ICAME34 2013
English corpus linguistics on the move: Applications and implications
Santiago de Compostela 22-26/5/2013

ICE-EA2.0

- Between new methodologies and old compatibilities

Josef Schmied
English Language & Linguistics
Chemnitz University of Technology
http://www.tu-chemnitz.de/phil/english/ling/presentations_js.php

1. Intro: Motivation and Development

1.1 ICE-EA almost 20 years old (1st ESL/EIL corpus) diachronic corpus models popular:

- FLOB/FROUN 30 years intervals
- COCA/COHA has 200 years (US; Davis)

1.2 Beginnings of ICE: concepts

- ICE categorisation discussions
- theoretical /developmental issues: mimimum age, English-medium education

1.3 ICE-EA (1990-95) in Tanzania + Kenya

- general categorisation maintained as far as possible
- cultural adaptation wherever necessary

→ ICE-Tanzania does not include social letters

1.2 Beginnings and concepts

May 1988 ICAME Birmingham:

Schmied, Josef. "Compiling a Corpus of East African English"

discussing of Brown/LOB/Kolhapur categories?

more sociolinguistics variables (gender, status, age, 1st language, etc.)

Greenbaum: October 1988 proposal

Greenbaum, Sidney. "A proposal for an international computerised corpus of English" .*World Englishes 7*, 315.

Schmied , Josef (1990). "Corpus-linguistics and the nativization of English". *World Englishes 9*, 255-268.

"corpus-compilation paradox":

a "national" corpus should contain culture-specific text(type)s,

but we can only identify them through corpus analysis

internal discussions in ICAME Newsletters (1990-96, etc.)

Spoken text categories in ICE corpora 1991 (SEU model)

Dialogues (180)

- **Private** (100)
  - Face-to-face conversations (90)
  - Phonecalls (10)
- **Public** (80)
  - Classroom Lessons (20)
  - Broadcast Discussions (20)
  - Broadcast Interviews (10)
  - Parliamentary Debates (10)
  - Legal cross-examinations (10)
  - Business Transactions (10)

Monologues (120)

- **Unscripted** (70)
  - Spontaneous commentaries (20)
  - Unscripted Speeches (30)
  - Demonstrations (10)
  - Legal Presentations (10)
- **Scripted** (50)
  - Broadcast News (20)
  - Broadcast Talks (20)
  - Non-broadcast Talks (10)

purpur problematic 1990s

1.3 Adapting ICE-EA in the 1990s (Hudson/Schmied *Manual*)

cf. http://www.tu-chemnitz.de/phil/english/chairs/linguist/real/independent/ICE-EA/index.html

	ICE	Kenya + Tanzania	
SPOKEN	**300**	**250**	**0**
DIALOGUE	180	130	0
(written as spoken		50)	0
private	**100**	**30**	0
direct conversation	90	30	0
distanced conv.	10	--	0
public	**80**	**100**	**0**

→ adapt to usage patterns at the expense of (micro-)compatibility

WRITTEN			
press editorials	10	--	--
institutional	--	10	10
personal columns	--	10	10

→ expand and increase sub-categorisation (+/- personal)

2. Adapting ICE-EA1990-2010s

2.1. ICE-EA limitations

several categories not possible in Tanzania

→ unequal Ke-Tz proportions

→ ICE comparative work included Kenya (ESL)
and still not 1M-word corpus (normalisation!)

→ replace more categories to bring both countries and ICE-EA1 and ICE-EA2 up to 1M words?

2.2 Adapting ICE to new East African usages

- private mass media expansion, esp. in Tanzania,
  esp. IPP Media: http://ippmedia.com/
  incl. Independent Television: http://www.itv.co.tz/
  incl. *Facebook*, etc. usually in Kiswahili!
- private school expansion
  → "popular demand" English-medium
  → "declining standards"
- changing political developmental contexts
  → CCM stability in Tanzania vs. "power-sharing" in Kenya

2.3 Adapting ICE to new international communication patterns since 1989

- "global" communication
  through internet, esp. WWW, chats, blogs
  replaces snail-mail / through letters, posts??
  or additional categories?
- English as a "global" language:
  ELF/EIL= English as an International Language
  esp. European Union, China
  contradicts ICE criterion: "(secondary) education through the medium of English"

3. Expanding ICE-EA2 into 2.0

3.1. New on-line option for data-gathering: Monitor Corpora

3.2. Newspaper-Corpus using HTTrack

3.3. Twitter-Corpus using own Twitter Corpus Creation Tool (Sven Albrecht)

3.4. Issues of compatibility and representativeness

3.1 Monitor Corpora

popular like COCA, BNC since they are big enough for lexical usage studies, idiomaticity, etc.

when they can be collected quickly, they are a necessary quick-and-dirty expansion to the small-and-beautiful ICE

→ internet corpora:
multi-medial? (less for East Africa)

→ well-established newspaper monitor corpora

→ less well-established Facebook/Twitter corpora

3.2 Newspaper-Corpus using HTTrack

Standard tool (cf. Nelson 2009, 2011)

but success depends very much on the on-line Web structure,

e.g. good: *Daily News* in Dar es Salaam

e.g. bad: *Daily Nation* in Nairobi

despite its claim

The *Daily Nation* is Kenya's leading newspaper and is a product of Nation Media Group (NMG) Limited.

