Christian Fandrych , Anna Iliash, Cordula Meißner, Franzsika Wallner, Kathrin Weigert — Universität Leipzig,
Elena Frick , Thomas Schmidt — IDS Mannheim
Hanna Hedeland, Daniel Jettka — Universität Hamburg

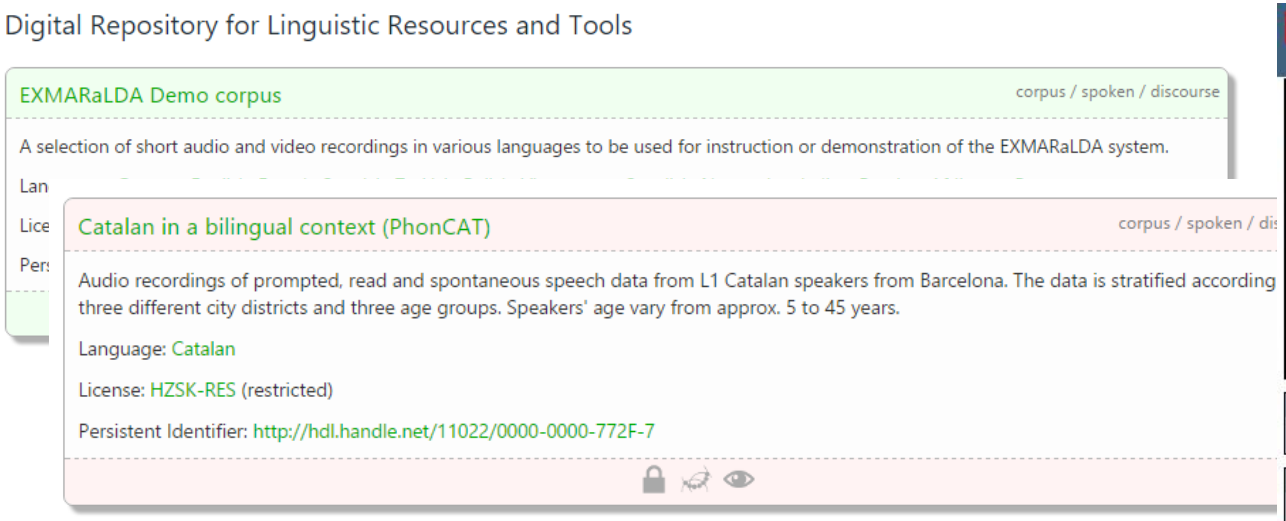
Wer bist Du, Nutzer?

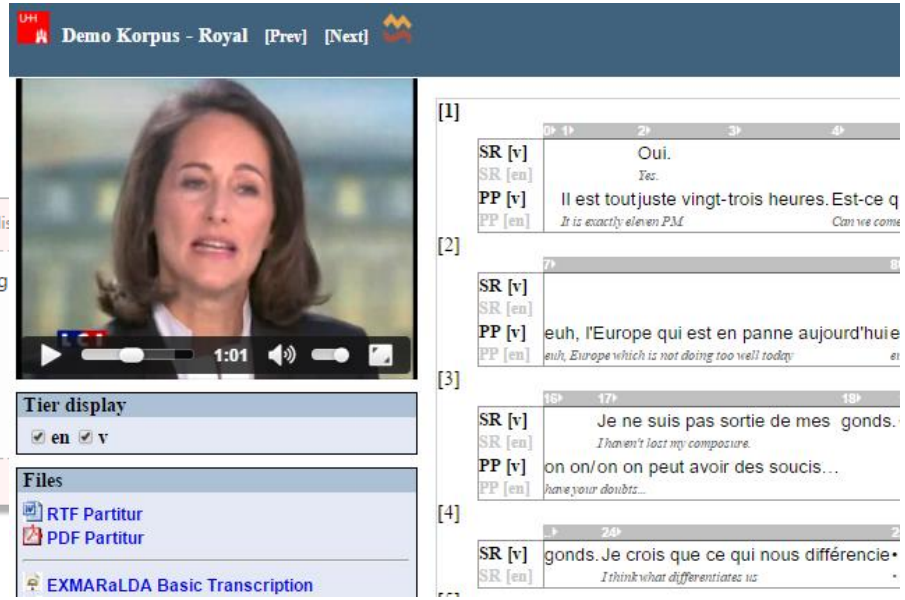
Eine Studie zur Nutzung dreier Korpus-Plattformen für mündliche Daten

