"Übrig sind noch 21 Gigabyte":
Perspektiven für professionelle Archivierung und Nachnutzung von soziolinguistischen Interviewdaten

Christian Mair, Freiburg
christian.mair@anglistik.uni-freiburg.de

Gefördert durch
DFG Deutsche Forschungsgemeinschaft

Grant MA 1652/12-1: *West African Englishes on the move*

UNI FREIBURG

1. Präludium: (Selbst)kritischer Rückblick auf die eigene 'Forschungsdatenspur'

Rohergebnisse zahlreicher Korpussuchen:

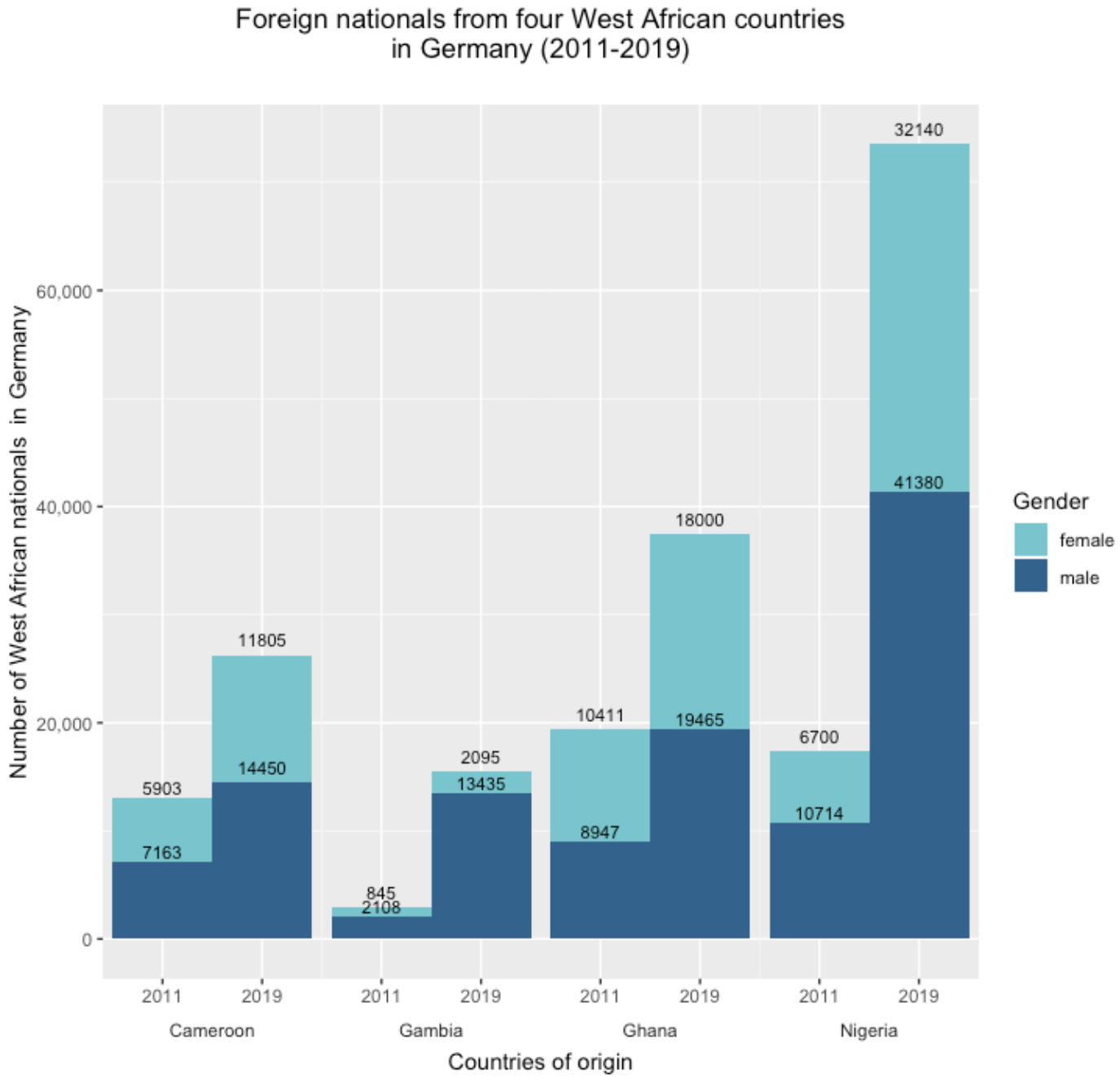
- verloren oder vernichtet
- unorganisierte Sammlungen von Papierausdrucken, aufbewahrt in Hinblick auf eigene Nachnutzung ...
- digital abgespeicherte unorganisierte Sammlung von Konkordanzen, zum Teil ohne technische Hilfe nicht mehr lesbar

