: quick, big, but dirty?
corpus quality depends on the social stratification
corpus size depends on frequency of phenomenon analysed
occurrence of linguistic structures is restricted by media features, e.g. tweet length in Twitter ("shorthand")

+ abbreviations, contractions are easy
- complex tenses, heavy NP modifications are difficult

→ well-known "caveats" apply:
   evaluate your data critically before you draw wide-rangings conclusions
**Opportunities: teaching perspectives**

- attracting students by “going with the customers”
- integrating real life into the academic world - leaving the “ivory tower”
- drawing students into a research approach from their own life-experience
- reaching students in their digital “reality”
Opportunities: teaching perspectives

Analysing the correlation between greetings, pre-farewells and text acts in our #EEE email data

Finally having the pictures of the complete staff online: http://www.linguistics.uni-bonn.de/people/ #bael

Choosing which coded features of our email corpus we will present at #JRM2011 in Essen #EEE

Had a great meeting with the #Stepin team! #BAEL #cooperation

Name Bonn Linguistics
Location Bonn, Germany
Web http://www.linguistics.uni-bonn.de
Bio Bonn Applied English Linguistics - empirical research of linguistic phenomena

29 following 197 followers 18 listed

Tweets 168

Favorites

Following

RSS feed of linguisticsbonn's tweets
Opportunities: research perspectives

- data collection/sampling seems easy
- style continua: between spoken and written (expansion of email research)
- access to youth speak: lol (expansion of instant messaging and texting research)
- instant evidence of language change (e.g. Tagliamonte, Kerswill/Cheshire research)
This article presents an analysis of Instant Messaging (IM), a one-to-one synchronous medium of computer-mediated communication. Innumerable articles in the popular press suggest that increasing use of IM by teens is leading to a breakdown in the English language. The analyses presented here are based on a unique corpus involving 72 teenagers and over a million words of natural, unmonitored IM. In addition, a corpus of speech from the same teenagers is examined for comparison. Targeting well-known IM features and four areas of grammar, we show that IM is firmly rooted in the model of the extant language. It reflects the same structured heterogeneity (variation) and the same dynamic, ongoing processes of linguistic change that are currently under way in contemporary varieties of English. At the same time, IM is a unique new hybrid register, exhibiting a fusion of the full range of variants from the speech community—formal, informal, and highly vernacular.
Wikipedia: self-definition

Wikipedia ... is a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. Its 18 million articles (over 3.5 million in English) have been written collaboratively by volunteers around the world, and almost all of its articles can be edited by anyone with access to the site. Wikipedia was launched in 2001 by Jimmy Wales and Larry Sanger and has become the largest and most popular general reference work on the Internet, ranking around seventh among all websites on Alexa and having 365 million readers.

The name Wikipedia was coined by Larry Sanger and is a portmanteau of wiki (a technology for creating collaborative websites, from the Hawaiian word wiki, meaning "quick") and encyclopedia.

http://en.wikipedia.org/wiki/Wikipedia (20/03/11)
Wikipedia: sociolinguistic discourse on reliability, non-expert driven

“Although the policies of Wikipedia strongly espouse verifiability and a neutral point of view, critics of Wikipedia accuse it of systemic bias and inconsistencies (including undue weight given to popular culture),[11] and allege that it favors consensus over credentials in its editorial processes.[12] Its reliability and accuracy are also targeted.[13] Other criticisms center on its susceptibility to vandalism and the addition of spurious or unverified information,[14] though scholarly work suggests that vandalism is generally short-lived,[15][16] and an investigation in Nature found that the science articles they compared came close to the level of accuracy of Encyclopædia Britannica and had a similar rate of "serious errors".[17]

Wikipedia's departure from the expert-driven style of the encyclopedia building mode and the large presence of unacademic content have been noted several times. When Time magazine recognized You as its Person of the Year for 2006, acknowledging the accelerating success of online collaboration and interaction by millions of users around the world, it cited Wikipedia as one of several examples of Web 2.0 services, along with YouTube, MySpace, and Facebook.[18] Some noted the importance of Wikipedia not only as an encyclopedic reference but also as a frequently updated news resource because of how quickly articles about recent events appear.[19][20] Students have been assigned to write Wikipedia articles as an exercise in clearly and succinctly explaining difficult concepts to an uninitiated audience.[21]”

http://en.wikipedia.org/wiki/Wikipedia (20/03/11)
Wikipedia: “editor”?

Here, as in other human endeavours, it is evident that the active attention of many, when concentrated on one point, produces excellence.
—Goethe, The Experiment as Mediator between Subject and Object, 1772

In departure from the style of traditional encyclopedias, Wikipedia employs an open, "wiki" editing model. Except for a few particularly vandalism-prone pages, every article may be edited anonymously or with a user account. No article is owned by its creator or any other editor, or is vetted by any recognized authority; rather, the articles are agreed on by consensus.[47]

Contributors, registered or not, can take advantage of features available in the software that powers Wikipedia. The "History" page attached to each article records every single past revision of the article, though a revision with libelous content, criminal threats or copyright infringements may be removed afterwards.[52][53] This feature makes it easy to compare old and new versions, undo changes that an editor considers undesirable, or restore lost content. The "Discussion" pages associated with each article are used to coordinate work among multiple editors.[54] ... Computer programs called Internet bots have been used widely to remove vandalism as soon as it was made,[16] to correct common misspellings and stylistic issues, or to start articles such as geography entries in a standard format from statistical data.

