Data from the internet for language and culture studies: A corpus-linguistic appraisal[1]

Josef Schmied
Chemnitz University of Technology

Abstract:

The internet today offers new opportunities of communication and information, it creates new text-types and text collections – and it makes available new data for language and culture studies. The entire Web can be used like a corpus and provides evidence for the usage of even rare lexemes and collocations. Facebook and Twitter allow data collection in new registers and networks, while Wikipedia records the negotiation of meanings in collective writing. However, the actual authors of Facebook, Twitter and Wikipedia texts are not clear, so that sociolinguistic variation analyses may be difficult. All these opportunities and challenges deserve a new appraisal from a corpus- and sociolinguistic perspective. This paper provides a general discussion and presents findings of small research projects to illustrate how old research questions may be answered by new research data from the internet.

Keywords:

Internet, Web as corpus, social media, Facebook, Twitter, Wikipedia, pronouns, coordination, culture-specific lexeme

1. Introduction

The Internet has been a research tool in linguistics for a long time. This has been documented in conference proceedings (e.g. Tomás et al.

1 Since the time of writing the internet, its tools and the social media have develped further, but Ochieng/Dheskali (2016) provide a similar survey with a few more recent tools and references that can be consulted in addition to the proposals here.

2005 on "web-mining"), journals (e.g. Sharoff 2006 on "fishing") and even handbooks (like Taiwo 2010 on digital communication in general). The internet today offers a variety of functions that are linguistically relevant in different ways: like an encyclopaedia, a network discourse and a corpus. The internet presents current up-to-date information, for instance; one of the major examples of this is the well-known Wikipedia project, which is linguistically interesting, because it is a platform for negotiating meanings. The internet also provides a basis for new forms of communication; one of the latest examples of this is Twitter, and again it is linguistically interesting, because it allows us to follow language innovations through communities of practice, for instance. Finally, the internet can simply be used as a gigantic text collection, which facilitates the linguistic comparison of usage phenomena in different domains and genres. The top level domains can be exploited easily by retrieval restrictions in search engines (like .cn for China and .edu for, usually American, educational websites) and by special programs (like WPC below), so that national and domain-specific usages can be compared. This has often been summarized as "Web as corpus" - although real empirical corpus linguists are wary of such phrases; they know that the internet is a massive digital text collection that may be used like a corpus, but it is by no means representative of language usage, not even in many digital communities. The three internet functions mentioned demonstrate that new forms of internet discourse will always create new challenges and new opportunities for corpus linguists – and this is what I intend to explore in this contribution.

2. New text collections – new forms of data retrieval: Web as corpus

2.1. The internet from a corpuslinguistic perspective

Every student knows how to use Google to find information on the internet. The word *to google* in fact has become a new lexeme in many languages, since it allows us to find out whether words or phrases are used regularly in the web (communities). The organization of the web according to relatively restricted top level domains (like .edu,.co,. biz, or.info) even allows us to compare usage in different domains like education, commerce and news by selecting the appropriate tags in an advanced search.

Although the "Web-as-Corpus" movement (Alegria, Leturia and Sharp eds. 2009 or Evert, Kilgarriff and Sharoff eds. 2008) has gained some momentum among specialists, students of English (linguistics) usually make only the most practical use of the resources available. This

Web use has been criticized heavily, because there are a large number of pitfalls (Thewall 2005 and Kilgarriff 2007). This is why Kilgariff developed Sketch Engine "for anyone wanting to research how words behave. It is a Corpus Query System incorporating word sketches, one-page, automatic, corpus-derived summaries of a word's grammatical and collocational behaviour." (http://www.sketchengine.co.uk/, 01/04/13). Since this is essentially a commercial enterprise, students usually fall back on simple Google searches.

Of course, one can only hope that all users are aware of the limitations of such top-level domain restrictions because they may include non-English websites. So, a search whether "*informations"* - the ungrammatical plural of the English mass or non-count noun, is used in Canada (.ca) reveals thousands of hits, because the equivalent French "*informations"* with the same spelling gives us enough hits from the Francophone webpages. Furthermore, the frequencies of Google websites are obviously very rough estimates; so, the indications for relatively low occurrences may be unreliable. A comparison of Web versus standard corpora search results revealed that "only domain-related searches yield results which are compatible with those from standard corpora" (Bergh 2005: 42).

It is not only that both contain a large number of newspaper texts because they are so accessible, it is also that the web has expanded into different types of texts (especially in chat rooms, blogs and the recent digital social media) that are more personal and informal. Thus the internet has expanded its registers and offers a wider stratification now than ever before. But this is not necessarily an argument against the web as corpus, but rather against older and limited corpora.

2.2. Webcorp for KWIC results

A more sophisticated tool, WebCorp (cf. Renouf 2003), has been developed by Birmingham City University over the last 15 years and it provides not only a convenient keyword in context (KWIC) structure but also the URL that indicates where a certain word or phrase is used and allows a quick verification even in larger contexts. Again, specific searches are possible by selecting Google or other search engines like Bing, for instance, and restricting the search to specific domains (like .co.uk) or text types (like BBC News), etc. Over the last 12 years or so, WebCorp has developed from a simple interface to a sophisticated linguistic research tool. In contrast to general search engines, WebCorp contains options (esp. customisable concordance span up to 100 characters, output format, etc.) that are specifically designed for linguistic

research; it even distinguishes between British tabloid and broadsheet newspapers and academic domains in the UK and the US, for instance.