Archivierung und Nachnutzung selbst erstellter Korpora:

Korpus/Korpora	Nachnutzung durch andere	Nachhaltige Archivierung
F-LOB/Frown (1 Mio Wörter Standardenglisch, Teil der "Brown"-Korpusfamilie	intensiv	CLARINO/Corpuscle (Norwegen), eduGain, CLARIN SPF; alternativ: Lancaster
ICE-Jamaica (Teil der International-Corpus-of-English-Korpusfamilie): geschriebene Texte und gesprochene Texte als orthographische Transkripte	rege	Archivierung und Zugang: Arbeitsgruppe M. Hundt (Uni Zürich), Web-Interface im Aufbau
ICE-Jamaica, Audio-Dateien der gesprochenen Texte	selten	nicht nachhaltig: persönlich/lokal in Freiburg
Corpora of Cyber-Cameroonian (CCC), -Jamaican (CCJ), -Nigerian (CCN): > 850 Mio Wörter aus diasprorischen postkolonialen Internetforen	sehr selten	nicht nachhaltig: persönlich/lokal in Freiburg; besonders kritisch bei CCC & CCJ
West African Englishes on the Move (DFG MA 1652/12)	bislang ein Fall	nicht nachhaltig

2. Fallstudie: WAfrE on the move

Demographie: Westafrikanische Einwanderung nach Deutschland (2011-2019)



Linguistik: (Westafrikanisches) Englisch, Deutsch, Pidgin – drei Lingue franche im Wettbewerb

	Englisch	**Deutsch**	**Pidgin**
"Migrantische" Perspektive	Postkoloniale Prestigesprache und Weltsprache; universell als LF einsetzbar	notwendig für Integration, aber ungeheuer schwer zu lernen; Quelle der Frustration	essentiell als informelle LF einer heterogenen Migrant:innengruppe; gelegentlich gruppenspezifischer Geheimcode
"Deutsche" Perspektive	Prestige als Weltsprache, LF-Gebrauch innerhalb der nationalen Grenzen suspekt	die **eine** Sprache (Singular) als Schlüssel zur Integration	Ist das Englisch? Versteh aber kein Wort!

Datenerbe: > 51 Stunden ethnographische Interviews

Category of interview	Number of interviews	Persons interviewed (totals per category)	Duration (hours: minutes)
Focus group	13	41 (16 Nigeria, 22 Cameroon, 2 The Gambia, 1 Sierra Leone)	16:25
Individual	36	36 (29 Nigeria, 1 Benin, 2 The Gambia, 3 Ghana, 1 Guinea)	32:51
Ancillary material	6	6 (all Nigerian)	1:45

Zustand

- Orthographische Transkription aller Interviews in ELAN
- Übersetzung aller nichtenglischen Passagen (Pidgin, Yoruba, etc.) für fast alle Interviews
- Projektspezifische weitere Annotationen (z.B. Verbal- und Nominalflexion) für wenige Texte
- Ca. 60 Prozent korrekturgelesen (projektinterne Benchmark: "to a level of correctness we don't have to be ashamed of")
- Zeitpfeile mit Hinweisen auf potentiell ergiebige Themen und Phänomene (ca. 25 Prozent des Materials)

Nachnutzung

- Linguistik: Noch zahlreiche interessante linguistische Phänomene, die im Projektrahmen nicht behandelt werden konnten
- Diskurse des Rassismus und der Diskriminierung
- Oral History
- Soziologie und Migrationsforschung
- …

Pidgin: useful, good vibes (and some lingering reservations)

I: May I ask you about your ah about Pidgin English (.) Naija

P 3: Ah if na dat one I dey house o @@@ [*ah if it's that, I'm ready*]
I dey Kampe [*I'm good to go*; Kampe = 'strong', 'solid']

I: But you're not teaching– you're not teaching it to your children

P 3: Dat one na street language [*that is street language*]

P 1: We don't speak it with the children
We speak it when we meet our African brothers and sisters outside (.) when we go out you know

(FG_4, 04:28-4:58)

P1, cam/f, FG 8

and they claim sometimes
not to un- understand you because they want to make you feel you have a heavy accent
and this happened especially during English class in in uhm in in my secondary school days uhm to me I I spoke the best English than any other kid could do in the whole school
but she always had to do like she is trying to correct me even though she herself had a German English-speaking accent (P1, Fg 8)

but on the street if I meet someone socially
uhm that wants me or that I get to speak English
with they don't make me feel that way they they feel very excited instead (P1, Fg 8)

