Culture in Newspaper Writing? Comparing journalistic text-types in Englishes around the world[1]

Susanne Wagner and Josef Schmied
Johannes-Gutenberg-Universität Mainz and
Chemnitz University of Technology

Abstract:

The internet today provides large amounts of data from every corner of the earth. English newspapers are published in almost every country, so they can be used as a convenient comparative database, since they make available large amounts of data relatively easily. However, newspaper corpus compilers and users have to be aware of some technical as well as linguistic issues. When these are considered and overcome they provide a good window into foreign cultures and can be used for language and culture teaching. This contribution uses examples from Tanzanian newspapers to illustrate the usefulness and the challenges. To improve comparisons in future research, better ways of making available comparative databases are necessary.

Keywords:

newspaper corpus, corpus compilation, text-types, culture-specific language, collocations, compatability

1 Since the workshop this paper is based on the development of research on the internet has been fast, but we refrain from changing this contribution because it is still very valid as a simple and direct hands-on approach. We have to add however the development of the BYU Corpora, especially the NOW corpus (News on the Web), which contains 3.5 billion words of data from web-based newspapers and magazines from 2010 to yesterday, because 4 millions words are added every day!) and the GloWbE (Corpus of Global Web-based English from 20 different countries), which is explained in detail in Davies/Fuchs (2015). They are particularly useful for comparison with the personally created newspaper corpora discussed here.

1. Why newspaper corpora?

The first question we should ask ourselves is why are newspapers not only a useful but also very interesting data source? It should be stressed here that *newspaper* in all of the following will first and foremost relate to electronic versions, not hard copies. As such, newspapers provide one of the most up-to-date, easily accessible and qualitatively as well as quantitatively most diverse data sources available to the (corpus) linguist today. Long gone are the days when copies of international newspapers were only available on planes or at well-sorted news stands and usually outdated by the time we got them.

The spread of online communication and digitisation has severe consequences for all of the traditional media. The traditional newspaper format – providing news once a day, usually in the morning – no longer reflects reading habits and interests. As a consequence, large news companies such as Guardian News & Media have changed their policies. "Digital First" (see Guardian Press Release 16 July 2011) means online publication is given priority over print publication. News have a short shelf life – thus, a story that was newsworthy at 3 p.m. but had already lost its "drive" at 6 p.m. may not even make it into tomorrow's print edition. Online and offline content no longer match; in terms of quantity, the online content by far outmatches the print edition.

Another huge advantage of newspapers is that they are published world wide. Given that newspaper writing as a genre is a (relatively) narrow or well-defined area, this allows us to compare newspaper language around the world – not only between languages, but also – and this is of particular interest to sociolinguists and variationists – between varieties. Although most differences will concern relative preferences and tendencies rather than absolute differences, many of the trends observed in newspapers may be indicative of incipient change or change in progress. Journalists are generally thought to be innovative language users, but they also tend to overdo things – the effects of which are known as journalese.

Various websites by such well-renowned publications as *The Economist* or *The New York Times* strongly advise their authors against using terms, collocations, metaphors and other figures of speech which are often "used only by journalists" ("Journalese and slang", http://www.economist.com/style-guide/journalese-and-slang; see also Corbett 2012), and the Johnson Language blog stresses that this style of writing makes articles often difficult to impossible to parse not only for non-native speakers, but also for natives: "Can any native English speaker correctly

parse "Perch 'Twitter abuse' probe" on the first attempt?" (http://www.economist.com/blogs/johnson/2012/04/journalese).

The range of different regional varieties we can investigate with the help of newspapers is sheer endless – English newspapers are not only published in Great Britain, the US, Canada, Australia, New Zealand and South Africa (non-exhaustive list for L1 varieties), but also in Tanzania, Kenya, India, the Philippines, Singapore and – going even further in terms of the status of English in the respective country – China. In how far newspapers of different countries with different writing traditions etc. are really comparable is a question to be addressed in the "issues" section below.

Although "newspaper language" is generally considered a genre itself, all larger newspapers will contain articles of different sub-genres, chiefly among them leading articles/editorials, feature articles and columns/opinions, which generally reflect different levels of subjectivity on part of the author. Moreover, newspaper contents can be arranged by topic, in print often reflected in the layout of the paper (international news vs. national/regional news; financial news; arts & culture; science; sports etc.). This also allows a comparison of (sub-)genres. A third but relatively problematic starting point for a comparison is one by intended readership. Here, we do not mean the broad difference between broadsheets and tabloids, but rather the different types of broadsheets available. In this respect, a large national newspaper (like *The Times* or *The New York Times*), which is not only read in their respective places of publication, but nationally (and even internationally), has a much broader and sociologically varied readership than a (relatively) small regional newspaper like *The Aberdonian*, which will not even be available a couple hundred miles from its place of publication.

Many online newspaper archives nowadays go back at least 10 years – this offers another avenue of research: real-time comparisons. Although these archives generally will not offer enough time depth to investigate changes in some areas (e.g. morphosyntax), changes in others happen more rapidly and can readily be traced even in a decade's worth of data (e.g. preference in hedges and boosters, intensifiers; preferences in modification patterns etc.). Lexical preferences and collocational patterns are very time-dependent and also offer some interesting research possibilities, particularly for students, who are often very keen on analysing "something current".