The great advantage of WebCorp output is that it displays (in addition to the URL) the key-word-in-context directly, so that the meaning of unknown words may be deduced. In our example (Figure 1), *matatus* must be a means of transport parallel to *busses*, *vans* and *cars* in Kenya.

1) http://mobile.nation.co.ke/-/1017178/1237730/-/format/xhtml/-/qs6c6q/-/index.html
Text, Wordlist, text/html, UTF8 (Content-type), date unknown

1: ... November 16, 2011 Lifestyle We should not allow matatus to take advantage of our fears By CAROLINE NJUNG'

2) http://www.mydeals.co.ke/cars-0-0-5-0-Buses,-Matatus-or-Vans-cars-for-sale-kenya.html
Text, Wordlist, text/html, ISO8859_1 (HTML source), 2011-01-01 (Copyright footer)

2: egory saloons (2937) 4 wheel drives (1055) buses, matatus or vans (313) pick ups (169) other (140) trucks &
3: (14) station wagons (11) Home > All Cars > Buses, Matatus or Vans Buses, Matatus or Vans Cars for Sale In K
4:) Home > All Cars > Buses, Matatus or Vans Buses, Matatus or Vans Cars for Sale In Kenya (313) (Page 1 of

3) http://www.cheki.co.ke/buses-matatus-vans
Text, Wordlist, text/html, UTF8 (Content-type), 2011-01-01 (Copyright footer)

5: Drives Station Wagons Pick-Ups Motorbikes Buses, Matatus & Vans Trucks Sell Your Vehicle Search Only displ
6: xNairobi KShs. 680,000/- Neg. New and Used Buses, Matatus & Vans for sale in Kenya Browse the latest new, u
7: rowse the latest new, used and second-hand buses, matatus, vans and people movers in Kenya. The most popula
8: | Website Terms & Conditions Top | Home | Buses, Matatus & Vans View Mobile Website | Unsubscribe Email |

4) http://www.the-star.co.ke/weekend/nairobi-living-/17878-irresistible-features-of-the-next-generation-of-matatus-
Text, Wordlist, text/html, UTF8 (Content-type), 2011-11-16 (Server header)

9: g Irresistible features of the next generation of matatus Skip to content Irresistible features of the next
10: t Irresistible features of the next generation of matatus Saturday, 19 March 2011 00:03 BY JAMES MURUA Rece
11: e useful disposal products. It turns out that all matatus have been ordered by our friendly and very busy e
12: dard location, with some hanging from the back of matatus beyond the reach of even the longest of arms. The
13: even the longest of arms. The next generation of matatus will probably sort this out with a standard place
14: especially the ones who refurbish the best of the matatus, they will take these trash improvements as an op
15: ith even more creature comforts. Most of the best matatus also known as 'nganya' already have some of the b
16: s or her attention and failing miserably. The new matatus will have an inbuilt bullhorn that one can grab a
17: will become a new pleasure in the new generation matatus. There shall be 'safe' and 'unsafe' seats where o

The Google Search API returned 62 hits (out of an estimated 4840000). WebCorp successfully accessed 50 web pages and generated 151 concordances.

Figure 1: WebCorp output for search term
"matatus" in.ke (truncated; 10/10/11)

2.3. WebPhraseCount for relative results

A more restricted, comparative statistics tool has been developed in our department under the name of WebPhraseCount (WPC). Like WebCorp, it uses internal Google output (API), but provides a new screen presentation that compares the absolute and relative occurrences of words and phrases on the WWW. It also allows specified searches in top level domains (cf. Schmied 2006).

The rough relative statistics can be used as plausible results, when it comes to user decisions, especially for non-native speakers. In our

case (Table 1), websites where the search item occurs have to be seen in proportion to the total websites included in Google, so that the intrasite percentages for the three options add up to 100 % in each top-level domain. If they have to decide whether *different from*, *different to* and *different than* have different co-texts or are used by different communities, they may find out that *different to* is "very British" and *different than* is more American than they had thought, as Table 1 illustrates:

Table 1: Preposition choice after *different* in British, Canadian and US top level domains (10/10/11)

	lexeme	*different from*	*different to*	*different than*
.uk 3,920M	absolute:	45,400,000	16,700,000	5,900,000
	intrasite:	66.76%	24.56%	8.68%
.ca 1,730M	absolute:	33,400,000	2,010,000	6,920,000
	intrasite:	78.9%	4.75%	16.35%
.us 564M	absolute:	21,400,000	1,110,000	5,220,000
	intrasite:	77.17%	4%	18.82%
.gov 566M	absolute:	52,300,000	483,000	14,100,000
	intrasite:	78.2%	0.72%	21.08%

Interestingly, the relatively rarely used domain.*us* has a similar distribution as the Canadian domain, whereas the extended US government web pages are more extreme.