http://en.wikipedia.org/wiki/Wikipedia (20/03/11)
Wikipedia: linguistic evaluation

**characteristics:**
the model of many semi-specialists producing excellence in popular academic writing, user-driven text production

**advantages:**
favours intellectual discourse, reveals differences of opinion, helps clarify concepts ...
in a relatively closed discourse community? student seminar group?
good data for popular academic writing, some data for constructivist text production in history (terminology, hedging)

**disadvantages:**
persuasive elements in informative texts?
"consensus over credentials"
→ students agree against the professor’s expert knowledge?
Facebook: discourse & identity

“Users can create profiles with photos, lists of personal interests, contact information, and other personal information. Users can communicate with friends and other users through private or public messages and a chat feature. To allay concerns about privacy, Facebook enables users to choose their own privacy settings and choose who can see specific parts of their profile.[67] The website is free to users, an generates revenue from advertising, such as banner ads.[68] Facebook requires a user's name and profile picture (if applicable) to be accessible by everyone. Users can control who sees other information they have shared, as well as who can find them in searches, through their privacy settings.[69]

The media often compare Facebook to MySpace, but one significant difference between the two websites is the level of customization.[70] Another difference is Facebook's requirement that users give their true identity, a demand that MySpace does not make.[71]"

http://en.wikipedia.org/wiki/Facebook (20/03/11)
Facebook: attitudes

Thema: How Social Networking is negatively effecting free discourse

Es werden alle 9 Beiträge angezeigt.

Nicholas
The rise of social networking seems to exist to create networks of like minded people. But the way they are set up, is that a small group of people with extreme views can actually censor and block those with opposing views from being heard.

Has anyone else experienced this sort of ghetto mentality that is being created on social networking sites.

vor über einem Jahr

Tony
And not so much affecting free discourse - but affecting negatively rational dialogue by negating the 'face to face' encounter. Discourse is not simply about exchaning ideas through words on a page, it is also about relationships and relationships are based upon a physical encounter with the Other.

vor über einem Jahr

Jenny
It's often not possible to have a physical encounter with everyone you would like, when you would like. Social networking, like other forms of communication or relationship, has both unique and shared characteristics. The nature or quality of interaction, information shared, relationships, ... depends upon what individuals and groups make them.
Facebook: linguistic evaluation

characteristics:
very different communication channels = audiences = styles?

advantages:
good for collective memory (symposium)
good data for in-group language = colloquial, implicit narrative style
good data for company2customers language = formal persuasive style
relatively wide discourse community or different discourse communities (breaking apart?)

disadvantages:
too much diversity → Will Facebook break apart?
restricted access for user - unlimited access for provider (prototype fairy/witch)
Twitter: discourse with followers

“Twitter is a website, owned and operated by Twitter Inc., which offers a social networking and microblogging service, enabling its users to send and read messages called tweets. Tweets are text-based posts of up to 140 characters displayed on the user's profile page. Tweets are publicly visible by default; however, senders can restrict message delivery to just their followers. Users may subscribe to other users' tweets – this is known as following and subscribers are known as followers[8] or tweeps[9] (Twitter + peeps).”
http://en.wikipedia.org/wiki/Twitter (20/03/11)

“Some examples of a discourse community might be those who read and/or contribute to a particular academic journal, or members of an email list for Madonna fans. Each discourse community has its own unwritten rules about what can be said and how it can be said ... members of the email list may or may not appreciate a Freudian analysis of Madonna’s latest single. Most people move within and between different discourse communities every day.”
http://en.wikipedia.org/wiki/Discourse_community (20/03/11)
**Twitter: linguistic evaluation**

**characteristics:**
“focussed” communication on “celebrity”, company, topic

**advantages:**
good for influence of individuals/celebrities on language change, innovation/diffusion studies among “followers”

relatively closed discourse community with restricted/specialised language

**disadvantages:**
relatively short, informal, written-like-spoken (“decay”?)
Conclusion: social digital media in English Studies

- new data to refine old concepts
- new data for old linguistic analyses
- new attractions to new students
- new practical opportunities in teacher - student, student - student discourses
Wikipedia: quality of writing

Because contributors usually rewrite small portions of an entry rather than making full-length revisions, high- and low-quality content may be intermingled within an entry. Critics sometimes argue that non-expert editing undermines quality. For example, Roy Rosenzweig had several criticisms of its prose and its failure to distinguish the genuinely important from the merely sensational. He said that Wikipedia is "surprisingly accurate in reporting names, dates, and events in U.S. history" (Rosenzweig's own field of study) and that most of the few factual errors that he found "were small and inconsequential" and that some of them "simply repeat widely held but inaccurate beliefs", which are also repeated in Encarta and the Britannica. However, he made one major criticism.

Good historical writing requires not just factual accuracy but also a command of the scholarly literature, persuasive analysis and interpretations, and clear and engaging prose. By those measures, American National Biography Online easily outdistances Wikipedia.\[107]\n
... A 2005 study by the journal Nature compared Wikipedia's science content to that of Encyclopædia Britannica, stating that Wikipedia's accuracy was close to that of Britannica, but that the structure of Wikipedia's articles was often poor.".\[17]\n
http://en.wikipedia.org/wiki/Wikipedia (20/03/11)