Social Digital Discourse: New Challenges for Corpus- and Sociolinguistics

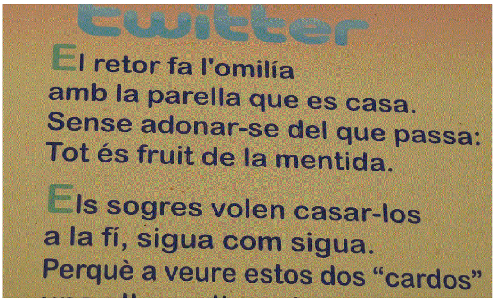
Josef Schmied
Chemnitz University of Technology, Germany

Abstract
This contribution attempts to link new forms of discourse with old linguistic sub-disciplines, in particular corpus- and sociolinguistics. It shows that social digital discourses can enrich the discussion of linguistic concepts as they pose new challenges for linguistic researchers – but they also offer new opportunities. It focuses on Wikipedia, Facebook and Twitter to show how new technical platforms can help us to expand our database to shed new light on old linguistic questions. This approach can make available data from places in Africa and Asia that are otherwise less accessible to the empirical English linguist. The study of social digital discourse may also help to attract media-oriented types of students to linguistic analysis when they realise that social media have something to contribute to linguistics just as linguistics has something to contribute to the new media world, which students and colleagues may perceive as outside of our academic world. This contribution aims to prove that they are not ...

Keywords
Corpus linguistics, empirical methods, Web-as-corpus, discourse analysis, Internet linguistics

1. Introduction: the Beauty & the Beast

The internet and its most recent communication forms and platforms such as Facebook and Twitter play an increasingly important role in popular public discourses nowadays. While students show great enthusiasm, many researchers have not yet realised the opportunities to collect new data and to attract students to standard questions in English language and linguistics, for instance. Of course, we must be aware of the Janus-faced nature of the phenomenon: This ambivalence of a fairy and a witch was beautifully captured during the famous Fallas in Valenica in 2011 by a huge *ninot* (Valencian for puppet or paper-mâché artistic monument) presenting the internet as a new version of the beauty and the beast (Photo 1), assembling all the modern stereotypes, - in the end "it's all a pack of lies" (Photo 2).



twitter

El retor fa l'omilía
amb la parella que es casa.
Sense adonar-se del que passa:
Tot és fruit de la mentida.

Els sogres volen casar-los
a la fí, sigua com sigua.
Perquè a veure estos dos "cardos"

Photo 1+2 The internet as Beauty&Beast and Twitter "discussion" during the Fallas in Valencia 2011 ("it's all a pack of lies")

This popular discourse has even predicted "wars" between the most successful protagonists in the internet (Fig. 1).

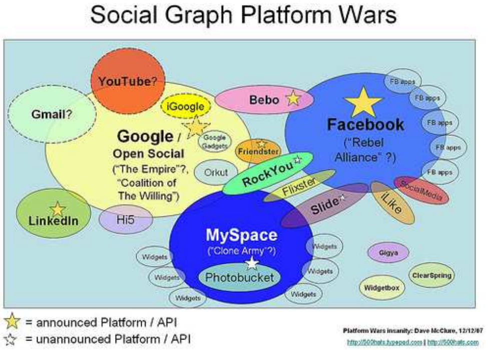


Fig. 1: Social Graph Platform Wars
Source: http://farm3.static.flickr.com/2248/2105757707_99dec8729a.jpg

If we want to take up the academic challenge, we do not have to start from scratch. For more than twenty years "computer-mediated communication" has been analysed in linguistics (e.g. Herring ed. 1996) and for more than ten years "the internet" has been discussed in popular and scientific scholarly contributions (e.g. Crystal 2006 and 2011 and Baron 2003 and 2008). Multimodal and semiotic approaches to digital discourses have been proposed by media specialists and linguists (e.g. Ferenčík 2011). Some scholars (e.g. Myers 2010) have even started the analysis of web 2.0 discourse; but many students and colleagues still think that this is not worth their linguistic efforts, although the first handbooks have become available on "different aspects of computer-mediated communication, such as electronic mail, instant message, chat, discussion forum, blog, video conferencing, YouTube, web-based learning and SMS, as well as aspects of behavior typically associated with online discourse like flaming, scamming, trolling, cyberbullying, language mixing, repelling and creativity" (Taiwo in his preface to Taiwo ed. 2010). Some scholars have shown that one can have "fresh perspectives on new media sociolinguistics", the subtitle for the editor's introduction in Thurlow/Mroczek eds. (2011: xix-xliv), a volume with sexy subsections such as "Multimodality: Beyond Language and into the Bedroom" (ibid: xxv) or "the notion of Foucauldian *discourses* – which we dub F-*discourse* as opposed to L-*discourse* ("language in use") ..." (ibid: xxvi). This contribution attempts to demonstrate that mainstream linguistics today, and in particular corpus- and sociolinguistics, can take up the challenge offered by the new social digital discourse.

2. Concepts

In the following section, I hope to demonstrate that key concepts from media and linguistic studies can be combined to shed new light on the modern forms of internet communication, particularly social digital discourse. In appropriate cases, I will use definitions from Wikipedia, since this platform will serve as an example later-on, in particular since text quality can be discussed on the basis of the evidence presented in this section.

2.1. **Media concepts and their linguistic applications**

2.1.1. Social networking service

Computer mediated communication, e.g. in email, in forums, etc., is already a well-established concept in applied linguistic research, and digital is almost tautological in modern communication today, while the focus on digital social networks is relatively new. The concept of social digital discourse is not yet defined in current dictionaries, not even the most current Wikipedia entries. But it obviously implies communication between several participants and a certain technology platform. This leads us to a useful Wiki definition under the keyword "social networking service":

A **social networking service** is an online service, platform, or site that focuses on facilitating the building of social networks or social relations among people who, for example, share interests, activities, backgrounds, or real-life connections. A social network service 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. Online community services are sometimes considered as a social network service, though 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)

Interestingly, the links offered in this description do not lead to entries such as "social media" or "web 2.0", which surface in many internet discussions today, but they demonstrate which Wikipedia entries are considered related concepts, such as "social networks", which can be useful for our linguistics analyses.

2.1.2. Social digital networks

Social networks have been a topic in socio-historical linguistics for a long time. Thus when we search for "social networks", "linguistics" and "English" in Google, we find the Paston letters, a very old social network from the 15th century and a good data base for linguistic research, since letters include more informal language styles, which are difficult to analyse but important for a better understanding of language development. The social networks we are interested in today are part of digital discourse, which today ranges from Skype and texting to micro-blogging and status updates on Facebook. These special languages and styles have are interesting comparative data for linguistics from different perspectives and from different parts of the world. Of course, the famous "digital divide" between the "haves" and the "have-nots", especially in its global version, is clearly visible in all maps demonstrating the internet penetration of the world (Fig. 2). Africa and parts of Asia do not (yet) have the same opportunities and (perhaps) threats as Europe and North America – and it is there where research in "New Englishes" is developing most vigorously (Schneider 2012: 366).

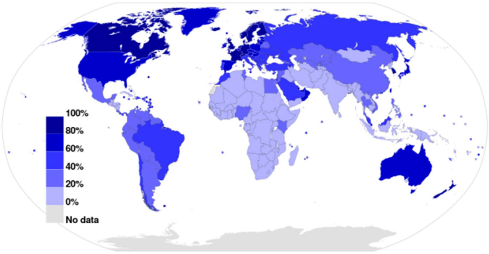


Fig. 2: World Map of Internet Penetration Source:http://en.wikipedia.org/wiki/File:InternetPenetrationWorldMap.svg

Of course, the issues of "digital technology" have also been discussed for Africa (e.g. Alzouma 2005), where some parts of the internet are more easily accessible than others. On-line newspapers are well established, and Anchimbe (2010) was able to use *The Post Newspaper Cameroon* (www.postnewsline.com) and particularly the interactive features there to analyse how the virtual community constructs its "Diaspora Anglophone Cameroonian identity online", which can be seen as African digital discourse. The more modern social digital networks are less frequently used, as Fig. 3 illustrates.