This is to say, we have to take into account that websites are very unevenly distributed around the world when we use the web to search for variety-specific lexemes – although the "digital divide" is only relative today. We also have to bear in mind that the number of websites has increased enormously everywhere, even in Africa, as exhibited from the comparative data for our next search examples. The WPC results for the East African lexemes *juakali* ("hot sun" as in: "… most organisations are run in a 'juakali' manner where people just turn up and work"), *ugali* (a maize dish), *matatu* (pl –*s*, see KWIC above) and *askari* (pl. –*s*). The comparison of the internet statistics that the results in Table 2 and Table 3 are based on shows that

- the web has expanded enormously in Africa,
- the culture-specific proportion has decreased, so that the texts appear less exotic, and
- the absolute frequency can be higher in absolute figures in mother-tongue varieties (e.g. because exotic lexemes occur frequently in travel blogs on East Africa written by UK travellers or hosted on

UK websites, etc.), but the relative frequency shows that they are not very prominent.

Table 2: Relative frequencies of EAfr lexemes on 10/06/04

	lexeme	*juakali*	*ugali*	*matatu*	*matatus*	*askari*	*askaris*
.ke 470	absolute:	29	10	32	28	9	7
	relative:	6.17%	2.128%	6.809%	5.957%	1.915%	1.489%
.tz 333	absolute:	2	7	20	1	16	1
	relative:	0.601%	2.102%	6.006%	0.3%	4.805%	0.3%
.za 737,000	absolute:	11	29	65	28	152	68
	relative:	0.001%	0.004%	0.009%	0.004%	0.021%	0.009%
.ca 3.5M	absolute:	19	55	33	23	170	12
	relative:	0.001%	0.002%	0.001%	0.001%	0.005%	0%
.uk 6.8M	absolute:	65	139	138	84	887	107
	relative:	0.001%	0.002%	0.002%	0.001%	0.013%	0.002%

Table 3: Relative frequencies of EAfr lexemes on 10/10/11

	lexeme	*juakali*	*ugali*	*matatu*	*matatus*	*askari*	*askaris*
ke 11.1M	absolute:	3460	19400	79100	70000	4610	16200
	relative:	0.03%	0.17%	0.71%	0.63%	0.04%	0.15%
.tz 2.09M	absolute:	67	588	3320	5	4300	90
	relative:	0%	0.03%	0.16%	0%	0.21%	0%
.za 234M	absolute:	186	1820	6440	640	69400	5180
	relative:	0%	0%	0%	0%	0.03%	0%
.ca 917M	absolute:	891	3840	7100	1300	78100	654
	relative:	0%	0%	0%	0%	0.01%	0%
.uk 2,400M	absolute:	5710	9400	91700	3360	229000	6830
	relative:	0%	0%	0%	0%	0.01%	0%

The comparison of the results of the WPC query clearly shows that

- the four lexemes (from Kiswahili) are clearly East African and do not occur frequently outside (not even in South Africa),
- the compound *juakali* and the two plural forms *matatus* and *askaris* (which do not have a plural *–s* in Kiswahili) are more integrated into Kenyan English than into Tanzanian English; possibly because there are alternatives (like *daladala* for *matatu*)

or because the English *–s* plurals are still considered unusual (if there are no rounding errors in the Google data).

2.4. The internet as a translation resource

Finally, it must be mentioned at least that the web provides convenient resources for translators to choose translation options "empirically" (Ferraresi 2009). On the one hand, ad hoc corpora can be used to extract collocations in comparable text types, on the other hand tools like *Linguee.com* juxtapose bilingual websites and present not only lexical options but also syntagmatic contexts. Thus the website does not "translate" phrases for translators but can rather inspire them to recreate their own version of phrases in similar text-types.

2.5. Internet limitations

Despite all the advantages of the web-as-corpus approach generally, we have to be specifically aware of its limitations, since the web is not a corpus, but simply a massive text collection. It is not reality, not even a straightforward reflection of reality, but it is constantly changing and developing new platforms of communication, which may be an opportunity for new linguistic data-mining (Koteyko 2010), as new users create new text-types and they expand the traditional analysis of language into new forms.

3. New forms of internet communication – new sources of sociolinguistic data: Facebook and Twitter

3.1. Social media from a discourse linguistic perspective

Social networks are not a new phenomenon to empirical linguists. The Paston letters have been conceived as an ideal network to analyse language variation and change in 15th century England. One of the most dominant features of the internet is that it is public. "Private invitations" on the internet can attract masses of "friends" or "guests" and "private homepages" are intrinsically exhibitionistic and always less private than one thinks. With the rapid growth of the internet, new smaller networks have gained enormous importance, the so-called social networks. Nowadays social network space (sns) is "digital" or even "virtual". However, there are overlapping areas.