Plattformen für Mündliche Korpora

- Audio/Videoaufnahmen natürlicher Interaktion
- Erschlossen durch Metadaten / Transkription / Annotation(en)
- Zur Nutzung für empirische linguistische Analysen
- International, z.B.
  - Frankreich: CLAPI, ESLO
  - Czech National Corpus
  - …
- In Deutschland…

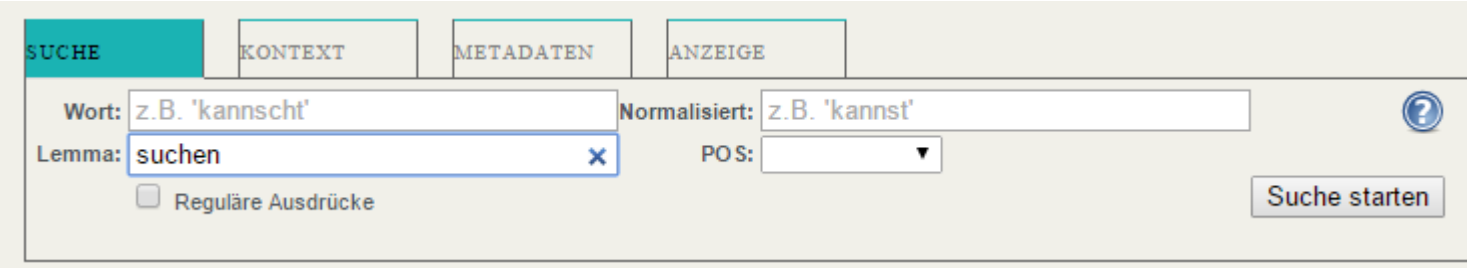
Hamburger Zentrum f. Sprachkorpora (HZSK)

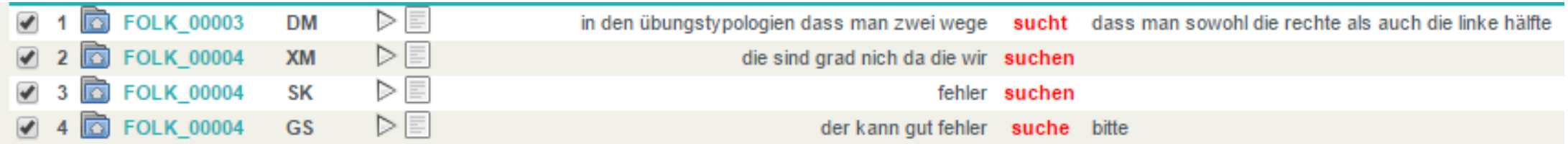




- Hervorgegangen aus dem „INF“-Projekt des SFB 538
- Schwerpunkt: Korpora zu individueller, gesellschaftlicher Mehrsprachigkeit, mehrsprachige Kommunikation
- CLARIN-Zentrum – Hosting weiterer Ressourcen, zukünftig: INEL (Sprachdokumentation)
- Aktuell ca. 650 registrierte Nutzer (seit 2011)

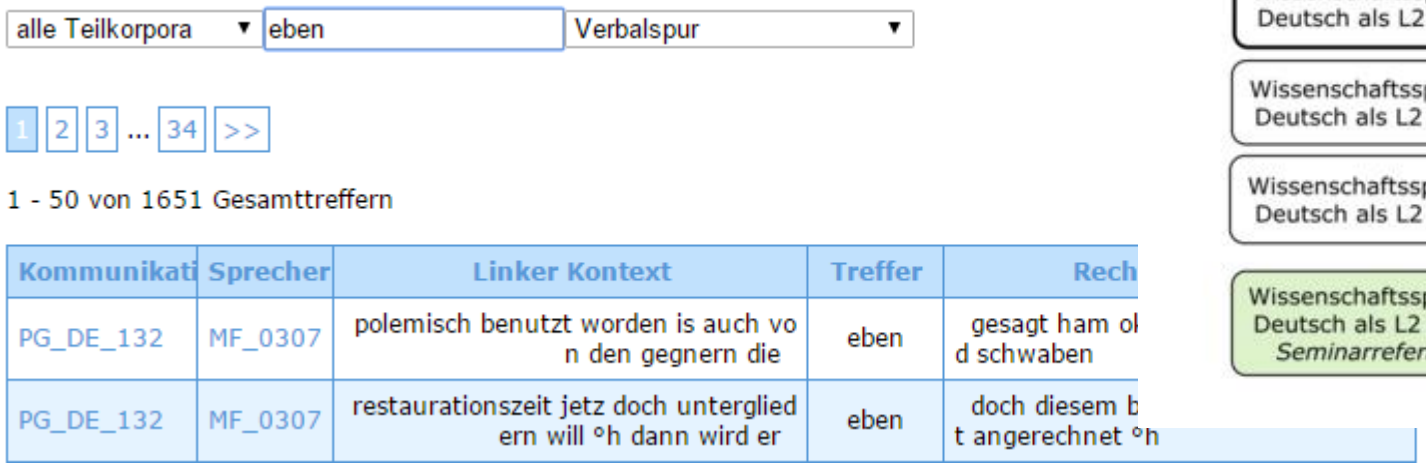
Datenbank für Gesprochenes Deutsch (DGD)