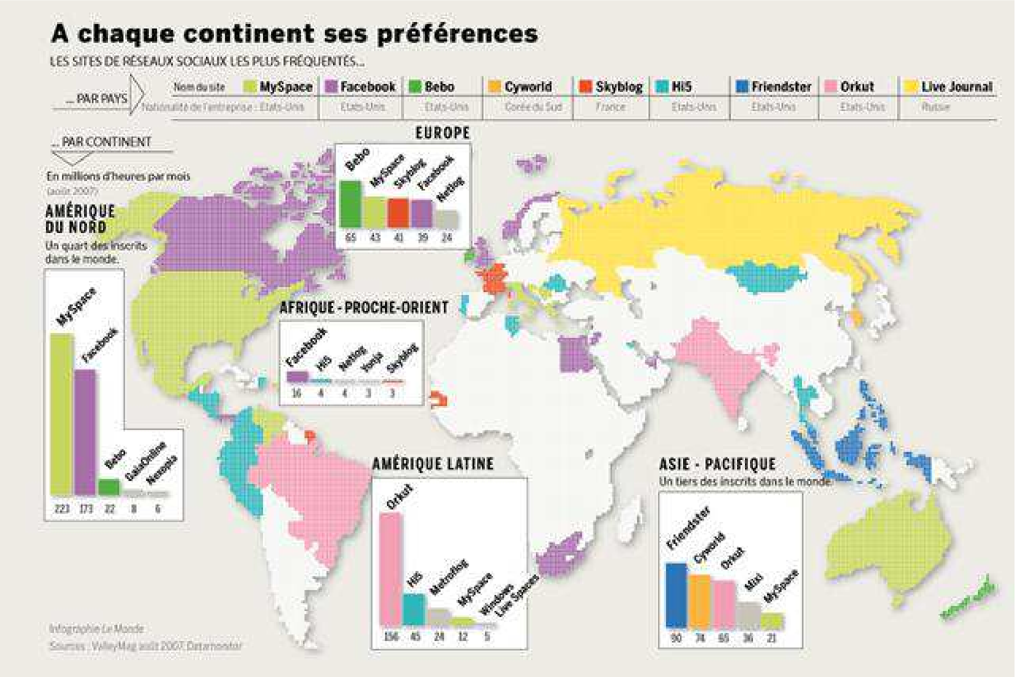


Fig. 3: Social network preferences by continent (*Le Monde* 2008)
Source: http://www.webthreads.de/article-data/uploads/2008/01/socialnetworks.jpg

In all these comparative considerations, we have to be aware of the fact that China is a special case: even if the international social networking services are available, they face strong “national” competition (as the respective Wikipedia pages summarize):

Baidu can rightly be called the “Chinese Google”, since it offers many services, including a Chinese language search engine for websites, audio files, and images. It also offers a special community service, the “Chinese Wikipedia” Baidu Baike, an online collaboratively written encyclopaedia and a searchable keyword-based discussion forum.

QQ is China's most popular IM software, a mixture of sms and email. It is linked with Qzone, a social networking website, which permits users to write blogs, keep diaries, send photos, and listen to music.

Renren is often referred to as the "Chinese Facebook". It also has a variety of functions with its own characteristics and, similar to Google, it changes the web page's design on special days, like Spring Festival or National Day.

These particular circumstances should be taken into consideration when comparing social digital discourse world-wide. Other national organisations (like vkontakte.ru in Russia) may have many users, but are not substantially different from the platforms described here.

2.1.3. User profile

From a corpus- and particularly sociolinguistic perspective, the most important feature of modern social media is the user profile, which is included in all three social media discussed below, i.e. Wikipedia, Facebook and Twitter. This is a collection of personal data associated to a specific user, his or her identity; but it is important to remember that it is the user's own explicit chosen identity. Although social media (esp. Facebook) would like to force their users to reveal as much as possible of their real identity, this publically visible profile is, of course, rather "the reflection of the shadows on the wall", as in Plato's cave allegory. The following Wikipedia entry does not sound too sceptical about this:

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)

For a sociolinguistic categorisation and correlation of language features, it is unclear whether this "subjective" self-description is less valuable than the "objective" socio-biographical data used in traditional sociolinguistics.

2.2. **Linguistic concepts and their social media application**

2.2.1. Internet communication networks as discourse communities

The most obvious linguistic concept that links with the media concepts discussed so far is the discourse community, which has been developed by Nystrand, Perelman and particularly Swales in the context of academic discourse (cf. Schmied fc.).

A discourse community:

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

http://en.wikipedia.org/wiki/Discourse_community (25/01/13)

This definition is partly based on the much older concept of a speech community as a group of people who share a set of norms and expectations regarding the use of language. It has been at the heart of the linguistic and sociolinguistic debate on uniform versus diversified norms over the last fifty years since Gumperz, Chomsky and Labov. The application of the concept to internet discourse is unclear and the debate has hardly expanded beyond the well-known netiquette (network etiquette) debate. In some internet discourse, the term blogosphere is used for part of the internet community:

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 (25/01/13)

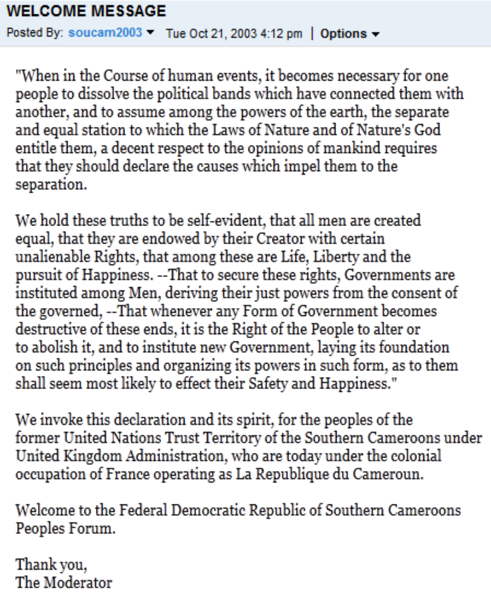
2.2.2. Forum participants as a community of practice

In most recent sociolinguistics, another potentially useful concept has gained attention in academic discussion, i.e. a community of practice, as defined in Wikipedia:

A **community of practice** (**CoP**) is, according to cognitive anthropologists Jean Lave and Etienne Wenger, a group of people who share a craft and/or a profession. The group can evolve naturally because of the members' common interest in a particular domain or area, or it can be created specifically with the goal of gaining knowledge related to their field. It is through the process of sharing information and experiences with the group that the members learn from each other, and have an opportunity to develop themselves personally and professionally (Lave & Wenger 1991). CoPs can exist online, such as within discussion boards and newsgroups, or in real life, such as in a lunch room at work, in a field setting, on a factory floor, or elsewhere in the environment.

http://en.wikipedia.org/wiki/Community_of_practice ((25/01/13)

When we look for such CoPs on the internet, we find them in different formats. In text 1, we see the Welcome page of a forum of Southern Cameroonians established as early as 2003. The "Moderator" evokes Cameroon's late colonial history, claiming that "colonial occupation" continues in "La République du Cameroun". The contrast to Francophone Cameroon establishes an Anglophone identity, in fact the forum's name SouCam refers to the part of the United Nations Trust Territory that voted to "unite" with the French part in the plebiscite of 1961. Such contrasting identities are typical for political opposition groups and provide a strong bond for communities of practice whose primary aim is obviously independence from (or at least greater autonomy in) Cameroon. The language used in the Welcome text is rather formal and intertextual, as this "declaration" style invokes the American Declaration of Independence.



WELCOME MESSAGE

Posted By: soucam2003 ▾ Tue Oct 21, 2003 4:12 pm | **Options ▾**

"When in the Course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to assume among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation.

We hold these truths to be self-evident, that all men are created equal, that they are endowed by their Creator with certain unalienable Rights, that among these are Life, Liberty and the pursuit of Happiness. --That to secure these rights, Governments are instituted among Men, deriving their just powers from the consent of the governed, --That whenever any Form of Government becomes destructive of these ends, it is the Right of the People to alter or to abolish it, and to institute new Government, laying its foundation on such principles and organizing its powers in such form, as to them shall seem most likely to effect their Safety and Happiness."

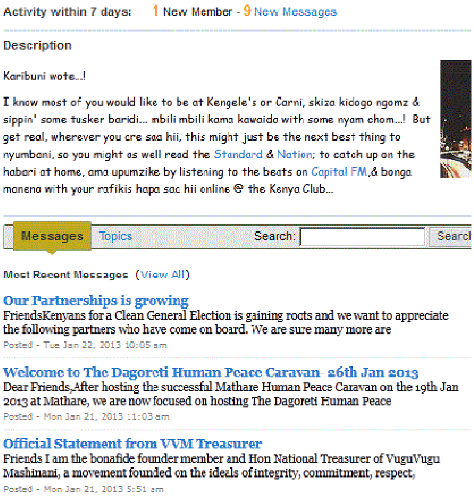
We invoke this declaration and its spirit, for the peoples of the former United Nations Trust Territory of the Southern Cameroons under United Kingdom Administration, who are today under the colonial occupation of France operating as La Republique du Cameroun.

Welcome to the Federal Democratic Republic of Southern Cameroons Peoples Forum.

Thank you,
The Moderator

Text 1: Welcome page for soucam yahoo forum (25/01/13)
Source: http://dir.groups.yahoo.com/group/AmbazoniaPeoples/message/1

Whereas the function of the soucam forum is typical, its language used in Text 1 is atypical. This becomes clear when we contrast it with Text 2, which is equally political (for "Clean General Elections"), but written in a much more oral style. The web community in this example identifies with Kenya and clearly practices a bilingual "life-style" and the colloquial "Welcome" includes the frequent code-switching between English and Kiswahili: The "Everybody Welcome" is in Kiswahili ("*Karibuni wote* ...!"), the *habari* ("News") from *nyumbani* ("Home") are obviously essential in order to create a common bond with friends outside of Nairobi and Kenya, and some Kiswahili words are integrated into English such as *nyam(a) chom(a)* ("Grilled Meat" or BBQ; with final vowels dropped) for *rafiki(s)* (with an English plural *-s* added). Similar phenomena have been found in other multilingual parts of the electronic world (such as Malaysia, cf. Hassan/Hashim 2009 and Norizah Hassan/Azirah Hashim/Phillip 2012).