With the advent of web 2.0 in the 21st century, new opportunities, but also new challenges have arisen for linguists, political as well as text- and corpus-linguistic. The question of how reliable and how sustainable

the use of social media is caused great concern, since linguists do not want to be manipulated or become dependent on transnational or US companies (Watters 2011). Gender issues have been discussed widely (e.g. Bamman, Eisenstein and Schnoebelen 2012 or Nguyen et al. 2013). In terms of text-linguistic classification, it is questionable whether Skype conversations are already digital discourse or whether modern internet telephone is something fundamentally different. Within the wild forms of computer-mediated communication, email and forum communication have been discussed widely, even from a linguistic perspective (e.g. Herring ed. 1996 and Frehmer 2008); new classifications distinguish between communication involving two or more participants, and between mono- or multi-modal forms. This raises the question whether our basic linguistic concepts have to be modified: Is digital discourse in social networking services described appropriately on the basis of the existing modals? It may sound surprising that even Wikia, which is part of the Wikipedia family, has a definition of discourse that includes these new networks already:

> **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.wiki/Discourse (20/03/11)

Of course, Wikipedia praises itself (rightly as the example shows) to adapt to current trends and reconceptualisations (cf. 4.1 below). Other linguistic concepts are equally suited as examples, like the old definition of discourse community based on Swales (1990):

A discourse community:
has a broadly agreed set of common public goals.

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

http://en.wikipedia.org/wiki/Discourse_community (07/12/2011)

The more recent concept like "community of practice" can also be integrated in our discussions on Wikipedia:

> A **community of practice** (**CoP**) is, according to cognitive anthropologists Jean Lave and Etienne Wenger, a group of people who share an interest, 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 and 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.
>
> This type of learning practice has existed for as long as people have been learning and sharing their experiences through storytelling. Wenger coined the phrase in his 1998 book, Communities of Practice: learning, meaning and identity.
>
> http://en.wikipedia.org/wiki/Community_of_practice (07/12/2011)

Special cases of digital discourses nowadays are the extremely topical social digital discourses, where the social default is the user group (which is of course also an option in previous internet communication like email).

Despite some fuzziness we can distinguish the following concepts prototypically:

- "multi-nodal concepts (like Wikipedia) integrate (theoretically) many active "editors" and many passive "consumers" or even all ("many2many/all"),
- focussed but still less central communication channels (like Twitter) integrate relatively few active producers (usually "leaders" or "celebrities") that have many passive followers ("one2many"), and
- more peripheral channels from email to other "peer2peer" ("few2few") networks include relatively few active producers and equally few passive recipients, possibly in different circles (as in Google+).

Linguistic variables and differences between these digital discourse networks include:

- Text sizes vary dramatically: while Wikipedia has an enormous and complex network of hypertext pages, Twitter is very restricted in the length of each text and Facebook or Google+ has valuable options in between.

- The number of texts is of course in reverse order: Twitter provides sometimes many feeds per day whereas the changes of the Wikipedia community are often slower than expected and its growth is not as exponential as it used to be.
- Text types are fundamentally different (despite some overlap, which is not always wanted and can be criticized or flagged, e.g. in "written like an advertisement"): Wikipedia aims at "a neutral point of view" in informative and occasionally (especially in the TALK section) argumentative texts, and is based on the belief that "subjective" versions will be eliminated by the community over time. Twitter tends towards narrative texts in its "celebrity" discourses (in which subjective actions can be followed narrowly over time), but towards instructive texts in its "professional" discourses (as in software instructions, etc.). Facebook may be narrative, but partly also persuasive, as in company or product status up-dates (below).
- The discourse cultures are theoretically very different: whereas Twitter consists basically of specific interest groups, Wikipedia explicitly has an "opening" philosophy "unlocking knowledge to the masses". Facebook is probably able to combine both, with a focus on more in-group communication style in the personal entries and a more open one in the company entries.

The most dramatic differences can probably be found in the text functions that the various communication channels focus on:

- Wikipedia focuses on the text production and the final product which will probably never be "final", because it is hopefully adapted to new societal conceptualizations immediately.
- Twitter helps to build or maintain communities, which may be defined by common past, future, interests, etc.
- Facebook again combines features of community building and maintenance with more "permanent lifelines" based on personal interpretation and "fabrication".

Of course, there are many hybrid functions, and in particular Facebook is famous for its many channels, which range from Wikipedia-like to email-like styles, which is often seen as one of the reasons why Facebook may break apart one day.

The following sections may illustrate how linguistically relevant variables of digital social networks can be used to compile and analyse new data to answer old language questions and even to document their sociolinguistic variation.

3.2. Case study: null-subjects in Facebook status updates

A concrete research example (from Beyer 2012) extracts data from Facebook to analyze null subjects. It 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 research 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 clauses have attracted some attention recently – and for this, Facebook style is ideal, since it is informal enough to gather enough data for a variationist analysis. The data gathered are not easy to categorise in sociolinguistic variables, but categorisation can be based on the intended readership; Facebook entries by companies, esp. their status updates can be assumed to be reader-adapted, e.g. magazine language aims to correlate with the social class of its readership, as Fig. 1 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

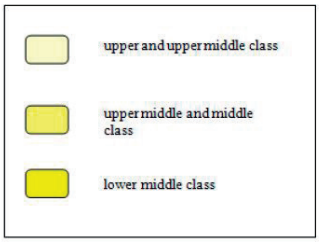

Figure 2: Classification of magazines in Facebook according to social class of the readership (Beyer 2012: 29)

On this basis, Beyer (2012: 68) was able to provide evidence (Fig. 2) for well-known sociolinguistic patters like

- 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 than in others, and

- lower middle class women's magazines used more null subjects than upper and upper middle class magazines.