NMG, founded by His Highness the Aga Khan in 1959, has become the largest independent media house in East and Central Africa. It has been quoted on the Nairobi Stock Exchange since the early 1970s.

As the leading multi-media house in the East African region, it has print as well as electronic media and the digital platforms which attracts a regular readership quite unparalleled in the region.

About us (http://www.nation.co.ke/meta/-/1194/1172/-/ojmv8c/-/index.html)

HTTrack WEBSITE COPIER

Free software offline browser

Version 3.46-1 (06/23/2012)

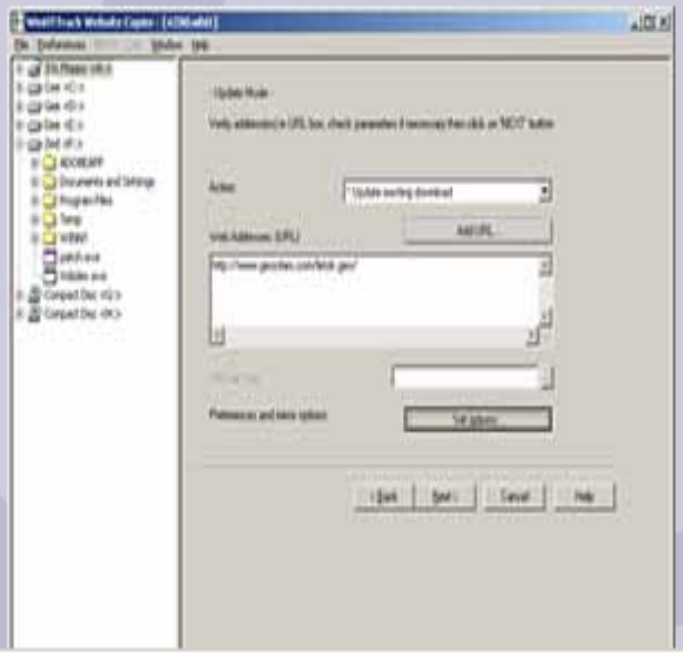
Unicode filenames handling, and many engine fixes

Installing HTTrack: Go to the download section now!

For help and questions: Visit the forum, Read the documentation, Read the FAQs

---

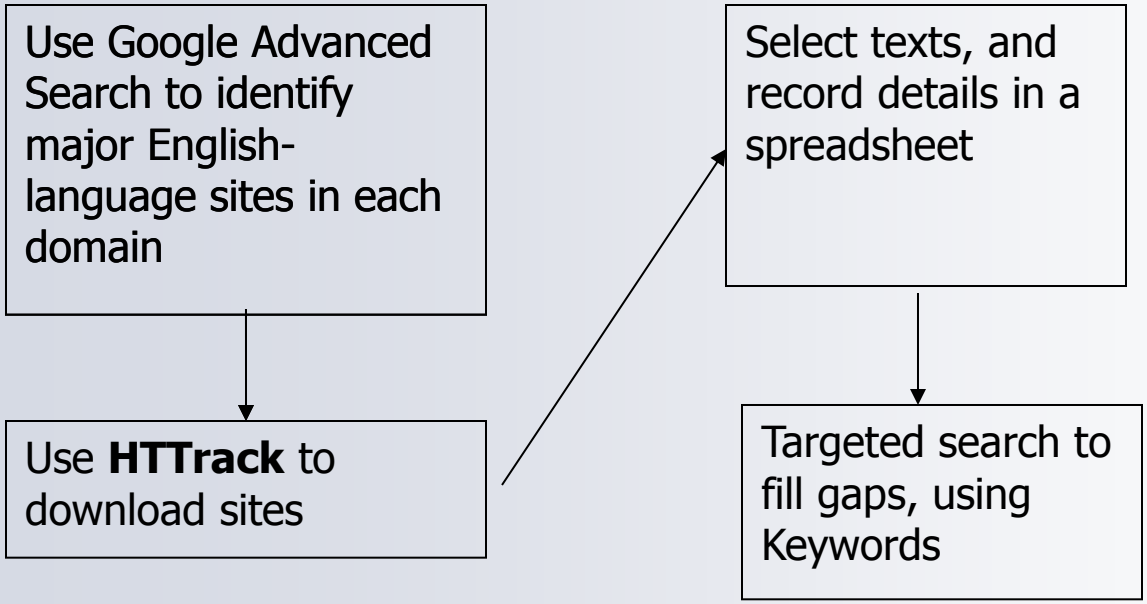
Welcome



HTTrack is a free (GPL, libre/free software) and easy-to-use offline browser utility.