Activity within 7 days: 1 New Member - 9 New Messages

Description

Karibuni wote...!

I know most of you would like to be at Kengele's or Carni, skiza kidogo ngomz & sippin' some tusker baridi... mbili mbili kama kawaida with some nyam chom...! But get real, wherever you are saa hii, this might just be the next best thing to nyumbani, so you might as well read the Standard & Nation; to catch up on the habari at home, ama upumzike by listening to the beats on Capital FM,& bonga manena with your rafikis hapa saa hii online @ the Kenya Club...

Most Recent Messages (View All)

Our Partnerships is growing
FriendsKenyans for a Clean General Election is gaining roots and we want to appreciate the following partners who have come on board. We are sure many more are
Posted - Tue Jan 22, 2013 10:05 am

Welcome to The Dagoreti Human Peace Caravan- 26th Jan 2013
Dear Friends,After hosting the successful Mathare Human Peace Caravan on the 19th Jan 2013 at Mathare, we are now focused on hosting The Dagoreti Human Peace
Posted - Mon Jan 21, 2013 11:03 am

Official Statement from VVM Treasurer
Friends I am the bonafide founder member and Hon National Treasurer of VuguVugu Mashinani, a movement founded on the ideals of integrity, commitment, respect,
Posted - Mon Jan 21, 2013 5:51 am

Text 2: Forum pages from Kenyaclub yahoo forum
Source: http://groups.yahoo.com/group/kenyaclub/ (23/01/13)

The great advantage of such internet data for the empirical linguist is that they are easily accessible, handy and even stratified: The data are directly accessible to the researcher, even though they are remote in time (2003) and place (Nairobi). Thus we can avoid the observer's paradox, which occurs when users adapt their styles as they perceive that their linguistic behaviour is observed; and since the texts are made available publically, analysing them does not seem to be an ethical problem. Internet data are topical, more oral and still written down, so that we do not need any transcription. However, how can we really compile a stratified corpus from different parts of the internet? And how do we deal with the textual variation? Should "digital discourse" be treated as one variable or several variables, or is it a genre? Is there one variety of social digital English?

2.2.3. Textual variation as genre or text-type?

Over the last few decades, several attempts have been made to differentiate texts according to their pragmatic functions. Early corpus compilations in the 1960s and 1970s used a classification of descriptive, narrative, expository and argumentative text-types. More recently, linguists have reinterpreted the traditional literary concept of genre as social action from a constructivist perspective. Neither has been applied convincingly in socio- and corpus-linguistic research (and the entries for text-type and genre in Wikipedia are not useful for our analysis of social digital discourse either). This means that we still have to use an ad hoc classification of styles in terms of textual variation as attention to the reader.

Today the extensive options of the internet make a comparison of different texts from the same source possible. Thus, the Daily News" from Dar es Salaam, Tanzania, has an on-line edition that we can read and comment on interactively, a Wikipedia entry with the basic historical and political background, a Facebook account that we can “Like”, and (Twitter) tweets that we can “follow” world-wide. This diverity allows us to monitor current events and language in Tanzania presented in different styles by different reporters and columnists. This is a welcome expansion of the traditional newspaper analysis used in the well-known ICE projects so far (Schmied 2011).

2.3. **Linguistic concepts applied to social digital discourses**

Social digital discourse today is defined by its social functions in the wide sense. In it, the default is a user group. In the following, we will apply our linguistic concepts to three well-known platforms with different types of discourses, i.e. Wikipedia, Facebook and Twitter. Their central idea is multi-nodal (one2many) communication, i.e. we may exchange emails with only one person, but it is not very useful to write a Wikipedia entry for one reader, construct a Facebook profile for one friend, or Twitter with one follower. All three platforms have a similar central "social" communication concept with an interactive, web2.0 component, but also some fuzziness: The central idea of Wikipedia is that this on-line encyclopaedia – ideally – has many active "editors" and many more passive readers. The central idea of Twitter is one active tweet producer and many passive "followers". Facebook has both options depending on the privacy settings: wide open visibility for company status-updates contrast with restricted in-group communication for party invitations.

Apart from these basic defining criteria, the three discourse platforms can also be characterised by different textual variables, like text length, number of texts, text-type and cultural background:

Wikipedia entries are usually long – and they are linked in hypertext format, so their length is not easy to define, potentially the reading is endless. Twitter texts are the shortest, a maximum of 140 characters, but they can also include links to more and longer texts. Facebook is a sort of compromise between a potentially very detailed personal lexicon, including its timeline history, and relatively short personal postings.

Obviously, the number of texts used for linguistic analysis must then be inverse: fewest in the case of Wikipedia, most in the case of Twitter.

The characterisation according to text type is relatively speculative: I would see Wikipedia as informative, as readers do not want to be persuaded; Facebook may be narrative and persuasive; Twitter is mainly narrative and instructive.

Similarly, the different platform discourses have different cultural backgrounds: Wikipedia is outgoing, the intended readership is broad and general; if information is "unsourced" or its "neutrality disputed", an entry is "flagged" and its editors may be "subjected to sanctions" (see below). By contrast, Twitter is "very in-group", since only followers receive the tweets. Facebook can, again, be both: the personal postings are in-group, the company status-updates are as general as possible.

Text functions focus on the product in Wikipedia, the presentation is ideally unbiased. The focus is on affiliation in Twitter, maintaining a thread ("fil") with followers. Facebook with both parts, the School Class Facebook and the Company Facebook, is a hybrid and it is multi-channel: it has Wiki-like functions and it has email functions.

3. **Challenges and new perspectives**

3.1. **Challenges for sociolinguistics**

Although the new data from social digital discourse may have attractions for sociolinguists, they also present new challenges. The socio-biographical data are a particular problem. Identities are, of course, constructed or "assumed", even multiple identities are possible. Thus they are more a "persona" than a real person; the gender issue (Baron 2004), e.g. male Facebook contributors posing as females, has been widely discussed in the media. So we must ask how real, how consistent a profile is, or how conscious or unconscious writer identity affects language behaviour and thus studies of language analysis and change. Primarily, such caveats concern the private sphere, but even Facebook has private and public spaces, personal and company profiles. Yet, it is also possible that language usage correlates well with assumed identities. All this emphasizes the constructivist view of language. When we analyse company web-communication, we know that it is a "constructed" profile: we are not interested in the identity of the person who actually wrote the text, we are interested in the "identity" the company wishes to portray. In my view, this is a logical expansion of sociolinguistic research away from studies based on supposedly objective data to studies incorporating more subjective indicators of identities.

In any case, the admittedly restricted sociolinguistic data from social media are better than the information we can hope for from many other texts from the internet.

3.2. Challenges for corpus-linguistics

The new opportunities for data collection from online discourse have been discussed on a broad social-science and comparative informatics basis (e.g. Goggins/Mascaro 2012).

The advantages of social digital data mentioned above are a great attraction for corpus-linguists. Corpus quality, of course, depends on the textual diversity and social stratification we can achieve. The corpus size necessary depends on the frequency of the phenomenon analysed. Since the text size is restricted in some media (especially Twitter cf. above), specific features and symbols are used, almost like shorthand. So it is easy to investigate abbreviations or contractions on the basis of a Twitter Corpus, but it is difficult to retrieve enough complex tenses, heavy noun phrase modifications – this is not Twitter usage. The most fruitful results can be expected for features that are typical of youth language, like *massive* intensification (cf. Martinez/Pertejo 2012). We have to be aware of these age- or media-specific over- and under-usages as special limitations or opportunities, if we do not want to be discredited as number crunchers. We have to evaluate our data critically before we draw wide-ranging conclusions.

An additional challenge may be retrieving the data effectively – if we do not want to resort to paste-and-copy techniques. How quickly and easily we can compile a corpus of texts from social media also depends on the technical design of these texts. Possible solutions must therefore be specific to the network service and have to be found and changed. Twitter has been tried and found relatively easy to use through its Application Programming Interface (API) in this respect (cf. below).

To sum up, although we have to be aware of the challenges and problems involved, the opportunities offered for innovative teaching and research perspectives are considerable, as well be shown below.

3.3. Teaching perspectives

The attractions of integrating social media into university teaching are obvious, since we may attract students approaching them as "customers". Integrating their real lives into the academic world also means leaving the "ivory tower". It even allows lectures to learn more from and about their students. We hope to reach students more personally in their digital "reality".