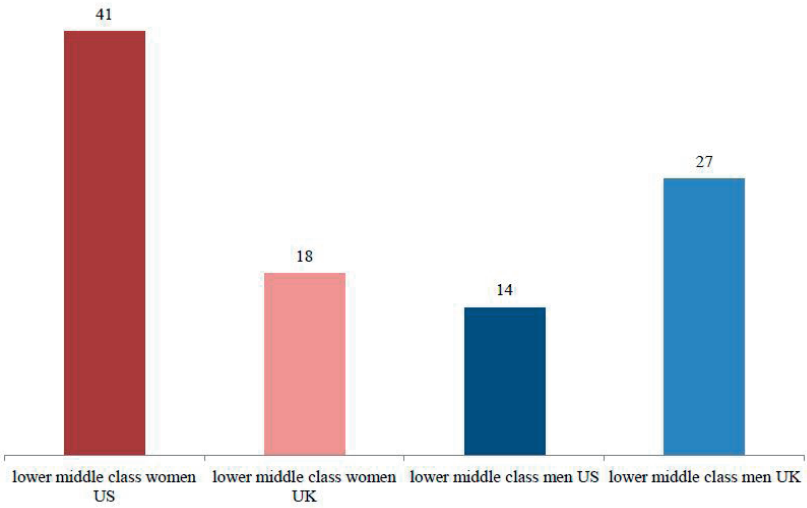

Figure 3: Overall occurrences of null subjects per 10,000 words in lower middle class US American and British women's and men's magazines (Beyer 2012: 69)

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.

3.3. Case study: pronoun coordination in Twitter

The 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 internet texts (like Wikipedia below). The register variation in a Twitter corpus has been discussed internally on Twitter web pages (Chinn n.d./2012?).

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, where they often confirm old and very well-known lexical choices (between *soda*, *pop* and *coke*, for instance). Thus, usage phenomena can be compared in different Twitter

circles all around the world (e.g. 50 km around Birmingham, Manchester and Glasgow). Whereas such regional variation is relatively well-known today, social variation is much more difficult.

A simple 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 handle.

Personal pronouns are a popular research topic in social discourse, e.g. Newman and Teddiman (2011) analyse it in online diary writing. It is well-known that the distribution of personal pronouns in informal social discourse is very uneven. This has been called "the personalisation of discourse" (Soffer 2012). Table 4 demonstrates clearly that writer and reader addresses (*I* and *you*) are by far the most frequent pronouns in Twitter:

Table 4: Frequency of relevant singular pronouns in the Twitter corpus (Schmidt 2012: 40)

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*

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 (per 100,000 words) figures in Tables 5 to 7 clearly show.

Table 5: 1sg. + 2sg. as subject coordinates in Twitter and 1sg. + 2sg. as prepositional complements in Twitter (*for*) (Schmidt 2012: 49/table 13 and 63/table 27)

	you and I	*you and me*	*I and you*	*me and you*

Table 5: 1sg. + 2sg. as subject coordinates in Twitter and 1sg. + 2sg. as prepositional complements in Twitter (*for*) (Schmidt 2012: 49/table 13 and 63/table 27)				
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**

Therefore, Table 5 shows that the traditional English grammar rules (in bold again in tables 6 and 7) 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 so far.

Table 6: 1sg. + 3sg. as subject coordinates in Twitter and 1sg. + 3sg. as prepositional complements in Twitter (Schmidt 2012: 50/Table 15 and 64/Table 29)								
	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: 3sg. + 3sg. as subject coordinates in Twitter and as prepositional complements in Twitter (Schmidt 2012: 52/Table 17 and 56/Table 21)								
	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**

Again, the social media data provide convincing results – as the Twitter based results in this study by Schmidt in 2012 correlated with the results of an internet questionnaire survey. This proves again that traditional sociolinguistic analyses on usage preferences can well be expanded into the new forms of internet communication.

3.4. New forms of identity construction in digital social media

A special challenge for sociolinguists in digital social media is ascertaining the biographical data, because everyone knows that socio-biographical data are self-references and can obviously be constructed on the internet even more easily than in reality. Users often feel that their internet activities as well as their profiles are in some way public, since even without leaks, which are reported as "mistakes" from time to time - the internet is a public space. The identities of members in social digital networks cannot be verified by personal signatures or identity cards. This is only an extreme manifestation of something that sociolinguists are aware of in all communication, i.e. that multiple identities can be constructed through language.

In every case therefore, we have to ask ourselves: The linguistic evaluation of Twitter data leads us to conclude that they are good for measuring the influence of individuals or celebrities on language change and allow us to record the innovation and diffusion of features among followers. Of course this is a relatively close discourse community which may use restricted unspecialised language. This advantage of Twitter data is that the individual texts are relatively short, informal and often "written like spoken", so they can be used or misused to illustrate the "decay" of English. Of course, the language of the social digital media is restricted by technical limitations like the tweets in Twitter, which encourages shorthand by its users and leads to abbreviations and constructions, the avoidance of complex sentences and heavy noun phrase modifications, etc.