It allows you to download a World Wide Web site from the Internet to a local directory, building recursively all directories, getting HTML, images, and other files from the server to your computer. HTTrack arranges the original site's relative link-structure. Simply

3.2 Workflow for Monitor Corpus



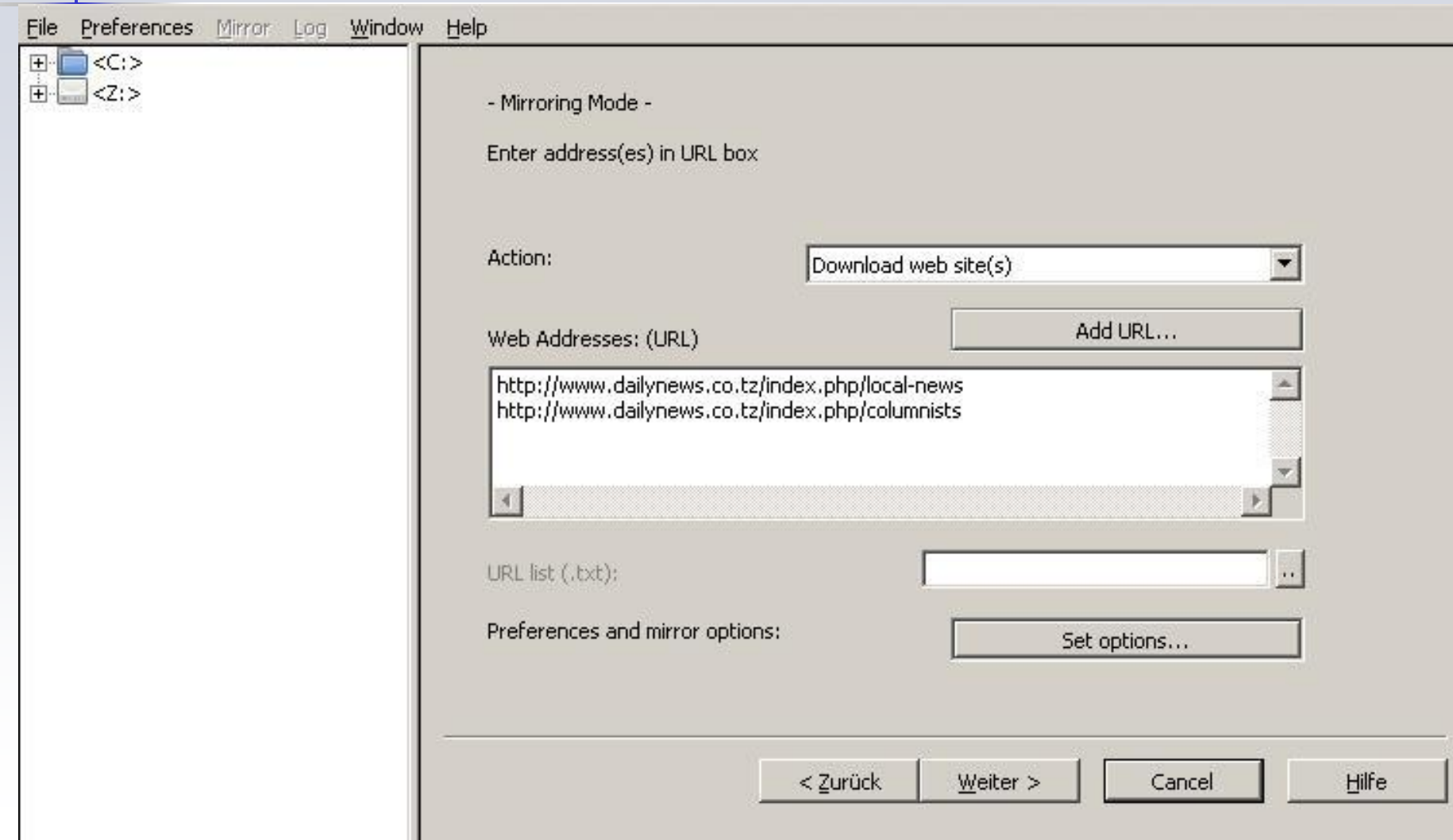
courtesy Nelson, Gerry 2009

http://ice-corpora.net/ice/icelite.htm

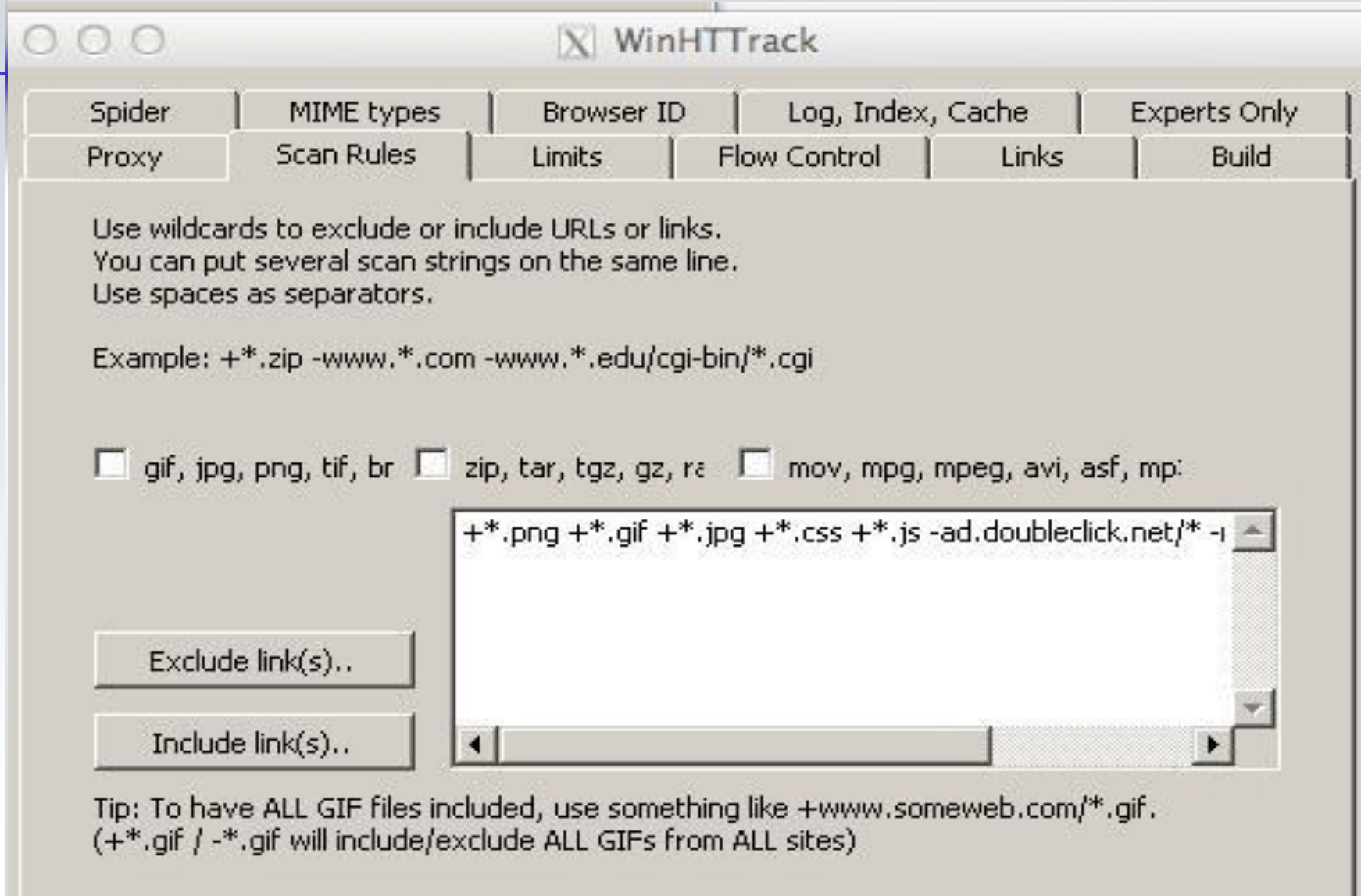
3.2 Advantages of HTTrack for ICElite

- **customisable**, to exclude unwanted files, e.g. images, sounds, movies, .exe.
  Customised settings can be saved in an "options" file [icelite.opt]
- **fast:** can download entire websites in a relatively short time (depending on the size of the site)
- **stable**: it never crashed, even when the download was aborted.
- can be run ‘in the background’, and won’t interfere with other processes
- can be run overnight, and will safely switch off your PC
- inserts **time & date accessed** in each downloaded file