Ebner et al. (2010: 99) conducted a study on informal and process-oriented learning in Austria and conclude (in formal German nominal style) that

the successful use of microblogging and the increasing value that results for students and teachers from the use of microblogging is substantial. For the students this can be summarized in the following points:

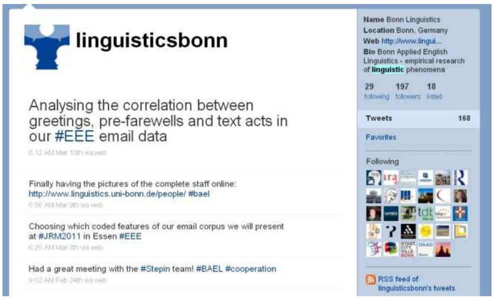
- Informal learning through informal communication.
- Support of collaboration.
- Feedback on thoughts.
- Suggestions to reflect one's own thoughts.
- Collaboration independent of time and place.
- Direct examination of thoughts and causes of learning.

For teachers the following factors are crucial.

- Current information on the status of learning.
- Possibility to steer the intervention in the learning process of individuals and groups.
- Possibility for immediate, direct feedback.
- Facilitation of student group work.
- Getting an impression of the learning climate.

These arguments were substantiated by my own observations during a recent project, where students organized all their discussions and meetings using Facebook – disregarding the Wiki project page. Maybe the Wiki stage will be used later for documentation, but the fact that students loved to organize themselves and their project work on the basis of new social media is most interesting.

Some English departments have already integrated social media into their staff-student communication channels, as can be seen from their websites. For example, Linguisticsbonn (Fig. 4) has been using Twitter for some time to invite students to Applied English.



linguisticsbonn

Analysing the correlation between greetings, pre-farewells and text acts in our #EEE email data
6:12 AM Mar 10th via web

Finally having the pictures of the complete staff online: http://www.linguistics.uni-bonn.de/people/ #bael
6:56 AM Mar 9th via web

Choosing which coded features of our email corpus we will present at #JRM2011 in Essen #EEE
6:20 AM Mar 8th via web

Had a great meeting with the #Stepin team! #BAEL #cooperation
9:02 AM Feb 24th via web

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

Fig. 4: Linguisticsbonn using Twitter
Source: http://www.linguistics.uni-bonn.de/ (23/03/11)

Whether using social media to distribute news also draws students into a linguistic research project that is based on their own life-experience with social media is another matter. However, a general critical assessment of teaching opportunities using Facebook is surprisingly positive and optimistic (Rambe 2012: 310):

The two [hierarchical and horizontal] discourses mentioned above have fundamental implications for pedagogical practice in higher education:

Academics should induct students into critical engagement, locating and interpreting the philosophy and ideologies behind different discourses they and peers activate. This could unlock student understanding of how disciplinary knowledge is constructed rather than passive reception of educator-generated content.

Educators should encourage learner discursive practices involving higher forms of knowledge (theoretical knowledge) as much as they deliberate on procedural issues. Facebook learning communities could be employed as vehicles for deconstructing theoretical propositions and perspectives through text-based interaction. As Salmon's (2000) five stage model of e-learning posits, it is the higher levels knowledge construction and development that allow for student selfregulation of on-task activities, responsibility for knowledge construction and cognitive growth.

Weak study skills and over-dependence on educators for information are addressed by developing an information sharing culture and valuing the strength of student contributions during collaboration. Through this, students learn to become principal knowledge brokers than information receivers. CDA exposed some hidden assumptions about power and implicit.

3.4. **Research perspectives**

Since data from social media seem to be available in abundance and sampling seems easy, we can use the data thus obtained to re-consider the old problems of style continua and English variation between spoken and written forms or preferences. This research focus is an expansion of up to 20 year old email research (Frehmer 2008). There is a considerable body of research on "lol" and smileys and several of our students are keen to expand early studies on instant messaging and texting. The popular linguist David Crystal has shown that this can provide instant evidence of language change. Others like Tagliamonte, Kerswill and Cheshire have shown the way for more thorough studies, as we can see from the example: “Linguistic Ruin? Lol! Instant messaging and teen language”. The fact that it is published in *American Speech* proves that new media language has entered into hard empirical linguistic research, as the following abstract shows:

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.

> Sali A. Tagliamonte and Derek Denis University of Toronto
>
> American Speech 2008 83(1):3-34; DOI:10.1215/00031283-2008-001
>
> 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.

Text 3: Abstract: Linguistic ruin? Lol! Instant messaging and teen language (Tagliamonte/Denis 2008)
Source: *American Speech* 83(1): 3-34; DOI:10.1215/00031283-2008-001

But this is only a beginning. In the following sections, we will look into three types of social digital networks and give examples of how they can be exploited for serious socio- and corpus-linguistic research in order to solve old questions of English language variation.

4. **Wikipedia**

4.1. **Non-expert driven discourse on reliability**

Wikipedia is one of the first (2001) and most prototypical examples of social digital discourse, which is evident from its 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.[3] Wikipedia was launched in 2001 by Jimmy Wales and Larry Sanger[4] and has become the largest and most popular general reference work on the Internet,[2][5][6][7] ranking around seventh among all websites on Alexa and having 365 million readers.[8][9] The name Wikipedia was coined by Larry Sanger[10] and is a portmanteau of wiki (a technology for creating collaborative websites, from the Hawaiian word wiki, meaning "quick") and encyclopedia.

Source: http://en.wikipedia.org/wiki/Wikipedia (24/01/13)

Wikipedia's departure from the expert-driven style of the encyclopaedia building mode and the large presence of “unacademic” content have been discussed in several forums, not least in Wikipedia itself (also under “Wikipedia”). Wikipedia even quotes Goethe to prove that their model of swarm intelligence works, that many semi-specialists can produce excellence: "Here as in other human endeavours it is evident that the active attention of many, when concentrated on one point, produces excellence" (The Experiment as Mediator between Subject and Object, Goethe 1772).

Although the policies of Wikipedia strongly emphasize reliability, verifiability and a neutral point of view, critics of Wikipedia accuse it of systemic bias and inconsistencies, especially undue weight given to popular culture, and allege that it favours consensus over credentials in its editorial process. This means in concrete terms: if the students or laypersons agree, they can agree against the professors' or experts' knowledge and enforce their views in the Wikipedia editing – until someone interferes.

Departing from the style of traditional encyclopaedias, Wikipedia employs an open "wiki" editing model. Except for a few particularly vandalism-prone pages, every article may be edited anonymously and with a user account. No article is owned by its creator, or any other editor, nor is it vetted by any recognized authority; rather, the articles are agreed on by consensus. This is not always easy, as the following entries prove:

- While most articles can be edited by anyone, semi-protection is sometimes necessary to prevent vandalism to popular pages.

- The reason for protection can be found in the protection log. If there are no relevant entries in the protection log, the page may have been moved after being protected.

This article and its editors are subject to Wikipedia general sanctions. See the description of the sanctions.

To-do list for Wikipedia: WikiProject Environment / Climate change task force:
The articles linked on this pages [!] can be monitored at Special:RecentChangesLinked/Wikipedia:WikiProject_Environment/Climate_change/to_do.

- Deletion discussions:
- Wikipedia:Articles for deletion/IPCC chapter 2
- Merge discussion:
- Talk:Climate change in the United Kingdom#Merge
- Discuss climate change articles for specific countries at Wikipedia:WikiProject Environment/Climate change/Climate change articles by country
- Comment on the Climate change discretionary sanctions proposal at Wikipedia:Administrators' noticeboard/Climate Change
- Articles
- Climate change - specific weasel words
- Global warming - broken digital object identifier (doi), rewrite discussion
- Temperature record of the past 1000 years - dead ext link(s)
- Cool Earth 50 - stub, orphan
- Mitigation of global warming - unsourced statements
- Climate change and agriculture - unsourced statements
- Climate change denial - neutrality disputed
- Global warming controversy - requested to be merged into Climate change denial
- Intergovernmental Panel on Climate Change - various fixes needed

...
http://en.wikipedia.org/wiki/Wikipedia:WikiProject_Environment/Climate_change_task_force (27/01/13)

This article illustrates how particularly controversial pages may be “flagged” and "monitored" by WikiProjects (cf. below). The procedure has been included here in detail since it shows how great the organizational effort is to ensure an ideal “cultured” discourse on this open platform.