4. New forms of knowledge construction – new writing analyses: Wikipedia

4.1. Wikipedia as collective knowledge construction

A special new type of internet text can be found in Wikipedia, not only because of its hyper-text structure (which can be seen from the underlined links in the texts below), but also its multiple author text production. This can be seen from the programmatic self-definition on the web already:

> Wikipedia … is a free, web-based, collaborative, multilingual encyclopedia project supported by the non-profit Wikimedia Foundation. Its 20 million articles (over 3.81 million in English alone) have been written collaboratively by volunteers around the world. Almost all of its articles can be edited by anyone with access to the site,[3] and it has about 100,000 regularly active contributors. As of July 2011, there are editions of Wikipedia in 282 languages. It has become the largest and most popular general reference work on the Internet, ranking sixth globally among all websites on Alexa and having an estimated 365 million readers worldwide. It is estimated that Wikipedia receives 2.7 billion monthly pageviews from the United States alone
> Wikipedia was launched in January 2001 by Jimmy Wales and Larry Sanger Sanger coined the name *Wikipedia*, which is a portmanteau of wiki (a technology for creating collaborative websites, from the Hawaiian word *wiki*, meaning "quick") and *encyclo*pedia.
> http://en.wikipedia.org/wiki/Wikipedia (07/12/11)

From a modern cognitive linguistic point of view, where meaning and identity are constructed in communities of practice, it is extremely interesting to see the Wikipedia managers' awareness of the importance of verifiability and neutral point of view, which is explicitly discussed on their webpages:

> 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), and because it favors consensus over credentials in its editorial processes, its reliability and accuracy are also targeted. Other criticisms center on its susceptibility to vandalism and the addition of spurious or unverified information; though some scholarly work suggests that vandalism is generally short-lived. A 2005 investigation in *Nature* showed that the science articles they compared came close to the level of accuracy of *Encyclopædia Britannica* and had a similar rate of "serious errors".
> Wikipedia's departure from the expert-driven style of encyclopedia building and the large presence of unacademic content has often been noted. In its 2006 Person of the Year article, *Time* magazine recognized the rapid growth of online collaboration and interaction by millions of people around the world. It cited Wikipedia as an example, in addition to YouTube, MySpace, and Facebook. Wikipedia has also been praised as a news source because of how quickly articles about recent events appear. Students have been assigned to write Wikipedia articles as an exercise in clearly and succinctly explaining difficult concepts to an uninitiated audience.
> http://en.wikipedia.org/wiki/Wikipedia (07/12/11)

A simple problem for a traditional corpus linguistic analysis is the "moving target" of Wikipedia, i.e. the texts change constantly and the author can rarely be identified. The central idea that many editors produce "excellence" is supported by a Goethe quotation:

> Here, as in other human endeavors, 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)

This model of many "semi-specialists producing excellence in popular academic writing or user-driven text production" has advantages and disadvantages. An advantage is obviously that it favours intellectual discourse and reveals differences of opinion, which hopefully clarifies concepts by making differences explicit – a central value in academic concept negotiations. This procedure fits well with our constructivist basis, since it should allow us to follow the development of terminology including author-specific hedging and differences in self-assuredness and commitment. However, this may also happen in a relatively close discourse community, even a student's seminar group may decide to publicize their perception of concepts to a world-wide readership. This illustrates the disadvantage of the approach, a "democratic" "consensus over credentials", in other words, a student group may agree to plot against their professor's expert knowledge. It may therefore not be surprising that many academics still hesitate to integrate Wikipedia into their ordinary teaching and students may feel obliged to hide the influence of Wikipedia, even if it is only used as a starting point for their own concept formation and writing.

Although the quality of writing has been a bone of contention since the beginning of Wikipedia at least a few studies have shown that in some parts it is not necessarily less reliable than traditional encyclopedia:

> 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, a history professor, stated that *American National Biography Online* outperformed Wikipedia in terms of its "clear and engaging prose", which, he said, was an important aspect of good historical writing. Contrasting Wikipedia's treatment of Abraham Lincoln to that of Civil War historian James McPherson in *American National Biography Online*, he said that both were essentially accurate and covered the major episodes in Lincoln's life, but praised "McPherson's richer contextualization… his artful use of quotations to capture Lincoln's voice… and… his ability to convey a profound message in a handful of words." By contrast, he gives an example

> of Wikipedia's prose that he finds "both verbose and dull". Rosenzweig also criticized the "waffling—encouraged by the npov policy—[which] means that it is hard to discern any overall interpretive stance in Wikipedia history." By example, he quoted the conclusion of Wikipedia's article on William Clarke Quantrill. While generally praising the article, he pointed out its "waffling" conclusion: "Some historians…remember him as an opportunistic, bloodthirsty outlaw, while others continue to view him as a daring soldier and local folk hero."
> Other critics have made similar charges that, even if Wikipedia articles are factually accurate, they are often written in a poor, almost unreadable style. Frequent Wikipedia critic Andrew Orlowski commented: "Even when a Wikipedia entry is 100 per cent factually correct, and those facts have been carefully chosen, it all too often reads as if it has been translated from one language to another then into to a third, passing an illiterate translator at each stage." A study of cancer articles by Yaacov Lawrence of the Kimmel Cancer Center at Thomas Jefferson University found that the entries were mostly accurate, but they were written at college reading level, as opposed to the ninth grade level seen in the Physician Data Query. He said that "Wikipedia's lack of readability may reflect its varied origins and haphazard editing. *The Economist* noted that the quality of writing of Wikipedia articles can be a guide to the reader: "inelegant or ranting prose usually reflects muddled thoughts and incomplete information."] 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 (07/12/2011)