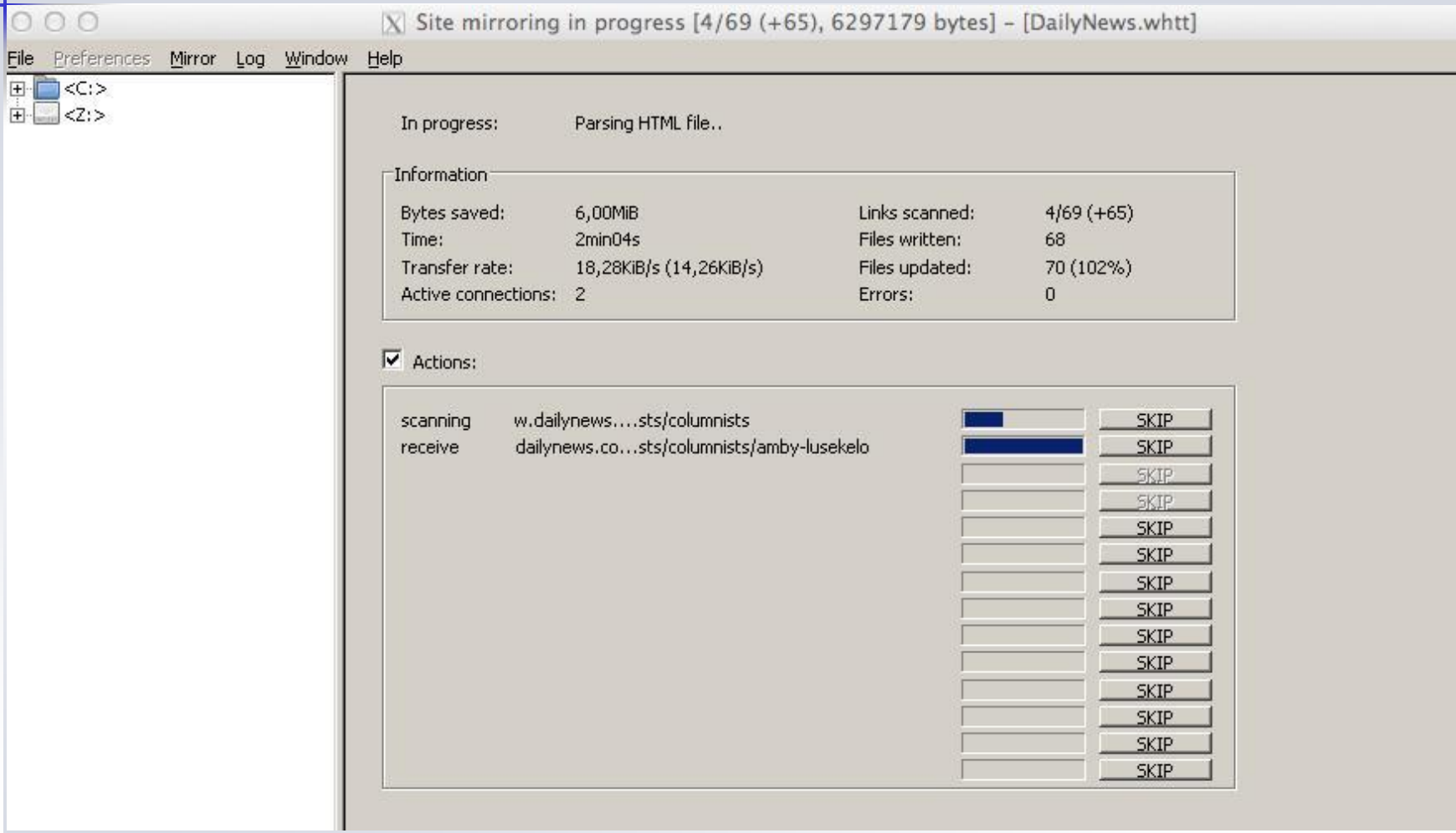
HTTrack webinput: *Daily News* (Dar es Salaam)



HTTrack file type selection



HTTrack in action
Daily News (Dar es Salaam)



HTTrack good structure → good output
Daily News (Dar es Salaam)

3.3 Twitter Corpus Creation Tool (TCCT)

designed to grab data from Twitter and store it locally

features:

- grabbing and saving tweets
- filtering tweets by language, date, location and type (recent, popular)
- interactive and non-interactive use (allows running the tool from a script)

capabilities:

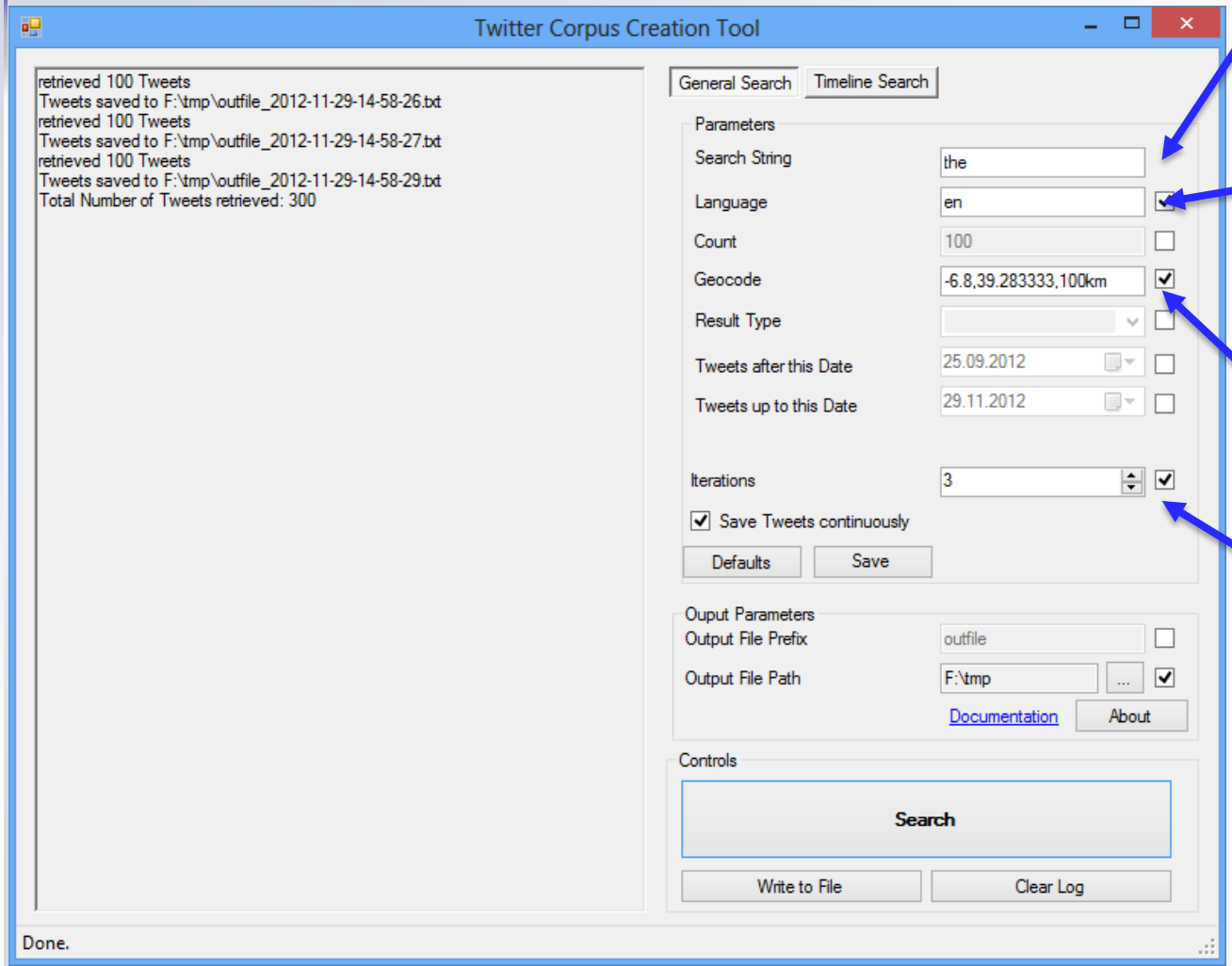
- maximum of 100 tweets per run → approx. 1500 words
- maximum of 100 consecutive runs

TCCT: Commandline parameters

https://twiki.tu-chemnitz.de/bin/view/English/TwitterCorpusCreationTool

- searchstring - **required**
- lang - Language, as defined by ISO-639-1
- count the number of tweets to return.
  a Twitter API internal option
  results of the Twitter search function are returned as pages that have a certain number of results on it
  *valid values range from 1 to 100 d*efault is 100
- geocode - This option allows the user to filter the results by geolocation. It uses the coordinates in the format of latitude and longitude and requires a radius that has to be either in miles or kilometers. Format example: **38.422,27.129,50km (latitude,longitude,radius)**
- type - The type of the returned tweets. Possibilities are *recent* (only returns in reverse chronological order), *popular* (returns tweets that Twitter ranks popular, nontransparent) and *mixed* (returns a mix of recent and popular Tweets).
- since + until
- prefix - The prefix of the output file. Output files are named *prefix-timestamp.txt*.
- path - Output path.
- iterations - The number of times to re-run the search. Convenient method to get more Tweets per run.