4.2. Linguistic evaluation of Wikipedia

A linguistic evaluation of Wikipedia has to take into account the general editing concept and style, the diversity of styles in the main Article as well as the Talk and the Revision History that belong to them. This favours intellectual discourse, reveals differences of opinion and may help to clarify concepts. However, in user-driven non-specialist academic text production, key persons are the editors, committed lay-persons or specialists who are interested in popularizing topics they are enthusiastic about. Again, this collaborative writing effort emphasises a constructivist approach to texts. Recently, many more editorial comments have been added to categorize Wikipedia articles (according to the 2009 strategic plan), which make it clear that Wikipedia has some implicit or explicit quality standards, which readers can influence by rating articles according to four standard criteria: trustworthy, objective, complete, well-written (and well-organized), which is explained through the following link:

The Article Feedback Tool (AFT) is a Wikimedia survey for article feedback, to engage readers in the assessment of article quality, one of the five priorities defined in the strategic plan.
This tool was created with the following goals:
Quality assessment – Article feedback complements internal quality assessment of Wikipedia articles with a new source of data on quality, highlighting content that is of very high or very low quality, and measuring change over time.
Reader engagement – Article feedback encourages participation from readers, offering a call to action for some assessors to improve the article.
http://en.wikipedia.org/wiki/Wikipedia:Article_Feedback_Tool (24/01/13)

This quality management is partly based on text-linguistic (text-type) criteria (see 2.2.3 above); in any case, the recent attempts to increase its reliability make it a good starting point for linguistic research – for content and argumentation structures.

4.3. Linguistic applications of Wikipedia

From a sociolinguistic perspective, it is interesting that "editors" can create profiles with photos, lists of personal interests, contact information, and other personal information. Through this discussion feature, users’ modifications to Wikipedia entries can be followed, compared and evaluated, however, the modifications are not as dramatic as one might assume. In a linguistic project on meta-discourse (Schmied fc.), I was hoping to follow the discussion of highly controversial pages like “Global Warming” or “Climate Change” (even the choice of term here suggests some ideological stance). Since the changes to the latest report issued by the Intergovernmental Panel on Climate Change (IPCC) have been debated widely in the media, I thought that comparing hedges (such as *may* and *probably* versus *must* and *usually*) would be a fruitful exercise, as they indicate author stance or even ideological bias. This project endevour has not been very successful so far, although we found some controversial on-line discussions, indicated in a Wiki:User profile and the corresponding user_talk:

I'm actively contributing Kenya-related articles. Also, I do watch new articles of all kinds, and if needed, do edit them, mostly categorising uncategorised pages. Yes, I do add speedy and proposed deletion templates quite often, but only when necessary. That is, copyright violation, or subject which obviously fails notability guidelines. (sic!)
My username has nothing to do with my real name.
http://en.wikipedia.org/wiki/User:Julius_Sahara
(25/01/13)

This reference to the name is important here, since in some Kenyan disputes tribal affiliation would determine the stance on such issues. In fact, whereas Julius_Sahara is still very active in this internal Wiki discussions, his opponent here, Xinunus, is not so easy to follow, for:

This account may be blocked due to abusive use of one or more accounts.

It was predictable that the controversial political discourse was on some leading politician, but the relatively formal and polite style suggests a well-developed discourse culture that Wikipedia can be proud of:

Raila Odinga
I am asking you kindly to stop removing information that is well resourced. The proper way to challenge an entry is to take it up on the discussion page. If you continue to remove information that is resourced by a valid newspaper I will have to get an admin involved. Again stop removing information just because you think it doesnt belong there. You must give more information other than "its not a valid resource" when you delete

other users information. Per wiki rules that is not the correct way to edit a page on here. --Xinunus (talk) 05:24, 1 September 2008 (UTC)

Please stop removing sourced material. You continue to do so without using the discussion page [sic!]. Next time you remove something I am reporting you to an admin. Please follow Wiki rules on challenging material posted. --Xinunus (talk) 02:39, 2 September 2008 (UTC)

The information I added to Raila Odinga was from three reliable sources (a BBC article, a Voice of America article and an Africa Business news article). Judging by your other entries, and previous complaints about your changes to this page, you seem to be a Kikuyu supremacist who is bringing ethnic and tribal animosity to an information page, where it is highly inappropriate. I would be fine with deletion of the whole paragraph, but as people interested in conveying reliable information, we both should find citing Robert Mugabe's propoganda minister for reliable information about Kenya (or anything else) utterly bizarre. Perhaps we could agree on a compromise solution, where we drop the whole paragraph? —Preceding unsigned comment added by 98.209.22.245 (talk) 20:36, 8 October 2008 (UTC)

My policy is to keep this, like any other article, as neutral as possible. I have consistently removed any biased text from this page, whether they have been added pro- or anti-Odinga editors. I agree that the whole paragraph is indeed unnecessary and it has been removed from the current version. Julius Sahara (talk) 14:40, 9 October 2008 (UTC)

http://en.wikipedia.org/wiki/User_talk:Julius_Sahara (25/01/13)

Although the interruption of a heated debate through sanctions or WikiProjects may be regrettable from a linguistic point of view, it also shows that Wikipedia has been very active to ensure that bias and conflict of interest editing are reported and the respective pages are eliminated, as this entry shows:

In the context of Wikipedia, conflict of interest editing is the editing of Wikipedia articles by people whose background means that their motives are likely to conflict with the encyclopedia's neutrality policy. Conflict of interest editing includes paid editing or paid advocacy, when employees, contractors, or those with financial connection to individuals, products, corporations, organizations, political campaigns or governments edit articles related to those subjects. Although these edits may often involve minor factual corrections and changes, significant media attention has revolved around the editing of articles which removes or downplays negative information and adds or highlights positive information by editors with a conflict of interest.

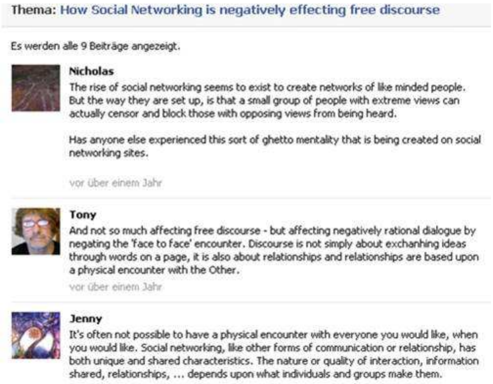
Wikipedia is free for anyone to edit, but the site maintains a neutral point of view policy. The encyclopedia's official stance on editors who have a conflict of interest strongly discourages them from working in areas where they may be intentionally or unintentionally biased. Wikipedia co-founder Jimmy Wales has argued that editors who have a clear political or financial conflict of interest should never directly edit articles, but instead propose edits to other editors on article talk pages, and seek their feedback.

http://en.wikipedia.org/wiki/Talk:Conflict_of_interest_editing_on_Wikipedia (25/01/13)

5. Facebook

5.1. Debates about privacy

Facebook is a social networking service launched in February 2004 by Mark Zuckerberg with fellow students from Harvard University. It is a very wide platform with a user profile, a news feed, messaging, voice and video calls (via Skype), and the famous LIKE button. Facebook is by far the most successful and linguistically the most stratified social network. But it is also the most criticized social network because of privacy violations, although it claims that safety of its users is top priority and requires users to give their true identity. This debate can be followed on the internet again and again (e.g. Fig. 5, which has been removed since).



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 exchanhing 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.

Fig. 5: How Social Networking is negatively effecting (!) free discourse (20/03/11)

To allow consensus about privacy, Facebook enables users to choose their own privacy. The (American) media often compare Facebook to Myspace, but one significant difference between the two websites is the level of customization. Perhaps this is the reason why Myspace (the old rival founded in 2003) has lost many followers recently in the US.

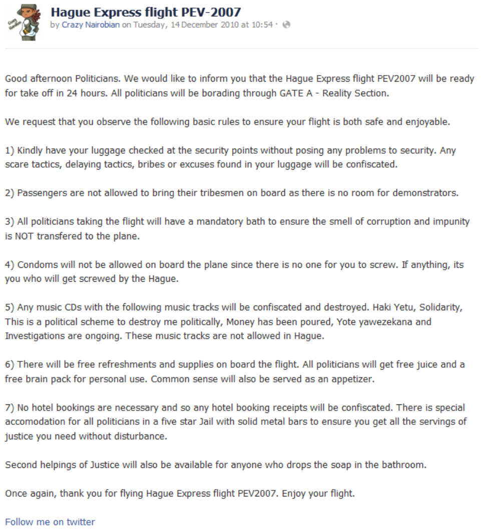
5.2. Linguistic evaluation of Facebook

Facebook combines different communication channels, different user groups, and different styles. Since its new timeline is good for collective memory, it is a good data base for in-group language change over the past few years. As the name suggests, Facebook includes some good data representing friends' talk in informal casual style, but it also provides good data for company2customers language, i.e. formal persuasive style. From a linguistic perspective, these are two very different styles.

The disadvantage for data collection is that there is too much diversity and the restricted access for users in contrast to the unlimited access for providers – the prototype of the fairy and witch mentioned at the beginning. So we have to find different ways of extracting stratified sociolinguistic language data on the basis of changing Facebook Graph APIs – and that requires specialist knowledge on Facebook technology as well as on cultural backgrounds.

Interestingly, some African cities like Yaoundé in Cameroon have "neighbourhoods" in Facebook, and it takes an insider to assess whether Bastos is more and Cite U Ngoa Ekelle and Titi Garage are less privileged, or the other way round. This we can simply compare English usage in these different networks to try a long-distance comparison of sociolinguistic informal language practices.