The following list summarises the key issues raised in the preceding section; newspaper corpora offer

- data generally accessible from anywhere in the world where access to the internet is available; sheer quantity of the data (→ web-based; daily, weekly)
- access to different varieties, genres (financial vs. general news), sociolects (levels of target audience among broadsheets)
- the option to trace recent developments in speech (→ "journalese"; incoming forms; change in progress)
- the possibility of diachronic research (real time studies), at least of some quickly changing phenomena

2. Issues with newspaper corpora

Despite the lures of easy availability, vast amounts of data, and a variety of research options, newspaper corpora also present researchers at all levels with a number of challenges. Very generally, we can summarise all of these as involving choices that any researcher has to make between quality and quantity: Should the data be pretty much left "as is", resulting in a large but "dirty" corpus, or should the resulting corpus include additional (primarily sociological) metadata on authors (age, gender, ethnicity, etc. – more on that later)? If yes, then the time involved in creating a newspaper corpus will increase dramatically. In everyday life, the balance for most researchers will be somewhere between these two extremes of the continuum. The following discussion uses examples from East Africa to illustrate where recent newspaper corpora can complement established comparative tools like the International Corpus of English (ICE), whose East African part (ICE-EA) with Kenya and Tanzania was completed over 15 years ago. For a recent comparative study in a world-wide context (Schmied 2012a), ICE-EA simply proved too small and too old, so that its 1.6 Million words had to be complemented by big and recent (7.7 Million words retrieved in 2012 from Tanzania, for instance) newspaper corpora. Of course, today (on-line) newspaper corpora can be complemented by other internet data, but they are even more difficult to classify according to classical socio- and text-linguistic criteria (Schmied 2012b and Schmied this volume).

2.1 Technical issues

The creation of any newspaper corpus will start with the transfer of online data into some kind of offline format. Traditionally, this means the creation of simple text files from the original format (mostly htm or html, but other file formats are also widely used). Nowadays, a wide range of programs are available to automate that process. Httrack is one example of such a software tool: a freeware program that can "grab"

all files within a certain server directory, httrack basically copies the structure of the source site.

Although httrack has been used in corpus building before (Nelson 2010), the potential can be demonstated in detail here through a few screenshots that illustrate the processes involved explicitly. If we want to use these technologies in teaching, we have to make users aware of all the implications from selecting and down-loading different formats to saving the texts in a logical structure with corpus- and text-linguistic meanings.

Fig. 1 below is a screen shot from the download of two sections of the *Daily News* in Dar es Salaam (Tanzania). Whether the two sections, local news and columnists, are already a corpus-linguistic subsection is not only a matter of intuitive genre awareness, but also of the comparative technical set-up, so that these genres can also be found easily in other newspapers. The comparable socio- and text-linguistic meanings are usage-specific for a given cultural context. Local news are particularly welcomed by the researcher interested in local culture but they may be relatively short and more cumbersome to collect than big reportage articles. Columnists' roles differ widely in different national cultures and the differences between text-types like editorial and institutional or personal column has interesting linguistic consequences, which have been discussed in the context of the ICE-East Africa compilation in Kenya and Tanzania (cf. Schmied/Hudson-Ettle 1996).

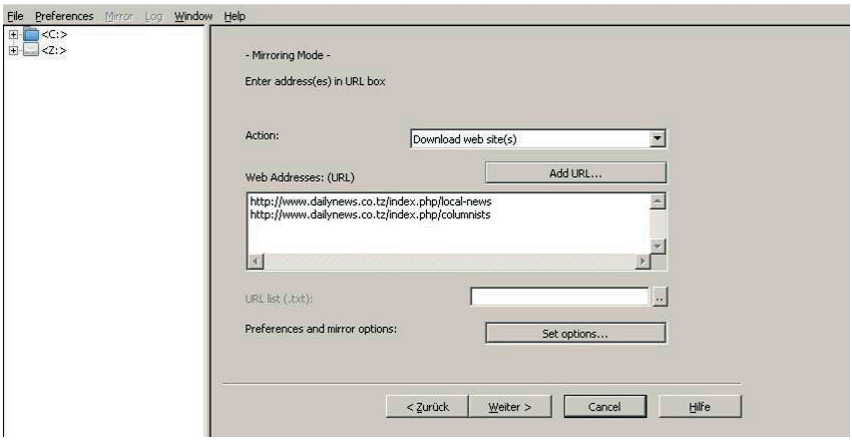

Fig. 1: Grabbing different *Daily News* sections with httrack

Fig. 2 shows the different formats that can be selected for or excluded from down-load. In our example, we exclude image formats, because we do not want a multi-media or multi-modal corpus. But

these new challenging options have been discussed intensively by corpus-linguistic and computer-science specialists (e.g. Allwood 2008 or Schmidt et al. 2009). For culture-specific concepts, they provide a particularly interesting way of complementing words with images or even sounds to get the (difficult) message across.

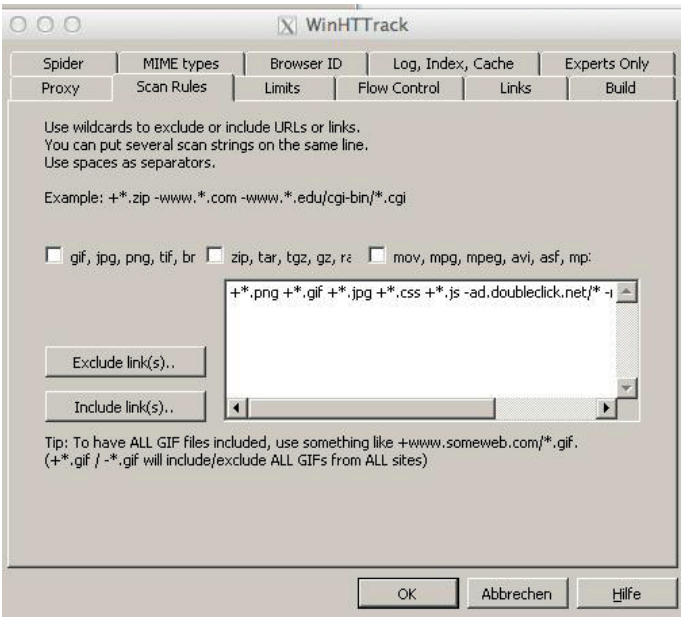

Fig. 2: Grabbing different formats with httrack

Fig. 3 demonstrates the down-load process itself, which may be relatively fast today, but for larger websites it is still useful to let httrack run over night, after checking whether it works well with a specific website.