I: who are fake oyinbo

P: all those that's not really white people you know we call them fake oyinbo because they are not ehn German or England or France you know all those uhm uhm Syrian Pakistan all those uhm uhm Iran @ @ all those are fake oyinbo oder uhm uhm Bulgaria (I_ng_6_f)

P1: make wi say di kontries dem wey e bi powerful fo Europe
na yi wi di call Sahra or Oyinbo
and di one wey dem kommot di other kontries wey dem no di do well like Slovakia Romania na fake Oyinbo wi di call dem
[*Let's say the countries that are powerful in Europe, it's them that we we call Sahra or Oyinbo. And those that come from the other countries that don't do well, like Slovakia Romania, we call them fake oyinbo*.]

I: ok

(FG_8_ng_ca)

3. Erste positive Erfahrungen mit zwei Text+-Repositorien

UdS, Saabrücken

- das ist zwar eine Menge Holz, sowohl was die reine Menge der Daten (21GB) als auch was die Art der Daten (gesprochene Sprache, eigentlich nicht unser Schwerpunkt) angeht, aber wir archivieren das gerne.

BAS, München

- danke für Deine Mail und das Word-Dokument. Das klingt sehr spannend!
-
- Zum einen: wunderbar, dass es a) Transkriptionen gibt, und b) dass die Daten bereits anonymisiert sind. Auch ist die Menge an Daten beeindruckend – ca. 50 Stunden.
-

Wichtig is daher die Trennung zwischen der Präsentationsschicht und den zu archivierenden Daten. Die Präsentation sollte in Freiburg entwickelt und gehostet werden, solange wie sie dort unterstützt werden kann. […] Wir müssen uns überlegen, wie tief wir den Datenschatz erschließen wollen, typischerweise sind wir (CLARIND-UdS) eher grob-granular unterwegs und die Feinstrukturierung ist innerhalb des Korpus angesiedelt, bei Audiodaten ist aber eine feinere Granularität durchaus gängig (schon weil sich niemand auf der Suche nach etwas Bestimmtem durch das ganze Material durchhören will

Der Idealweg […]

(https://clarin.phonetik.uni-muenchen.de/BASWebServices/interface/Coala)

5 Mustertabellen des Webdienstes COALA (https://clarin.phonetik.uni-muenchen.de/Bas/BASWebServices/EXAMPLES/examples-COALA.zip)

Diese Tabellen enthalten die Metadaten zu den Aufnahmen, den Transkripten, den Sprecher/innen, den Sitzungen, usw. Dann ladet ihr die Dateien hoch und COALA erstellt daraus CMDI-kompatible Metadaten, die u.a. einen Import in unser Repository ermöglichen.

Im Anhang wird klar, dass in dem Projekt viel Potential steckt, ich kann mir sofort einen sehr interessanten und lebendigen Webauftritt vorstellen, mit fein-granularer Suchfunktion und guter Erschließung der einzelnen Takes. Dieser Webauftritt (die Präsentationsschicht) ist praktisch nicht nachhaltig archivierbar, erfahrungsgemäß ist alle 5–10 Jahre eine Grundrenovierung der dahinterliegenden Software notwendig.

In der Praxis wird es nicht ganz glatt gehen und an verschiedenen Stellen Anpassungen benötigen.

[…] verteilen wir die Aufgaben, klären die Zugriffsmöglichkeiten (unbeschränkter Zugriff, Zugriff nur für akademische Nutzer/innen, individuelle Lizenz) und kopieren die Daten nach München, wo sie erst einmal für Dritte unzugänglich zwischengespeichert werden. Sobald die Erstellung der Metadaten abgeschlossen ist, können wir den eigentlichen Import-Vorgang starten – am Ende stehen die Korpora dann im BAS Repository.

5. Schluss

Erste Kontaktaufnahme mit NFDI

- geprägt von Offenheit auf Seiten der Repositorien (so, wie es sein soll)
- zeigt mir die Abgründe meines sprachtechnologischen Unwissens, was Langzeitarchivierung betrifft
- motiviert mich sofort zu Recherche, um Jargon und Akronyme zu entschlüsseln
- Management der Archivierung wesentlich einfacher als Förderung von Nachnutzung und kumulativem Fortschritt der Forschung auf der Basis der einzelen Korpora, Datenressourcen, "Sammlungen"