TCCT: interface command options



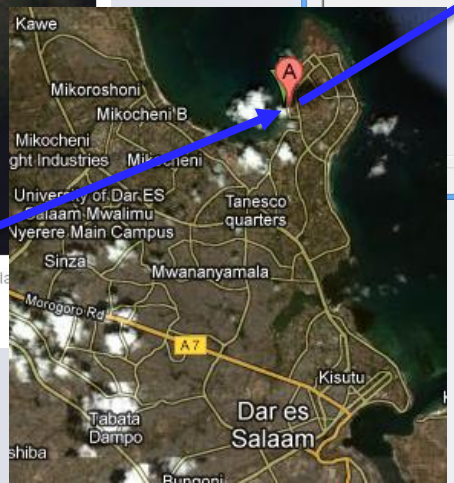
search string is always required

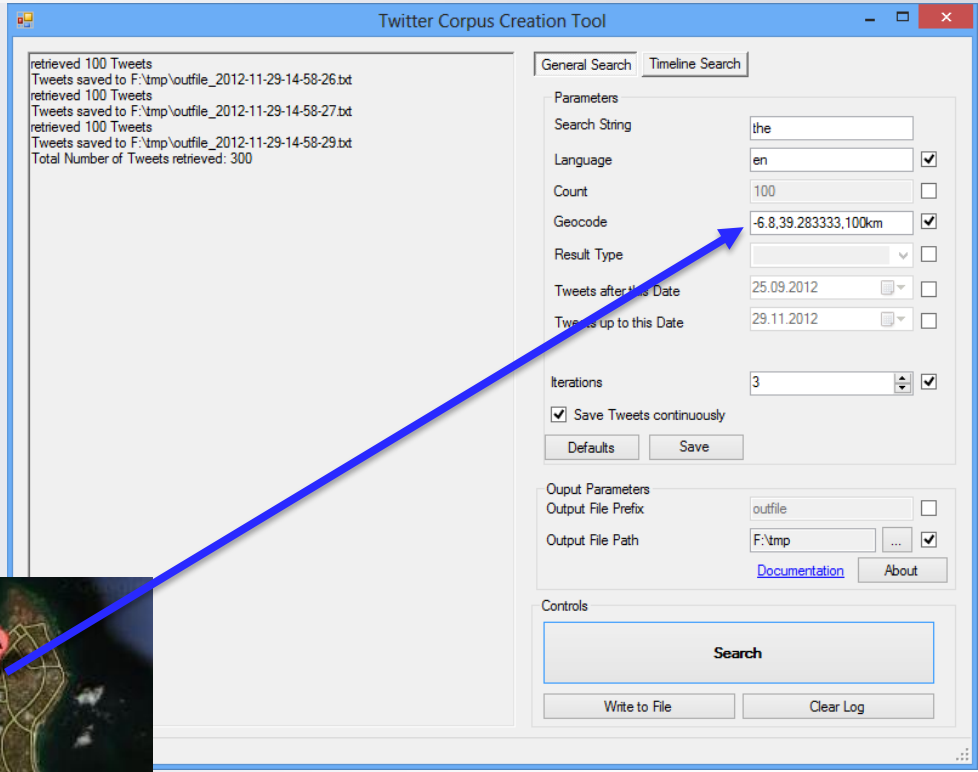
language, as definded by ISO-639-1

geo location latitude,longitude, radius

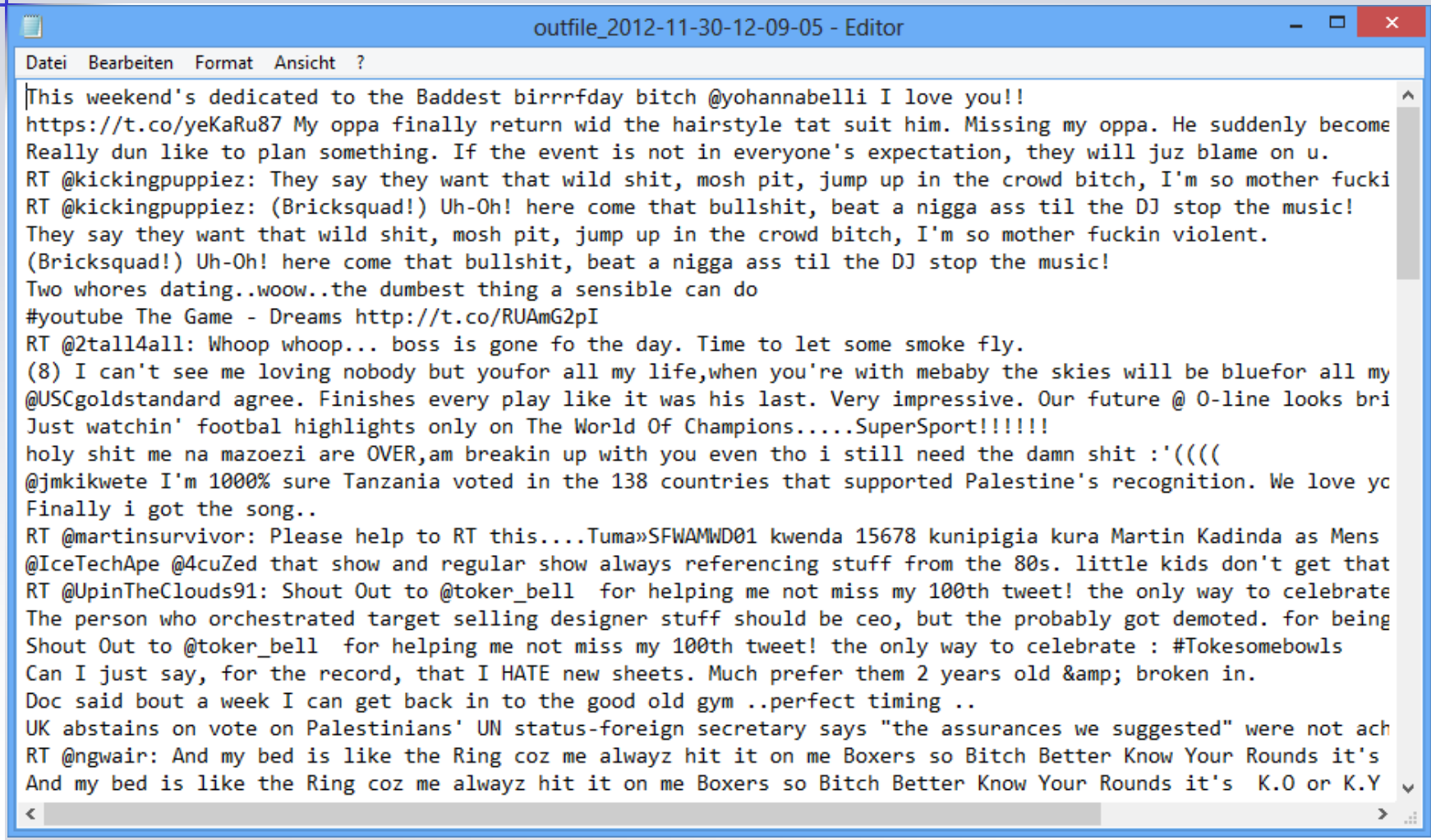
number of runs

TCCT: example tweet+location





TCCT: results stored as plain text



TCCT: 100 tweets word frequency visualized

geolocation Dar es Salam, radius 100km, search string "the"

TCCT: limitations

due to Twitter API restrictions

- search string obligatory (*the* occurs in all English texts)
- results limited to ~1% of all tweets within the last 7 days
- geolocation falls back to location in the Twitter User profile

other

- very low number of geotagged tweets (privacy concerns?)

TCCT: development

possible future features:

- storing tweets and metadata as XML file / SQL database
  → store metadata in a sociolinguistic database
- include multimedia information (pictures, videos) → provide context
- special treatment for tweets with hyperlinks → detect automatically generated tweets

3.4. Issues of compatibility and representativeness

- trans-ICE compatibility and national adaptation are in contrast
- maintain ICE compatible (1M words Ke/Tz each?)) and adapt monitor corpora (1 M texts?)?