In most cases, however, Facebook entries show diverging styles. An extreme case is text 4: The entire discourse consisting of the author's initial narrative input and many friends commentaries clearly comprises two different parts: The first part is written in formal official English, although it is presented as an oral announcement at an airport. The oral part can be seen from the direct address "Good afternoon" politicians to the final farewell: "Enjoy your flight".



Hague Express flight PEV-2007
by Crazy Nairobian on Tuesday, 14 December 2010 at 10:54 ·

Good afternoon Politicians. We would like to inform you that the Hague Express flight PEV2007 will be ready for take off in 24 hours. All politicians will be borading through GATE A - Reality Section.

We request that you observe the following basic rules to ensure your flight is both safe and enjoyable.

1) Kindly have your luggage checked at the security points without posing any problems to security. Any scare tactics, delaying tactics, bribes or excuses found in your luggage will be confiscated.

2) Passengers are not allowed to bring their tribesmen on board as there is no room for demonstrators.

3) All politicians taking the flight will have a mandatory bath to ensure the smell of corruption and impunity is NOT transfered to the plane.

4) Condoms will not be allowed on board the plane since there is no one for you to screw. If anything, its you who will get screwed by the Hague.

5) Any music CDs with the following music tracks will be confiscated and destroyed. Haki Yetu, Solidarity, This is a political scheme to destroy me politically, Money has been poured, Yote yawezekana and Investigations are ongoing. These music tracks are not allowed in Hague.

6) There will be free refreshments and supplies on board the flight. All politicians will get free juice and a free brain pack for personal use. Common sense will also be served as an appetizer.

7) No hotel bookings are necessary and so any hotel booking receipts will be confiscated. There is special accomodation for all politicians in a five star Jail with solid metal bars to ensure you get all the servings of justice you need without disturbance.

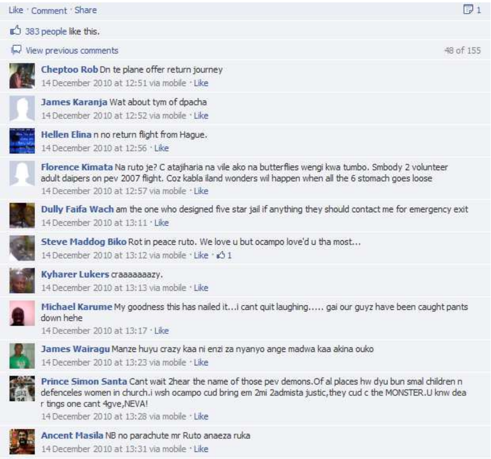
Second helpings of Justice will also be available for anyone who drops the soap in the bathroom.

Once again, thank you for flying Hague Express flight PEV2007. Enjoy your flight.

Follow me on twitter

Text 4 Hague Express flight PEV-2007
Source: http://www.facebook.com/notes/crazy-nairobian/hague-express-flight-pev-2007/494044029760 (24/01/13)

The text is almost entirely written in formal English, there are only two cases of code-mixing: *Haki yetu* (Kiswahili: "our rights") and *Yote yawezekana* (Kiswahili: "Everything is possible"), Kiswahili expressions that add local colour to the inviting discourse. The text also presupposes some understanding of Kenyan culture and political background, since the flight number PEV2007 immediately recalls for all Kenyans the "post-election violence" from the year 2007, which showed to the entire world the political contrasts in the country. In view of these events, it may be not surprising that the resentment of politicians in Kenya is still great and the "*Crazy Nairobian*", which is actually the name of a journal, would like to send them all to a five star Jail. The title "Hague Express" also assumes some general world knowledge, i.e. that war criminals are sent to The Hague in the Netherlands for trial. The political culture is also characterized by the "standard luggage" of politicians in Kenya, i.e. "scare tactics, delaying tactics, frights and excuses". The text, however, also demands some good knowledge of English, since the play on the world "screw", with the sexual meaning first and the police meaning second, cannot be taken for granted in a second-language country like Kenya.
The more oral language in the numerous spontaneous commentaries from the same day underneath illustrate not only some pronunciation characteristics of Kenyan English like "*bun*" and "*admista*" in the contribution by Prince Simon Santa but also some internet-specific language like "2hear" by the same contributor. The example of code-switching includes a few Kiswahili inclusions in the English contribution by Prince Simon Santa as opposed to a few English inclusions in the Kiswahili contribution by Florence Kimata. From the different names, the linguistic analyst can also draw two conclusions: a few names, like Steve Maddog Biko look assumed, whereas the vast majority of names appear real since they show the expected tribal forms: *Cheptoo* for Kalenjin, *Wairagu* for Bantu/Kikuyu, for instance.



Cheptoo Rob Dn te plane offer return journey
14 December 2010 at 12:51 via mobile · Like

James Karanja Wat about tym of dpacha
14 December 2010 at 12:52 via mobile · Like

Hellen Elina n no return flight from Hague.
14 December 2010 at 12:56 · Like

Florence Kimata Na ruto je? C atajiharia na vile ako na butterflies wengi kwa tumbo. Smbody 2 volunteer adult daipers on pev 2007 flight. Coz kabla iland wonders wil happen when all the 6 stomach goes loose
14 December 2010 at 12:57 via mobile · Like

Dully Faifa Wach am the one who designed five star jail if anything they should contact me for emergency exit
14 December 2010 at 13:11 · Like

Steve Maddog Biko Rot in peace ruto. We love u but ocampo love'd u tha most...
14 December 2010 at 13:12 via mobile · Like · 1

Kyharer Lukers craaaaaaazy.
14 December 2010 at 13:13 via mobile · Like

Michael Karume My goodness this has nailed it...i cant quit laughing..... gai our guyz have been caught pants down hehe
14 December 2010 at 13:17 · Like

James Wairagu Manze huyu crazy kaa ni enzi za nyanyo ange madwa kaa akina ouko
14 December 2010 at 13:23 via mobile · Like

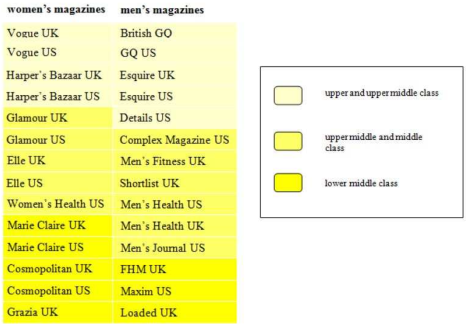
Prince Simon Santa Cant wait 2hear the name of those pev demons.Of al places hw dyu bun smal children n defenceles women in church.i wsh ocampo cud bring em 2mi 2admista justic,they cud c the MONSTER.U knw dea r tings one cant 4gve,NEVA!
14 December 2010 at 13:28 via mobile · Like

Ancent Masila NB no parachute mr Ruto anaeza ruka
14 December 2010 at 13:31 via mobile · Like

Text 5 Facebook commentaries on Hague Express flight PEV-2007
Source: http://www.facebook.com/notes/crazy-nairobian/hague-express-flight-pev-2007/494044029760

5.3. Linguistic applications of Facebook

My own Facebook example (from Beyer 2012) uses specific company webpages, status updates by British and American men's and women's magazines, to be precise. The data was collected between July and August 2011. The linguistic question was unusual for English, since English is typologically not seen as a null-subject or pro-drop language, but in informal and oral contexts subject-less clause have attracted some attention recently. The sociolinguistic research component is based on the assumption that magazine language can be categorized by social class, as Fig. 6 indicates:



women's magazines	men's magazines
Vogue UK	British GQ
Vogue US	GQ US
Harper's Bazaar UK	Esquire UK
Harper's Bazaar US	Esquire US
Glamour UK	Details US
Glamour US	Complex Magazine US
Elle UK	Men's Fitness UK
Elle US	Shortlist UK
Women's Health US	Men's Health US
Marie Claire UK	Men's Health UK
Marie Claire US	Men's Journal US
Cosmopolitan UK	FHM UK
Cosmopolitan US	Maxim US
Grazia UK	Loaded UK

Fig. 6: Classification of magazines according to social class of the readership (Beyer 2012: 29)

On this basis, Beyer (2012: 68) was able to provide evidence that

- lower middle class magazines applied fewer null subjects in their Facebook status messages than upper middle and middle class magazines,
- British and US American men's magazines used fewer null subjects in lower middle class magazines,
- lower middle class women's magazines used more null subjects than upper and upper middle class magazines, and

- the investigation of lower middle class magazines confirmed the findings of the previous group of upper middle and middle class magazines.

This is a convincing result for a small-scale study, since it is in line with the standard sociolinguistic expectations on English variation. It shows that Facebook data can be used for traditional sociolinguistic variation studies.