In general, this may mean for an academic context simply that, as with other pieces of academic writing, the basic issue is that students have to learn to evaluate text, so that they can distinguish between good and bad Wikipedia texts as traditionally they had to distinguish between good and bad book texts. Not everything that is published even in book form today is really worth quoting by students in the discourse context they are in. This is also supported by Wiki managers' comments and writer guidance (which is actually against the original principles of a self-regulating autonomous academic community):

> This article **appears to be written like an advertisement**. Please help improve it by rewriting promotional content from a neutral point of view and removing any inappropriate external links.

For linguistic research, the innovative linguistic challenge is that Wikipedia may allow us to follow the process of knowledge construction through an analysis of persuasive elements in informative texts. This can be illustrated in a famous special case.

4.2. Case study: "Climate change" and "Global warming" in Wikipedia

As we know, public opinion about topical issues is strongly influenced by the way in which information is presented to them, especially when scientific data have to be interpreted for the general academic user. Even for academic users, it is difficult to develop their own independent stance on complex phenomena like climate change or global warming. This is why it is extremely important to analyse the metalanguage that authors add to guide the reader through their text. Author involvement, author commitment, and hedging have been identified as crucial persuasive elements (e.g. in constructions like "The data suggest" vs. "I am convinced"). In this context, it may be interesting to compare changes in Wikipedia articles (and the related discussion visible in TALK or HISTORY) in particular in the case of internationally heated debate like global warming or climate change. If we look at the Wikipedia entry for *climate change*, we can find the interesting cases of hedging at the beginning of the relatively long article and we can compare a few recent changes, esp. the ideologically important link to "global warming", which is sometimes seen as a rival concept:

> **Climate change** is a significant and lasting change in the statistical distribution of weather patterns over periods ranging from decades to millions of years. It may be a change in average weather conditions or the distribution of events around that average (e.g., more or fewer extreme weather events). Climate change may be limited to a specific region or may occur across the whole Earth. http://en.wikipedia.org/wiki/Climate_change (07/12/2011)

> **Climate change** is a significant and lasting change in the statistical distribution of weather patterns over periods ranging from decades to millions of years. It may be a change in average weather conditions, or in the distribution of weather around the average conditions (i.e., more or fewer extreme weather events). Climate change is caused by factors that include oceanic processes (such as oceanic circulation), biotic processes, variations in solar radiation received by Earth, plate tectonics and volcanic eruptions, and human-induced alterations of the natural world; these latter effects are currently causing global warming, and "climate change" is often used to describe human-specific impacts.
> http://en.wikipedia.org/wiki/Climate_change (07/08/2013)

For the discussion of Wikipedia articles, there are specific spaces that document the discussion or allow users to "Talk" about the articles on Climate Change and Global warming, which are evidence of a heated debate, which deserves further linguistic analysis:

Figure 4. Wikipedia talk page

5. Conclusion and outlook

This contribution has shown that the Internet today is such a powerful resource of knowledge construction and dissemination that applied linguists can use it for their purposes (cf. Lüdeling, Evert and Baroni 2007). It may be debateable whether there is already a sub-discipline called "web linguistics" (Bergh and Zanchetta 2008), but the opportunities and challenges are obvious, as I tried to demonstrate in this contribution.

Of course, the principles outlined in this sketch have to be adapted constantly to the latest technological developments. Thus Google recently updated its "social networking and identity service" Google+, which combines features of social media (like Messenger and Hangouts, a video chat system for up to 10 people), a cloud storage system and even a white-board for joint work on a text, for instance. Such systems are important for linguists not only for communication but also for recording communication interaction for further analysis.

These examples show that these new options of data collection will expand. The whole audio-visual dimension could not be discussed in this contribution, since it poses different challenges despite the new opportunities. Usually YouTube recordings or other internet recordings (like the Hangouts mentioned above) are not good enough for acoustic phonetic analyses, they open new opportunities for collecting web corpora (e.g. Thelwall 2005) that the early discussants of the International

Corpus of English 25 years ago would not have dreamt of (Schmied 2011). I hope that linguists interested in recent discourse and cultural changes reach out to grasp the opportunities of data from the internet, bearing in mind that all these useful innovations may also include a challenge, so that they have to be viewed critically and consciously, in order to avoid the disadvantages and use the advantages that modern internet technology offers.

References

Alegria, I., Leturia, I and Sharp, S. (eds.) (2009). Proceedings of the Fifth Web as Corpus Workshop (WAC5*). Pre-SEPLN Workshop*. Donostia-San Sebastian, Spain. Retrieved from http://www.sigwac.org.uk/attachment/wiki/WAC5/WAC5_proceedings.pdf (11/11/11).