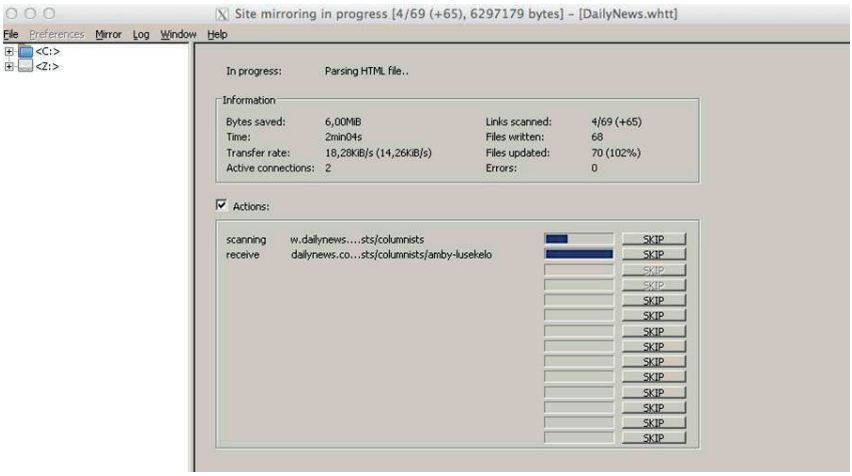

Fig. 3: Down-load process with httrack

Generally, httrack is relatively straightforward and the software is user-friendly. Still, caveats include the following:

For one, not all websites can be copied starting from their top-level domains – robots (as which httrack is classified by most websites) are disabled. This can be circumvented by entering sub-level domains to be grabbed (e.g. /sports/national rather than /sports only for the sports section), but certainly increases the researcher's work. Next, "copying" is exactly that – although httrack allows the exclusion of certain file types, the researcher has to sift through the files to separate the articles from any additional files/material. While this can be automated as well, different websites will require different versions of e.g. a Perl script that simplifies this job. Fig. 4 is an example of a relatively good structure of a newspaper website that makes the work of the researcher who wants to down-load only the relevant text sections easily, since we can see the set-up down to the individual *Daily News* articles in html format.

Fig. 4: Structure of the *Daily News* website down to individual articles

A major hassle is the fact that many newspaper websites publish the same article in different subsections, sometimes with different dates, so that duplicates have to be identified and removed before proceeding. Again, there are technically sophisticated ways to do this, but particularly at a student level, none of the issues raised in this paragraph can be expected to be handled automatically. Most students would have to rely on – very time-consuming – manual work at this stage.

Once the websites have been copied to a local hard drive, further steps need to be taken in order to be able to use the texts with text retrieval programs and concordancing software like AntConc (freeware) or WordSmith (shareware). Some programs work with htm and html files as input format, but most will require simple ASCII text files in txt format. A much more relevant issue is that practically all websites contain more information than the article the researcher is interested in. These may be headers and/or footers with copyright or contact information etc. The inclusion of this information would seriously skew the distribution of words in the resulting corpus. Once more, programs are available that convert different input file types to txt; however, the removal of headers and footers can at most be a semi-automatic process, which again must be modified for each website.

2.2 Linguistic issues

In addition to the mostly technical issues raised in the previous section, researchers have to deal with many peculiarities of newspaper articles that concern linguistic issues. These are related to various sub-fields, and one or the other may be more important in some contexts

than in others. The order of presentation here does thus not follow any logical scheme.

One issue that concerns any corpus compilation is which texts the researcher ultimately decides to include. Criteria for in- or exclusion can differ widely, and for newspapers, they may have to be modified depending on the kind of newspaper(s) one is dealing with. For example, text length might serve as a criterion of in-/exclusion, with all texts of over 2,000 words to be included in the corpus. This would not be a problem for larger national papers or even big regional papers. But in small regional newspapers, it may be difficult or even impossible to find a feature article of, say, 2,000+ words. Other criteria for in-/exclusion may involve content-related decisions in the widest sense, such as including national news but excluding international news. (This concrete example is of particular relevance as many newspapers buy their non-regional pieces from one of the big news agencies such as Reuters or AP, meaning that the same article will appear in countless newspapers around the world in an identical format, which of course makes any comparison completely useless.) Whatever the formal guidelines for text selection ultimately look like, they – like all other decision – are unlikely to be set in stone, but are rather likely to be modified from case to case (or newspaper to newspaper).