- Korpusplattform des Archiv für Gesprochenes Deutsch
- Browsing, Query, Download für 24 Korpora des Deutschen
- Variationskorpora (Dialekte, Sprachinseln), Gesprächskorpora (Sprache in Interaktion)
- Forschungs- und Lehrkorpus Gesprochenes Deutsch (FOLK)
- Aktuell knapp 4.700 registrierte Nutzer (seit 2012)

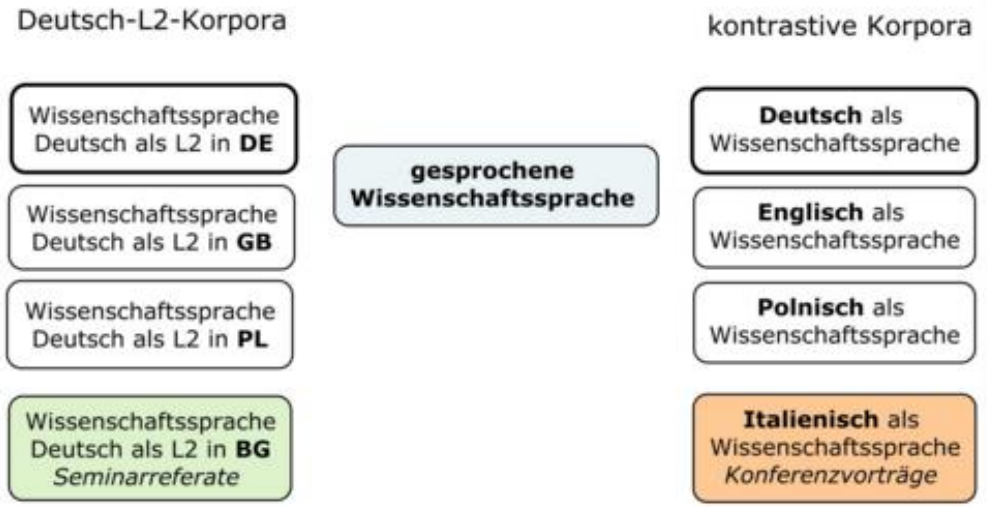
Gesprochene Wissenschaftssprache Kontrastiv (GeWiss)


alle Teilkorpora ▾ | eben | Verbalspur ▾

1 2 3 ... 34 >>

1 - 50 von 1651 Gesamttreffern

Kommunikati	Sprecher	Linker Kontext	Treffer	Rech
PG_DE_132	MF_0307	polemisch benutzt worden is auch vo n den gegnern die	eben	gesagt ham ol d schwaben
PG_DE_132	MF_0307	restaurationszeit jetz doch unterglied ern will °h dann wird er	eben	doch diesem b t angerechnet °h



- Herder Institut Leipzig, Aston University, Universität Wrocław
- Empirische Ressource für vergleichende Analysen gesprochener Wissenschaftssprache
- L1/L2-Daten in Deutsch, L1 in Englisch und Polnisch, Prüfungsgespräche, Vorträge
- Aktuell 500 registrierte Nutzer (seit 2013)

„The digital revolution has the potential to do for spoken language what the printing press did for written language. […] Researchers have the tools to transform access to the spoken word, preserving an essential aspect of cultural heritage, and **stimulating a diverse set of communities**.“

“[However, we know] far less about how best to support access to extended sessions of spontaneous speech. There is also a need for **focused assessment of the needs of specific user groups** that to date have been understudied. Some examples include teachers, scholars in the humanities and social sciences [...].”

Goldman, J. / Renals, S. / Bird, S. / de Jong, F. / Federico, M. / Fleischhauer, C. / Kornbluh, M. / Lamel, L. / Oard, W. / Stewart, C. / Wright, R. (2005)

Accessing the Spoken Word.