Bamman, D., Eisenstein, J. and Schnoebelen, T. (21012). Gender in Twitter: Styles, stances, and social networks. Retrieved from http://www.cc.gatech.edu/~jeisenst/papers/GenderInTwitter923.pdf (02/11/16).

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

Bergh, G. (2005). Min(d)ing English language data on the Web. What can Google tell us? *ICAME Journal* 29, 25-46.

Bergh, G. and Zanchetta, E. (2008). Web Linguistics. In: A. Lüdeling and M. Kytö (eds.), *Corpus Linguistics 1. An International Handbook*. Berlin: Mouton de Gruyter, 309–327.

Chinn, J. (n.d.). Twitter Register Variation: Creating and Processing the Corpus. Retrieved from http://twitter.obdurodon.org/papers.html (01/04/13).

Eisenstein, J., O'Connor, B., Smith, N. and 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.

Evert, S., Kilgarriff, A. and Sharoff, S. (eds.) (2008). Proceedings of the 4th Web as Corpus Workshop (WAC-4) "Can we beat Google?" Marrakech, Morocco. Retrieved from http://webascorpus.sourceforge.net/download/WAC4_2008_Proceedings.pdf (11/11/11).

Ferraresi, A. (2009). Google and beyond: Web-As-Corpus methodologies for translators. *Tradumàtica* 07. Retrieved from http://webs2002.uab.es/tradumatica/revista/num7/articles/01/01.pdf (11/11/11).

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

Herring, S. C. (ed.) (1996). *Computer-mediated Communication: Linguistic, Social and Crosscultural Perspectives*. Amsterdam: Benjamins.

Hundt, M., Nesselhauf, N. and Biewer, C. (eds.) (2007). *Corpus Linguistics and the Web*. Amsterdam: Rodopi.

Koteyko, N. (2010). Mining the internet for linguistic and social data: An analysis of 'carbon compounds' in Web feeds. *Discourse Society* 21: 655-675.

Kilgarriff, A. (2007). "Googleology is bad science". *Computational Linguistics* 33: 147-151.

Lüdeling, A., Evert, S. and Baroni, M. (2007). Using Web data for linguistic purposes. In: M. Hundt, N. Nesselhauf and C. Biewer (eds.), 7-24.

Newman, J. and Teddiman, L. (2011). First Person Pronouns in Online Diary Writing. In: R. Taiwo (ed.), 281-295.

Nguyen, D., Gravel, R., Trieschnig, D. and Meder, T. (2013). How Old Do You Think I Am?" A Study of Language and Age in Twitter. In Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media, 2013. Retrieved from http://www.dongnguyen.nl/publications/nguyen-icwsm2013.pdf (27/08/13)

Ochieng, D. and Dheskali, J. (2016). Simple tools on the internet for English non-native academic writers in Africa. In : J. Schmied and D. Nkemleke (eds.), *Academic Writing and Research across Disciplines in Africa*. Göttingen: Cuvillier, 59-80.

Renouf, A. (2003). WebCorp: Providing a renewable data source for corpus linguists. *Language and Computers* 48: 39–58.

Russ, B. (2012). Examining large-scale regional variation through online geotagged corpora. *Proceedings of the 2012 ADS Annual Meeting*. Columbus, OH: 1-63.

Schmidt, S. (2012). But it's all subjective anyway, or is it? Case variation in coordinated pronouns. MA thesis. Chemnitz University of Technology.

Schmied, J. (2006). New ways of analysing ESL on the www with WebCorp and WebPhrasecount. In: A. Renouf and A. Kehoe (eds.), *The Changing Face of Corpus Linguistics.* Amsterdam: Rodopi, 309-324.

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, 21-37.

Schmied, J. (2012). Social digital discourse: New challenges for corpus- and sociolinguistics. *Topics in Linguistics 10: Approaches to Text and Discourse Analysis*, 43-56.

Sharoff, S. (2006). Open-source corpora: Using the net to fish for linguistic data. *International Journal of Corpus Linguistics* 11, 435-462.

Soffer, O. (2012). Liquid language? On the personalization of discourse in the digital era. *New Media & Society* 14(7), 1092-1110.

Swales, J. (1990). *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Taiwo, R. (ed.) (2010). *Handbook of Research on Discourse Behavior and Digital Communication: Language Structures and Social Interaction*. IGI Global.

Thelwall, M. (2005). Creating and using web corpora. *International Journal of Corpus Linguistics* 10: 517-541.

Tomás, J., Sánchez-Villamil, E., Lloret, J. and Casacuberta, F. (2005). WebMining: An unsupervised parallel corpora web retrieval system. In: P. Danielsson and M. Wagenmakers (eds.), *Proceedings from the Corpus Linguistics Conference 1.* Birmingham. Retrieved from http://www.corpus.bham.ac.uk/PCLC/WebMining.pdf (11/11/11)

Watters, A. (2011). How recent changes to Twitter's Terms of Service might hurt academic research. Retrieved from http://www.readwriteweb.com/archives/how_recent_changes_to_twitters_terms_of_service_mi.php (11.11.11)