If we want to move beyond a merely genre-based decision process, a number of other features of newspaper texts also make good candidates. Most of them, however, require the researcher to deal with another notoriously problematic aspect of newspaper writing, namely the (sociological) metadata available for newspaper articles. First and foremost, this will affect everything to do with the author(s) of a text. Even a supposedly straightforward, 'simple' category such as sex turns out to be anything but that. Let's say we want to divide our corpus into articles written by men and compare them to articles written by female authors. We can be relatively certain that someone with the first name *Richard* will be a man in a Western/European context, but what about *Jati*? A quick Google search informs us that *Jati* is an Indonesian boy's name. So maybe if we want to include newspapers from Indonesia in our corpus, all *Jatis* will be put into the category 'male'. But what about an article written in a German newspaper by someone named *Jati*? Can we assume that the parents were aware of the gender association and that the author is thus also a man? No, of course we can't. The same holds for simpler cases such as shortened forms or nicknames – *Sammy* can be *Samantha* as well as *Samuel*, *Tony* can be *Anthony* and *Antonia*, etc. etc. And if we are completely unfamiliar with the naming culture of a society, we'll be relatively helpless whether we're dealing with

men or women based on their names alone. What about *Sun Xiaochen* for example? A Google image search returns pictures of both men and women. Is this a reliable way to identify authors' sex? No, of course it is not. But while some newspapers offer information about their authors on their websites, often including photos, this is by no means the rule. Instead, for many newspaper articles, it is already difficult to associate the article with any name at all. Many newspapers use author pseudonyms, often in the form of initials or other shortened forms. Others – at least in the online version – don't give any information at all. This makes it difficult to impossible to add any metadata to a newspaper text. From sex, we could move on to age, ethnicity, education etc., but all of this is equally pointless. However, this should be taken as a challenge rather than a problem because from a corpus-linguistic perspective it is highly desirable to take a forensic approach and see whether similar socio-biographical data really produce texts with similar language features in similar social and technical production contexts.

In summary, every researcher should be hyperaware of the fact that any newspaper corpus compilation will be guided not only by content-based principles, but also – and sometimes even primarily – by (ultimately) technical considerations. The more tech-savvy can relatively easily circumvent at least some of these issues. For everyone else, manual sifting through files will (still) take up a large part of the compilation process.

In summary, the following technical and content issues must be kept in mind when compiling a newspaper corpus:

- What if any tagging should the corpus receive?
- Can we readily identify genres for each text?
- Formal guidelines for the in-/exclusion of texts need to be established
- Unclear metadata (e.g. author identification and with that age, sex, ethnicity etc.) make large-scale sociolinguist studies difficult to impossible.

3. Newspaper corpus applications and analyses

Once the corpus has actually been compiled, we need to ask ourselves which questions can be addressed to it. In most corpus compilation, this question actually predates the actual compilation, with a research question guiding the corpus compilation process. However, this often means that the resulting corpus will be biased towards answering that one specific research question, making it not very useful for other (future) research.

In those cases where the corpus is compiled without any guiding research questions, typical questions that will arise during the compilation are those discussed in any introductory textbook to corpus linguistics, and the interested reader is referred to those for further information. In brief, we are talking about width vs. depth (both of the corpus design and of the research questions) and we need to cater for two different types of academic users, students and researchers interested primarily on form or in content. For sociologists, psychologists and cultural specialists in the widest sense newspaper corpora expand their methodological spectrum. For all these, the language specialist can extract the formal and functional features that can be used as indicators for cultural conventions. Of course, grammatical features are less variable in this perspective than lexical, phraseological-collocational and idiomatic features, which may indicate a more culture-specific style. Whereas grammatical and even lexical features can usually be investigated in small (1 Million-word) corpora, collocations and styles in the widest sense can only be analysed in much larger data-bases. As long as a corpus is not POS tagged, complementation studies also require large data-bases when the co-texts of the language choices have to be compared qualitatively and quantitatively. The much-quoted redundant preposition in *discuss(ed) about*, for instance, only occurs three times in over 1 000 usages. The contrast, deleted prepositions in idiomatic phrases, is more difficult to detect; but *protested* is only used a few times with *against*, but also with *at* and *about*, and most often with no preposition in front of the object (*protested the norm* with *against* seems to be clearer). This is why newspaper corpora are seen as necessary, not only quick, convenient and dirty.

This is only one reason why a hard quantitative analysis is often difficult for newspaper corpora. A second reason is that the variables of text producer and text production can usually only be filled in partially. However, qualitative research is often rewarding enough.

In summary, the following issues must be kept in mind for newspaper corpus applications:

- Can we use an existing corpus or do we have to compile a specific or even a general-purpose corpus?
- Is the corpus so big that the language features analysed occur frequently enough?
- Are the socio-biographical and textual features documented well enough to allow a quantitative multi-variant analysis?

4. Newspaper language – a window to the "cultural soul"?

Just like Rissanen (1992) demonstrated that historical corpora can be a window to the history of English, we can say that newspaper data provide a window to the "cultural soul" of its community. Topical issues may be short-lived in newspapers, but cultural values and traditions come out very clearly in newspaper corpora. The "cultural discourse" in East Africa, for instance, comes through in code-switching and loan-words. If the data-base is big enough completely unknown words can be disambiguated from the context. The (Kiswahili) phrase *asante sana* is for instance very often used in English contexts to express a cultural "thank you", although the English equivalent is still more often used, especially since newspapers are written data; but only a large-scale corpus can provide enough cases to explore differences in usage. Of course, Twitter (cf. Schmied 2012b) could provide an equally big amount of data (even in countries like Tanzania today) with a more oral style, so that the relative choices in a Twitter and a newspaper corpus could be explored.

A similar case is *pole*, the Tanzanian politeness marker, often translated as *sorry*. This *pole*[2] has to be distinguished from the English *pole*[1] by looking at each occurrence. Thus Fig. 5 shows the KWIC concordance for *pole* (in alphabetical order of right collocates) and although only about one tenths of all occurrences have this African meaning, a following *sana* (*very sorry*) disambiguates it clearly; but even names and honorific or address terms (like *mzee, ndugu*) in the context indicate this.