4. Conclusion

- historical ICE-EA is possible
- ICE-EA 2.0 is possible
- ICE-EA 2.0 is necessary:
  50 years after independence+20 years after ICE-EA1
  to test hypotheses about “dynamic” developments” -
  maybe less dynamic than we thought

- but still in the brainstorming phase

References

Greenbaum, Sidney (1988). A proposal for an international computerised corpus of English. *World Englishes 7, 315.*

Hudson-Ettle, Diana/Josef Schmied [1996]. Manual to accompany The East African Component of The International Corpus of English (ICE-EA). Background information, coding conventions and lists of source texts.

McEnery, Tony/Andrew Wilson (1996*). Introduction to Corpus Linguistics*. Edinburgh U.P.

Meshrie, R./R.M. Bhatt (2008). *World Englishes: The Study of New Language Varieties.* Cambridge: CUP.

Schmied, J. (1990). Corpus-linguistics and the nativization of English. *World Englishes 9, 255-268.*

Schmied, J. (1991). *English in Africa. An Introduction.* Harlow/London: Longman.

Schmied, Josef (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, Josef (fc.)/Susanne Wagner. 'Comparing English on the Web, the case of ICEweb-East Africa'.

Schneider, E. W. (2007). *Postcolonial English: Varieties around the World*. Cambridge: C.U.P.