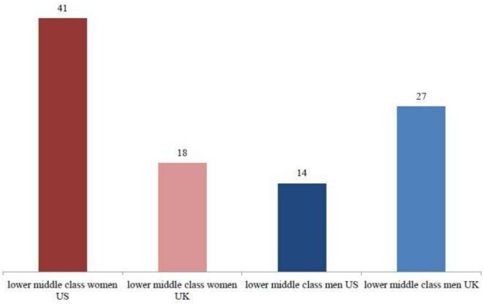


Fig. 7: Overall occurrences of null subjects per 10,000 words in lower middle class US American and British women's and men's magazines (Beyer 2012: 69)

6. Twitter

6.1. Discourse with followers by celebrities, companies and service centers

Twitter, also called "the SMS of the Internet", is a web platform that offers a social networking or 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. These tweets are publically visible by default; however, senders can restrict the delivery to their followers. Although the most popular accounts are celebrities from showbiz and politics, companies also have their own Twitter account today, especially from the US (from CNN to Amazon, but also F.C. Barcelona and Brose Baskets, the German basketball champion). The social and political impact of Twitter became famous during the "Twitter revolutions" in North Africa in 2011. Although there were some security breaches in the past, Twitter is not as controversial as facebook. Twitter launched a verification program in 2008, allowing celebrities to get their accounts verified. Twitter has expanded in 2011 to an integrated photo sharing service and in 2013 a short video attached makes it more multimodal, revealing more personal identity. Among academics, Twitter gained some reputation as a measurement of popular topics and debates, since its "trending topics" (despite some controversies about fan-group manipulations) indicate what is discussed "in the world" (like Google searches). Although this usually has a strong North American bias or has be restricted to specific areas (e.g. Germany), it reports about the current usages almost immediately and thus helps linguists interested in following the diffusion of new words and word meanings.

Fig. 8 shows a Twitter query "my brother and I" (a linguistic research question pursued in Schmidt 2012 below) in the Dar es Salaam area. Among the results was this nice text with picture of the writer and his brother. Through the geocode (of the registration or the sending location), the writer can be located on the Google map.

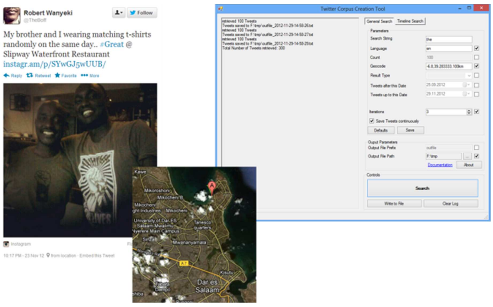
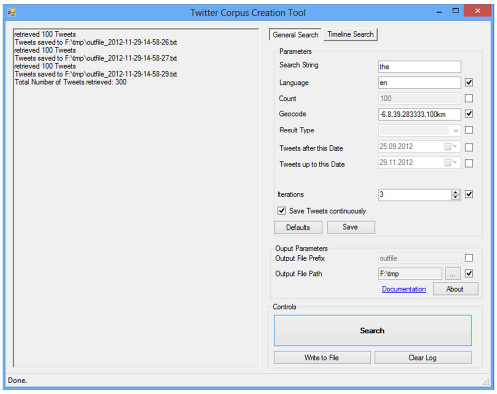

Robert Wanyeki
@TheBoff
My brother and I wearing matching t-shirts randomly on the same day.. #Great @ Slipway Waterfront Restaurant instagr.am/p/SYwGJ5wUUB/
Instagram
10:17 PM - 23 Nov 12 from location · Embed this Tweet

Fig. 8: Robert Wanyeki and his tweet from Dar es Salaam retrieved by Twitter Corpus Creation Tool

Fig. 9 is a screenshot from our Twitter Corpus Creation Tool, which uses the Twitter API. It illustrates how the Search String "may" is sent using the default language (English) and the default number of tweets to return (100, the maximum for Twitter's research API). The Geocode (here Latitude: -6.8 / Longitude: 39.283333 for Dar es Salaam and a 500 km radius) is only possible when the user has opted-in to use the Tweeting With Location feature (turned ON). The time for the Tweet collection can be set between two Dates (maximum 1 week ago) and several Iterations can be used. Results can be saved in an Output File continuously, whose name can be changed from the default "outfile" (e.g. mybrotherandI_2013-01-30-00-47-13.txt, with the search string and the time stamp), so that each collection file can be clearly identified. The link to Documentation leads to the related Manual in Wiki.


Twitter Corpus Creation Tool

```
retrieved 100 Tweets
Tweets saved to F:\tmp\outfile_2012-11-29-14-58-26.txt
retrieved 100 Tweets
Tweets saved to F:\tmp\outfile_2012-11-29-14-58-27.txt
retrieved 100 Tweets
Tweets saved to F:\tmp\outfile_2012-11-29-14-58-29.txt
Total Number of Tweets retrieved: 300
```

Parameters
Search String: the
Language: en ☑
Count: 100 □
Geocode: -6.8,39.283333,100km ☑
Result Type: □
Tweets after this Date: 25.09.2012 □
Tweets up to this Date: 29.11.2012 □
Iterations: 3 ☑
☑ Save Tweets continuously

Ouput Parameters
Output File Prefix: outfile □
Output File Path: F:\tmp ☑

Done.

Fig. 9: Twitter Corpus Creation Tool extracting "may" in tweets from Dar es Salaam, Tanzania

6.2. Linguistic evaluation of Twitter

These advantages of Twitter data are that they are relatively short, informal, written-like-spoken – and there are many. Although the content of many tweets may be considered "pointless babble" or "social grooming", the language used is very interesting for language researchers since most texts are clearly written in "conversational" style and closer to spoken English than other social media texts.

Since tweets are so frequent and often have a geo-location tag, they have been used for Twitalectology studies (e.g. Eisenstein et al. 2010 and Russ 2012), especially in the US. Fig. 9 shows an impressive result that answers a very old and well-known dialectology question on the distribution of lexical alternatives in US English:

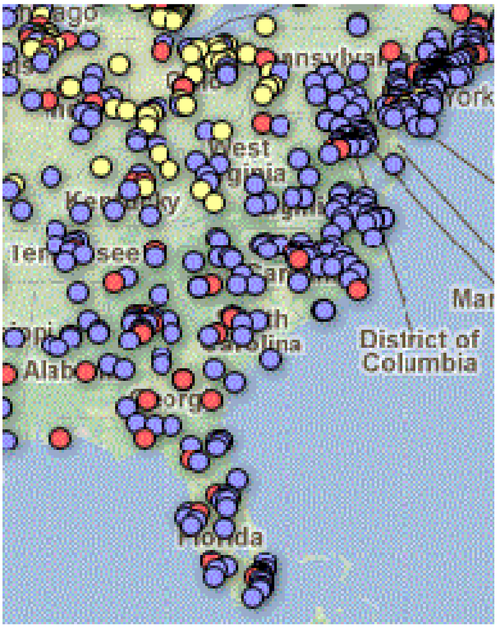


Fig. 10: Dominant lexical choice of *soda* (blue), *coke* (red) and *pop* (yellow) in US tweets
Source: http://briceruss.com/ADStalk.html

However, such API applications always depend on the restrictions imposed by the social media companies, and these can be seen as ambivalent again, balancing individual privacy demands and academic research demands is not easy – and Twitter has recently restricted the research options again.

6.3. Linguistic applications of Twitter

My own Twitter example (from Schmidt 2012) uses data on coordinated personal pronouns (like *he and I*) collected with the help of the Twitter API during one week (April 7 to 14, 2011). Although this may not sound like a long collection period, the amount of data collected was overwhelming. Due to the frequency of personal pronouns in Twitter discourse, these archives created Excel files up to 150 megabytes in size, which made them difficult to process.

Personal pronouns are a popular research topic in social discourse, e.g. Newman/Teddiman (2011) analyse them in online diary writing. It is well-known that the distribution of personal pronouns in informal social discourse is very uneven. This preference for the *me&you* perspective has been called "the personalisation of discourse" (Soffer 2012). Table 5 demonstrates clearly that writer and reader address (*I* and *you*) are by far the most frequent in our Twitter corpus:

Rank	***Frequency***	***Word***
2	256,488	*you*
3	240,532	*I*
4	140,977	*me*
10	34,825	*her*
11	33,815	*him*
17	27,739	*he*
21	26,246	*she*

Table 5: Frequency of relevant singular pronouns in the Twitter corpus (Schmidt 2012: 40)

Table 6 shows that the traditional English grammar rules (the variants in bold in the following tables) are still adhered to by Twitter users, and yet the alternatives are chosen surprisingly often. For the first time, we can gain an insight into the gradience of the phenomenon in informal English styles. Social digital discourse gives us easy access to "liquid language" (Soffer 2012) that has been very difficult to grasp before.