30 pole out of kitchen Piotr Kardys is a Pole with a problem -- a pole; in fact, a telephone pole in his kitchen. It was erected without his permission by telephone operator TPSA and when
31 refer to me by my first name. Is seems you are standing on very weak professional brick. Pole, it appears this game is alien to you. Bye Alfred Comment surely Mr Ngotezi you
32 Your comment: Total Comments on the above stories (19) Comment Pole Kaka Madaraka, Nami imenisikitisha baaday ya kusoma mail yako. Comment Madaraka,
33 wa odok paco. Pi man dong amito ni dano pe gubed ki lworo, jami ma obedo kakati I gajeti ni pole lok ameli. Lok kikome en aye lok ma en acoyo i ceng acyo (The New Vision) ma aveco kamal
34 e) image which he ostensibly seeks to salvage?? Answer: zilch, no cigar! Quite the reverse! Pole mzee!! Daily News reader, Biharamulo, Tanzania Comment Mr. Ngotezi as you had s
35 for ugandans in particular in this predicament, u see the damn floods, run, climb a tree, a pole'n what not. this is not a puzzle, just pure common sense. ....>> DAN HARRISON, WHAT ARE Y
36 ment any news coming ? Comment nice! i'm gonna make my own journal Comment Pole Ndugu Madaraka Nyerere! I revered the late Mwalimu, even in his death; my hero he was a
37 laugh at a naked other forgetting how miserably naked you yourself are!!! Wacha wa seme, pole ni mwendo Bwana!â!. TO : ERARARIT authored by on.16. June 2009 at 18:20 Who decides
38 the horizon, unseen for another half year. Throughout this time, it is visible at the other pole on the opposite side of the planet doing the same circulation along the horizon. Therefo
39 8,000 euros (19,400 pounds) to Canadian-based internet casino Goldenpalace.com. Pole wants pole out of kitchen Piotr Kardys is a Pole with a problem -- a pole; in fact, a telephone po
40 e tenets of objectivity are violated to the degree to which the story appears to favour one pole over the other. Editors worth the name always insist to their reporters on the need to
41 ls are reporting a huge surge in enrolment, supposedly inspired by model Kate Moss's sultry pole performance in the recent pop video by White Stripes and the film "Closer", in which Nata
42 muâ.â http://kampala.usembassy.gov/latest_embassy_news.html Wewe Jude abari gani? Pole pole bwana. Â Â ÂâÂ Copyright The New Vision 2000-2010. All rights reserved.
43 "Apwoyo (ma tek)." I'll be back. Abiroduogo Slowly, slowly. Mos, mos. This is equivalent to pole, pole in Kiswahili. We'll see each other [later]. Wabironenore. Let's see each other tomo
44 Acholi+Labwor=the united republic of Ugandaâ!..Â Jal Alotimbitek, sasa wewe uniambie pole pole! Please tell me. Where is &Garamba Republic& in your equation above? Is it a (-) or (+)
45 ing captain John Terry again overshadowed the match, but Ancelotti knows his side remain in pole position and clearly regards United as their serious rivals for the championship. &I thi
46 "But I'm exuberant on ELS - Summer literary seminars," I yelled against the wind, from my pole position on the dhow. ELS, that's not a new psychedelic drug on the market -- although t
47 uses to glance over her shoulder at the audience. But contrary to appearances, Debbie is no pole pro. She's a 42-year-old married banker in London who has just completed a six-week begin
48 ed?" "I had a drink too many and was hit by the door." Ojwang explains. So you give him a 'pole sana' although you quietly wonder how doors could be so accurate. Bang on the eye or th
49 then he cried out his eyes!! That is a genuine and rightful act of love ya enemy deserve!!!Pole sana Mahika3ji!!!!!! I know how it is , when someone treat you unjustly at working place.Yo
50 0, Hits; 1180 Recovering from your hangovers and other excesses other past two days, eh? Pole sana. I had a great time, watching the world go by. I even had fun watching some idiosyncr
51 ractive thread)) Comment well.. it's like I knew! Comment Ndugu, Madaraka! Pole sana. Congratulation for your wisdom, you have analysed an issue very postively, in my
52 10. On Friday February 5, 2010, 14:53 PM , Kephas Matoke, United States wrote: Mwakere pole sana ndugu.Very soon rigging will be behind us.Is a great relief to have a non performer
53 her â the lousy extremist secterian tribal dimwits in our midst&!&!&! Sasa mwendo ni pole sana sasana! Jude authored by on 7. July 2009 at 15:55 "You say Edi said,"once said that t
54 . Let's see each other tomorrow. Wanare kiny. I'm (very) sorry. Mos (shinya). Equivalent to pole (sana) in Kiswahili. Sleep well. Nindi ma ber. I want... Adwarc... You want... Odwarc...
55 mammal weighing 157 kg. If after all that you don&t go to see the National Museum, well, pole sana! The US Embassy library is a wonderfully advanced library, particularly for Science st

Fig. 5: Excerpts from the KWIC concordance for *pole* in East African newspapers

In a test corpus of over 7 million words or 14,000 files from ten East African newspapers, we can see that *matatu* is a very wide-spread, basically Kenyan, phenomenon. It occurs 214 times, while other, but related East Africanisms like *dala-dala* or *boda-boda* occur only 14

and 30 times, respectively. On such a basis, collocational analyses start becoming fruitful. Thus the nouns after *matatu* are not only *driver*, but also *owner, operator, business* and even *industry*, *system,* and *madness*; after *boda-boda* there is *cyclist*, *rider, trip*, only occasionally *driver*. Such small-scale investigations can be an entry-point to an African culture for any student of English. It is also a reliable basis for an entry in modern (on-line) dictionaries of World Englishes (cf. the model entry for *matatu* in Schmied 2004: 259). The excerpt (Fig. 6) from the KWIC concordance for *matatu* in East African newspapers illustrates three phenomena: code-switching (line 70 is completely in Kiswahili), the reuse of texts (lines 53 to 55, which may be weeded out for the definite final newspaper corpus) and some of the unusual collocations mentioned above. It also illustrates a hyper-text retrieval problem (*nasdaq* is not a collocate of *matatu* in line 71).