International Journal on Digital Libraries, 287-298, [http://hdl.handle.net/1842/947]

Nutzerstudie

- WER? Nutzerprofile, Nutzergruppen
- WAS? Funktionalität, Daten
- WIE? Arbeitsweisen, Usability

Fragebogenstudie, webbasiert, 128 Fragen an ca. 5000 Nutzer

Qualitative Interviews mit „Power Users“ (10 für DGD, 5 für GeWiss)

POSITIVE VERZERRUNG!

Ergebnisse: Allgemeines

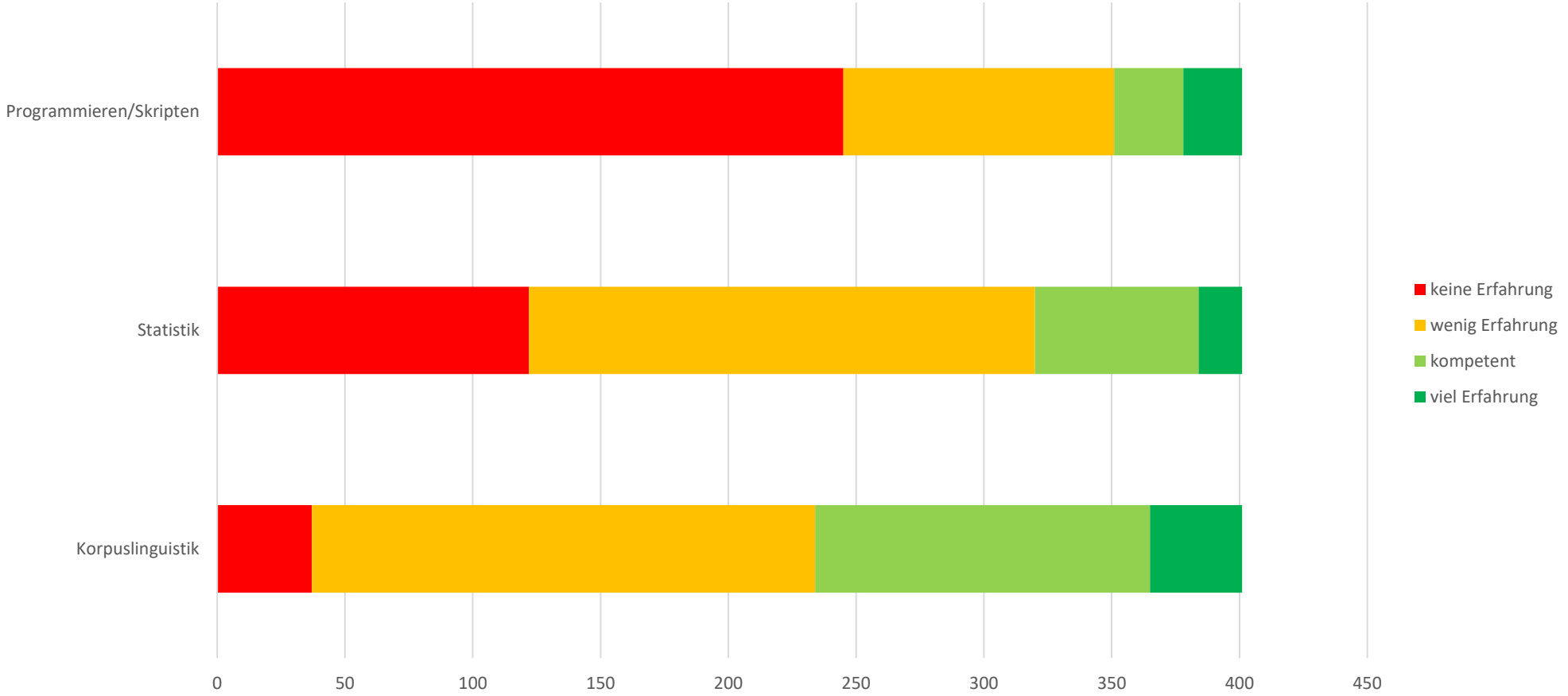
- 669 Umfrageteilnehmer, 401 vollständige Rückläufe (=8%)
- Typischer Nutzer
  - weiblich (67%)
  - zwischen 21 und 30 Jahren (54%)
  - L1 Deutsch (66%)
  - lebt und arbeitet in Deutschland (71%)
  - im Studium (59%) bzw. graduiert (40% auf Doktorandenniveau oder darüber)

Welche Bereiche interessieren Sie?