The variables we are used to from traditional sociolinguistic studies based on sociolinguistic interviews and corpus-linguistic analyses also apply to Twitter English usage, but whereas some usages occur very frequently, others can hardly be found even in the vast Twitter database used for this case study, as the normalised figures in tables 6 to 8 clearly show.

	you and I	*you and me*	*I and you*	*me and you*
subject coordinates	6,504 37.0%	**7,572** **43.2%**	10 0.1%	3,458 19.7%
per 1M.words	1,662	**1,935**	2	883
prepositional complements	**1,122** **26.1%**	1,550 36.0%	0 0%	**1,630** **37.9%**
per 1M.words	**287**	396	0	**417**

Table 6: 1sg. + 2sg. as subject coordinates in Twitter and 1sg. + 2sg. as prepositional complements (*for*) in Twitter (Schmidt 2012: 49/table 13 and 63/table 27)

	he/she and		*him/her and*		*I and*		*me and*	
	I	*me*	*I*	*me*	*he/she*	*him/her*	*he/she*	*him/her*
subject coordinates	**435** **31.8%**	3 0.2%	166 12.1%	15 1.1%	**0** **0%**	16 1.2%	9 0.7%	**724** **52.9%**
per 1M.words	**111**	<1	42	4	**0**	4	2	**185**
prepositional complements	4 1.5%	1 0.4%	**23** **8.7%**	22 8.4%	0 0%	0 0%	1 0.4%	**212** **80.6%**
per 1M.words	1	<1	**6**	6	0	0	<1	**54**

Table 7: 1sg. + 3sg. as subject coordinates in Twitter and 1sg. + 3sg. as prepositional complements (*for*) in Twitter (Schmidt 2012: 50/table 15 and 64/table 29)

	him and		***he and***		***her and***		***she and***	
	her	*she*	*her*	*she*	*him*	*he*	*him*	*he*
subject coordinates	**24** **41.4%**	1 1.7%	2 3.4%	**23** **39.7%**	**0** **0%**	1 1.7%	3 5.2%	4 6.9%
per 1M.words	**6**	<1	<1	**6**	**0**	<1	<1	1
prepositional complements	**21** **70.0%**	0 0%	1 3.3%	**5** **16.7%**	**1** **3.3%**	0 0%	0 0%	**2** **6.7%**
per 1M.words	**5**	0	<1	**1**	**<1**	0	0	**<1**

Table 8: 3sg. + 3sg. as subject coordinates in Twitter and as prepositional complements (*for*) in Twitter (Schmidt 2012: 52/table 17 and 56/table 21)

Again, the social media data provide convincing results – which in this study correlated with the results of an internet questionnaire survey, and this proves again that traditional sociolinguistic analyses on usage preferences can well be expanded into the new social digital discourses.

7. Conclusion: Evaluating social digital media in English Studies

I hope to have demonstrated that social digital media are a good topic in English studies. Of course, new data force us to refine our old concepts. New data also allow us to pursue our old linguistic analyses on a new basis. Maybe we can attract more media-oriented types of students to linguistic analysis when they realise that social media have something to contribute to linguistics just as linguistics has something to contribute to the new media world, and maybe we can also exploit the practical opportunities in teacher – student, student – student discourses.

The most controversial issue in academia are references to Wikipedia in academic writing. Students find it an easy starting point and professors often do not accept Wikipedia as an academic source – and both are correct: Wikipedia must be based on reliable sources, and students have to learn to go back to the original source wherever possible. The Wikipedia controversy only accentuates a problem that may occur in all publications whether in traditional printed books or on-line. This quality issue has two parts, content and presentation style; and this is raised for discussion in the Wikipedia entry in Wikipedia:

The opportunity for vandalism provides a number of unique challenges to Wikipedia. One criticism is that, at any moment, a reader of an article cannot be certain that it has not been compromised by the insertion of false information or the removal of essential information. Former *Encyclopædia Britannica* editor-in-chief Robert McHenry once described the predicament using a simile:

The user who visits Wikipedia to learn about some subject, to confirm some matter of fact, is rather in the position of a visitor to a public restroom. It may be obviously dirty, so that he knows to exercise great care, or it may seem fairly clean, so that he may be lulled into a false sense of security. What he certainly does not know is who has used the facilities before him.

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

It is clear that Wikipedia contributors usually cannot be experts in a wide field, so it is actually amazing that Wikipedia articles have been “surprisingly accurate” in many respects, even compared to traditional expert writing in encyclopaedias, as Wikipedia praises itself:

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. ···

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.".

http://en.wikipedia.org/wiki/Wikipedia (20/03/11) in section "Quality of Writing"

Thus social digital discourse is obviously a challenge and an opportunity for linguists, in teaching and in research. The research perspective seems to be particularly attractive, because social media allow us quick access to new types of language data (even from places where fieldwork may be difficult) that can help to pursue old questions of English variation. Linguistic concepts can be expanded to be profitably used to describe social media discourse. In teaching, this form of English texts brings up a discussion of old scholarly virtues like critical reading and thinking, diligence in empirical work, accuracy in documentation, etc. Thus social media may still be more a Beauty than a Beast for modern linguistics.

References

ALZOUMA, G. 2005. Myths of digital technology in Africa: Leapfrogging development? Global Media and Communication 2005 1: 339. http://gmc.sagepub.com/content/1/3/339.

ANCHIMBE, E. A. 2010. Constructing a Diaspora Anglophone Cameroonian Identity Online. Taiwo, Rotimi (ed.), p.130-144.

BARON, N. S. 2008. Always On: Language in an Online and Mobile World. Oxford: Oxford University Press.

BEYER, D. 2012. But it’s all subjective anyway, or is it? Case variation in coordinated pronouns. MA thesis. Chemnitz University of Technology.

CRYSTAL, D. 2006. Language and the Internet. Cambridge: Cambridge University Press.

CRYSTAL, D. 2011. Internet Linguistics: A Student Guide. London: Routledge.

EBNER M. – LIENHARDT, C. – ROHS, M. – MEYER, I. 2010. Microblogs in Higher Education – A chance to facilitate informal and process-oriented learning? Computers & Education 55, p.92-100.

EISENSTEIN, - O’CONNOR, B. - SMITH, N., - XING, E. 2010. A latent variable model for geographic lexical variation. Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. Cambridge, MA: 1277-1287.

FEREN□ÍK, M. 2011. Click Here (to Find Out More): E-Escape as an Indexable World. Topics in Linguistics. Discourse Analysis in a Digital World. Issue 7 – August 2011. Constantine the Philosopher University in Nitra. Faculty of Arts.

Frehmer, Carmen. (2008). Email – SMS – MMS: the Linguistic Creativity of Asynchronous Discourse in the New Media. Bern: Lang.

GIRIDHARADAS, A. 2011. India calling. An intimate portrait of a nations's remaking. New York:Times Books, Henry Holt and Company, p.74-77.

GOGGINS, S. P. – MASCARO, CH. 2011. Social Media Discourse and Culture: A Proposal for Comparative Informatics Research. Group Informatics Lab, Drexel University. Retrieved November 6, 2012 from http://www.itu.dk/people/rkva/IWCI-2011/positionPapers/ComparativeInformaticsStudyProposal-FINAL-V3.pdf

HERRING, S. C., (ed. 1996). Computer-mediated Communication: Linguistic, Social and Crosscultural Perspectives. Amsterdam: Benjamins.
MARTINEZ-PALACIOS, I. – NUNEZ–PERTEJO, P. 2012. He's absolutely massive. It's a super day. Madonna, she is a wicked singer. Youth language and intensification: a corpus-based study. In Text&Talk 32(6), p.773-796.
MYERS, G. 2010. Discourse of Blogs and Wikis. London: Continuum.
NEWMAN, J. – TEDDIMAN, L. 2011. First Person Pronouns in Online Diary Writing. Taiwo, Rotimi (ed.), p.281-295.
RAMBE, P. 2012. Critical discourse analysis of collaborative engagement in Facebook postings. Australasian Journal of Educational Technology 28(2), p.295-314.
RUSS, B. 2012. Examining Large-Scale Regional Variation Through Online Geotagged Corpora. Proceedings of the 2012 ADS Annual Meeting. Columbus, OH: 1-63.
SCHMIED, J. 2011. Using Corpora as an innovative tool to compare varieties of English around the world: the International Corpus of English. Rassegna Italiana di Linguistica Applicata - 1-2/2011, p.21-37.
SCHMIED, J. (fc). English for Academic Purposes: Contrastive Perspectives in the Curriculum. In Haase Christoph/Josef Schmied (eds.), English for Academic Purposes: Practical and Theoretical Approaches. Göttingen: Cuvillier.
SCHMIDT, S. 2012. But it's all subjective anyway, or is it? Case variation in coordinated pronouns. MA thesis. Chemnitz University of Technology.
SCHNEIDER, E. W. 2012. Editor's report 2007-2012. English World-Wide 33, p.363-368.
SOFFER, O. 2012. Liquid language? On the personalization of discourse in the digital era. New Media & Society 14(7), p.1092-1110.
TAGLIAMONTE SALI, A. – DENIS, D. 2008. Linguistic ruin? Lol! Instant messaging and teen language. American Speech 2008 83(1), p.3-34
TAIWO, R. (ed.) Handbook of Research on Discourse Behavior and Digital Communication: Language Structures and Social Interaction. IGI Global.
THURLOW, C. - MROCZEK, K. eds. 2011. Digital Discourse: Language in the New Media: Language in the New Media. Cambridge: C.U.P.