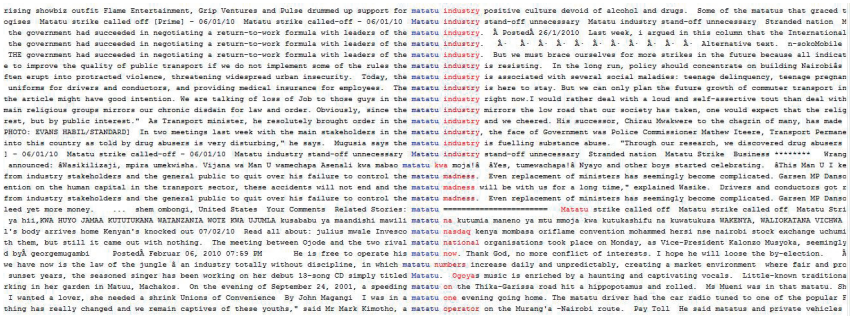
rising showbiz outfit Flame Entertainment, Grip Ventures and Pulse drummed up support for matatu industry positive culture devoid of alcohol and drugs. Some of the matatus that graced t
ogises Matatu strike called off [Prime] - 06/01/10 Matatu strike called-off - 06/01/10 Matatu industry stand-off unnecessary Matatu industry stand-off unnecessary Stranded nation M
the government had succeeded in negotiating a return-to-work formula with leaders of the matatu industry. Â Posted Â 26/1/2010 Last week, I argued in this column that the International
the government had succeeded in negotiating a return-to-work formula with leaders of the matatu industry. Â Â Â Â Â Â Â Â Â Â Â Â Â Â Â Â Alternative text. n-sokotobile
THE government had succeeded in negotiating a return-to-work formula with leaders of the matatu industry. But we must brace ourselves for more strikes in the future because all indicat
e to improve the quality of public transport if we do not implement some of the rules the matatu industry is resisting. In the long run, policy should concentrate on building Nairobiâs
ften erupt into protracted violence, threatening widespread urban insecurity. Today, the matatu industry is associated with several social maladies: teenage delinquency, teenage pregnan
uniforms for drivers and conductors, and providing medical insurance for employees. The matatu industry is here to stay. But we can only plan the future growth of commuter transport in
the article might have good intention. We are talking of loss of Job to those guys in the matatu industry right now.I would rather deal with a loud and self-assertive tout than deal with
main religious groups mirrors our chronic disdain for law and order. Obviously, since the matatu industry mirrors the low road that our society has taken, one would expect that the relig
rest, but by public interest." As Transport minister, he resolutely brought order in the matatu industry and we cheered. His successor, Chirau Mwakwere to the chagrin of many, has made
PHOTO: EVANS HABIL/STANDARD] In two meetings last week with the main stakeholders in the matatu industry, the face of Government was Police Commissioner Mathew Iteere, Transport Permane
into this country as told by drug abusers is very disturbing," he says. Mugunia says the matatu industry is fuelling substance abuse. "Through our research, we discovered drug abusers
1 - 06/01/10 Matatu strike called-off - 06/01/10 Matatu industry stand-off unnecessary Matatu industry stand-off unnecessary Stranded nation Matatu Strike Business ******** Wrang
announced: âNasikilizaji, mpira umekwisha. Vijana wa Man U wamechapa Arenali kva mabao matatu kva moja!â ÂYes, tumewachapa!â Nyayo and other boys started celebrating. âThis Man U I ke
from industry stakeholders and the general public to quit over his failure to control the matatu madness. Even replacement of ministers has seemingly become complicated. Garsen MP Danso
ention on the human capital in the transport sector, these accidents will not end and the matatu madness will be with us for a long time," explained Wasike. Drivers and conductors got r
from industry stakeholders and the general public to quit over his failure to control the matatu madness. Even replacement of ministers has seemingly become complicated. Garsen MP Danso
leed yet more money. ... shem ombongi, United States Your Comments Related Stories: matatu ========== Matatu strike called off Matatu strike called off Matatu Stri
ya his,KWA HUYO JAMAA KUTUTUKANA WATANZANIA WOTE KWA UJUMLA kusababu ya maandishi mawili matatu na kutumia maneno ya mtu mmoja kwa kutukashifu na kuwatukuza WAKENYA, WALIOKATANA VICHWA
l's body arrives home Kenyan's knocked out 07/02/10 Read all about: julius mwale Invesco matatu nasdaq kenya mombasa oriflame convention mohammed hersi nse nairobi stock exchange uchumi
th them, but still it came out with nothing. The meeting between Ojode and the two rival matatu national organisations took place on Monday, as Vice-President Kalonzo Musyoka, seemingly
d byâ georgemugambi Posted Februar 06, 2010 07:59 PM He is free to operate his matatu now. Thank God, no more conflict of interests. I hope he will loose the by-election. Â
we have now is the law of the jungle â an industry totally without discipline, in which matatu numbers increase daily and unpredictably, creating a market environment where fair and pro
sunset years, the seasoned singer has been working on her debut 13-song CD simply titled Matatu. Ogoya's music is enriched by a haunting and captivating vocals. Little-known traditiona
rking in her garden in Matuu, Machakos. On the evening of September 24, 2001, a speeding matatu on the Thika-Garissa road hit a hippopotamus and rolled. Ms Mueni was in that matatu. Sh
I wanted a lover, she needed a shrink Unions of Convenience By John Magangi I was in a matatu one evening going home. The matatu driver had the car radio tuned to one of the popular F
thing has really changed and we remain captives of these youths," said Mr Mark Kimotho, a matatu operator on the Murang'a -Nairobi route. Pay Toll He said matatus and private vehicles

Fig. 6: Excerpts from the KWIC concordance for *matatu* in East African newspapers to illustrate code-switching and collocations

It has to be pointed out that such "cultural investigations" are not restricted to Africa. The current usage and new meanings of traditionally English words in China, for instance, like *city* as in *shoes city* (which is very unusual in other Englishes) and *village* as in *urban village* (which is very different from the British concept, but so well-known that even Wikipedia has different entries) only indicate that nowadays international newspaper readers have to be careful and investigative to understand the cultural implications that are best "deciphered" in a concordance window. Of course, occasionally, "cultural language" is only meant for the "insider" and the cultural novice is grateful for cases where an explanation is provided, as in "you can't get any available job unless you bribe/pay ("toa kitu kidogo") to the senior officers". However, we have to admit that such phenomena are difficult to track in the written published language of newspapers. It is important to be aware of the limitations of our own approach.

Obviously, newspapers are only a "window" to the local culture and not the complete representation of it, especially for African newspapers in a second or even foreign language. Newspaper corpora nicely complement other approaches to compatible language data for East Africa like the available ICE corpora and the internet using other tools like Web Phrase Count (cf. Schmied 2004). But all comparisons are relative and depend on the data-base. In this context, it is a good starting point to compile quick&dirty ad-hoc corpora (Bertaccini/Aston 2001), but for many cultural purposes a large stratified reference corpus would be even better.

In summary, the following issues must be kept in mind for cultural studies based on a newspaper corpus:

- What are the technical and cultural contexts of the corpus compiled?
- Can an existing corpus be re-used?
- Is the cultural breadth and corpus stratification wide enough to generalise?

5. Learning and Teaching

When we want to use our corpus in learning and teaching, we can use a large body of relevant publications for inspiration (e.g. Sinclair ed. 2004). The special challenge of corpus-based teaching is that many exercises in inductive learning (from examples to rules) can be constructed on the basis of real-language newspaper data. This has the advantage of authenticity, but it also has disadvantages, especially of structural complexity and cultural distance. Thus in addition to other teaching and learning problems we are faced with problems in a research field that is dominated by technologically driven quantitative research: many students, particularly at undergraduate level, will not have any experience in working with real language in general, or corpora in particular. Usually, lack of time does not allow for an integration of a detailed methodology section in the course of a one-term seminar. Moreover, it would not make much sense for lecturers to teach the same tools usage to all students. Thus, many universities have compiled online tutorials or offer workshops where students can learn the basics of corpus research in a couple of hours and are then left to explore the field further on their own. Well-designed textbooks with many hands-on exercises are available which help students in their endeavour.

Experience shows that a lot hinges on students' experience with language in general and varieties in particular when they work with corpora. Having read numerous term papers and final theses, a lecturer

can typically tell very easily whether the student in question has looked at more than 10 examples and is actually able to judge whether something – a lexeme, a morphosyntactic pattern, a collocation etc. – is exceptional or not. All too often, students judge numbers too readily, and wording like "much more frequent" or "is the rule" suggest that the exposure to data has been relatively limited and the sensitivity to judge the results adequately is not fully developed yet.

In summary, the following issues must be kept in mind when newspaper corpora are used for learning and teaching:

- Are students familiar enough with data retrieval methodologies and do they have enough cultural background knowledge to be able to interpret results plausibly?
- Are students familiar with inductive learning strategies and are they explorative personalities that enjoy such approaches?
- Is the corpus manual or newspaper background material detailed enough to allow sophisticated interpretations?

6. Ensuring compatibility through infrastructure

Finally, although *ad-hoc* corpora have been mentioned as a suitable quick&dirty data-base for some investigations above, it would be desirable if electronic resources including newspaper corpora could be deposited in "digital repositories" and "digital archives" to ensure a sustained development of corpora. Although it is hoped that newspaper corpora will remain available on the internet for "grabbing" for scholarly purposes, it seems a pity to separate data-bases from analyses and "reinvent the wheel" whenever a researcher starts to use newspapers as a data-base. It would be much more effective and convenient to go to a repository where the publications and the data they are based on would be publically available for every researcher. Then researchers could decide whether to re-use deposited old data or whether to collect new project-specific data and add them later to the treasure.

Although similar ideas have been around for a long time and the Oxford Text Archive was founded as early as 1976 (for literary texts initially), a modern electronic infrastructure is being realised only now under a coordinated European project, CLARIN (the Common Language Resources and Technology Infrastructure). The German component, CLARIN-D, has this specific objective:

> Using the Open Science platform it shall be possible to download accompanying data and analysis/visualisation scripts while reading a publication without having to ask the author. Also, it should be possible to implement methods of data analysis and visualisation. An

> Open Science platform increases the transparency of research findings, facilitates reproduction of experiments and comprehension of the exact details of an experiment. It results in easier error correction and suggestion of more efficient ways of data analysis.
> The project will focus on integrating the R statistical language for a start.
> (Implementation of a platform for Open Science and Reproducible Research http://de.clarin.eu/en/curation-project-5-1-psycholinguistics; 30/05/2013)

The technical details of such data-bases have been discussed and documented in an EScience Seminar (2009):

> Repositories need to be trustworthy and therefore it is required that their repository system supports suitable security and access permission mechanisms.
>
> Repository systems need to be integrated into quality assessment procedures and therefore need to provide access statistics.
>
> Repository systems need to offer structuring options that support managing large amounts of related resources where the incarnation of a resource is its rich metadata description that adheres to a community wide accepted standard.
>
> Researchers increasingly often use individual resources in various collections for various purposes, i.e. the repository system needs to support mechanisms to build arbitrary collections of "atomic" resources and to identify such collections by metadata descriptions for citation and other purposes.
>
> Repository systems need to support the "live archives" principle, i.e. they need to cope with extensions of existing resources without changing stored objects (thus versioning) and with additional resources at various dimensions, including enrichments such as commentary and relations between resources.
>
> Each resource (version) needs to be uniquely identified by a persistent identifier and information enabling authenticity checks, so that a user – when accessing a resource with a certain identifier – is ensured that he will get the same resource.
>
> Repository systems need to support resource replications within trusted federations to support long-term preservation and access optimization. All instances share the same identity, i.e. a persistent identifier would typically indicate all existing copies.
>
> Repository systems need to support resource curation by communities, which is needed because of the continuing changes of formats and encoding principles.

> Repository systems need to have mechanisms to guarantee a high availability and persistence so that researchers can rely on the services they offer. These services need to be explicitly documented and APIs need to be available for program access.
>
> Repository systems need to have facilities to ensure format and encoding consistency based on explicit specifications.
>
> Repository systems will increasingly often be embedded in federations and therefore need to ensure interoperability at various levels to allow users for example to build virtual collections.
>
> Since repository systems often need to support long-term interpretability, resource encapsulation mechanisms for unreadable and proprietary formats need to be carefully documented and should be avoided whenever possible. Software will have a limited life-time and data survival will depend also on the maintenance costs.

This is very relevant to our newspaper corpora. It would make comparisons in future research much easier if comparative databases could be made available. In more general terms, this would not only make corpus-work more effective, it would also make it more transparent. Then researchers could continue building corpora on the basis of previous work and explore cultural variation in space and time more thoroughly. Then corpus-linguists could concentrate on the analysis and interpretation of their data instead of worrying about technicalities of corpus-compilation and data-retrieval - but still, exploring culture in newspaper writing is a challenge linguists have to take up.

References

Allwood, J. (2008). Multimodal Corpora. In: A. Lüdeling and M. Kytö (eds.) *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter. p 207-225.

Bertaccini, F. and Aston, G. (2001). Going to the Clochemerle: exploring cultural connotations through ad hoc corpora. In: G. Aston (ed.) *Learning with Corpora*. Houston TX: Athelstan, 198-219.

Corbett, P. (2012). Avoiding 'Journalese'. *After Deadline*. Blog, New York Times. http://afterdeadline.blogs.nytimes.com/2012/02/14/avoiding-journalese/Corbett, Philip. 2012. [last accessed 19.07.2012]

Davies, M. and Fuchs, R. (2015). Expanding Horizons in the Study of World Englishes with the 1.9 Billion Word Global Web-Based English Corpus (GloWbE). English World-Wide, 36, 1-28.

EScience Seminar 2009/EScience-Seminar Repository Systems http://colab.mpdl.mpg.de/mediawiki/EScience_Seminar_2009/EScience-Seminar_Repository_Systems [accessed 19.07.2012]

Guardian News & Media to be a digital-first organisation *Guardian* Press Release, 16 July 2011. http://www.guardian.co.uk/gnm-press-office/guardian-news-media-digital-first-organisation [last accessed 19.07.2012]

Implementation of a platform for Open Science and Reproducible Research for Psycholinguistics and Cognitive Psychology (WG 5). http://de.clarin.eu/en/discipline-specific-working-groups/wg-5-human-speech-processing-psycholinguistics-cognitive-psychology/curation-project-1.html [accessed 19.09.2013]

Journalese – A strange English dialect. *Johnson blog*, *The Economist,* April 3rd 2012. http://www.economist.com/blogs/johnson/2012/04/journalese [last accessed 19.07.2012]

Journalese and slang. *Economist Style guide*

http://www.economist.com/style-guide/journalese-and-slang [accessed 19.07.2012]

Nelson, G. (2010). ICELite: An Internet-sourced International Corpus of English. End of Project Report. http://ice-corpora.net/ice/icelite.htm [accessed 19.07.2012]

Rissanen, M. (1992). The diachronic corpus as a window to the history of English. In: J. Svartvik (ed.), *Directions in Corpus Linguistics: Proceedings of Nobel Symposium 82. Stockholm 4-8 August 1991.* Berlin: Mouton de Gruyter, 185-205.

Schmidt, T., Duncan, S., Ehmer, O., Hoyt, J,. Kipp, M., Magnusson, M., Rose, T. and Sloetjes, H. (2009). An Exchange Format for Multimodal Annotations. In: Michael Kipp, J-C. Martin, P. Paggio and D. Heylen (eds.), *Multimodal Corpora, Lecture Notes in Computer Science*, 207-221. Berlin: Springer.

Schmied, J. and Hudson-Ettle, D. (1996). Analysing the style of East African newspapers in English. *World Englishes* 15:1, 103-113.

Schmied, J. (2004). Cultural Discourse in the Corpus of East African English and beyond. *World Englishes 23:2*, 251-260.

Schmied, J. (2012a). Tanzanian English. In: Kortmann, Bernd (ed.), *World Atlas of Varieties of English*. Berlin/New York: de Gruyter Mouton, 454-465.

Schmied, J. (2012b). Social Digital Discourse: New Challenges for Corpus- and Sociolinguistics. *Topics in Linguistics 10: Approaches to Text and Discourse Analysis* (Constantine the Philosopher University in Nitra), 43-56.

Schmied, J. (this volume). Data from the internet for language and culture studies: A corpus-linguistic appraisal, 177-201.

Sinclair, J. (ed.) (2004). *How to Use Corpora in Language Teaching.* Amsterdam and Philadelphia: John Benjamins.