Germanistische Linguistik	238	59,35%
Deutsch als Fremdsprache	199	49,63%
Korpuslinguistik	196	48,88%
Gesprächsforschung	195	48,63%
Spracherwerb	172	42,89%
Soziolinguistik	154	38,40%
Pragmatik	145	36,16%
Fremdsprachenunterricht	132	32,92%
Konstrastive Linguistik	122	30,42%
Dialektologie	114	28,43%
Phonetik	93	23,19%
Computerlinguistik	84	20,95%
Wissenschaftssprache	83	20,70%
Lexikographie	67	16,71%
Korpustechnologie	65	16,21%
Sonstiges (bitte angeben)	46	11,47%

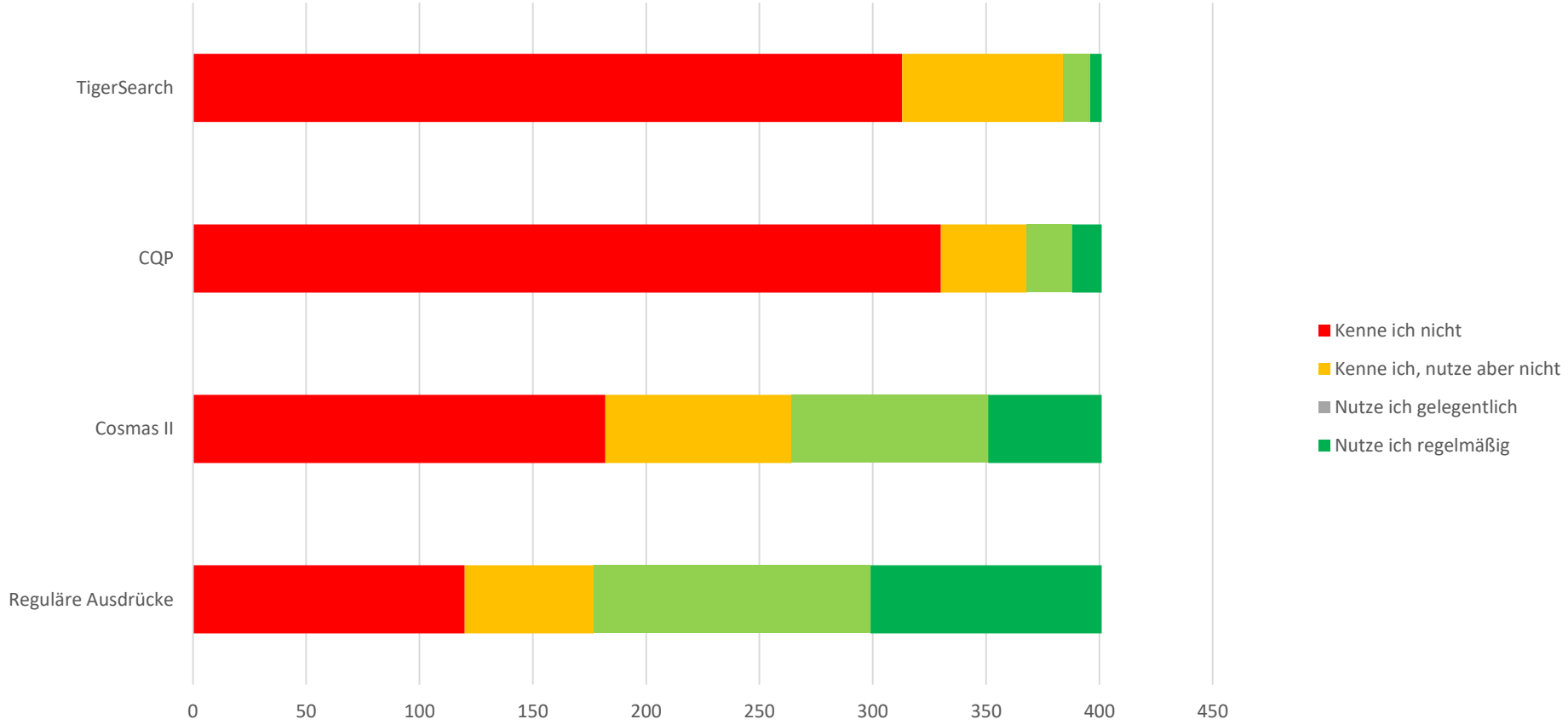
Erfahrungen

Frage 10 – „Bitte beurteilen Sie Ihre Erfahrung in folgenden Bereichen“

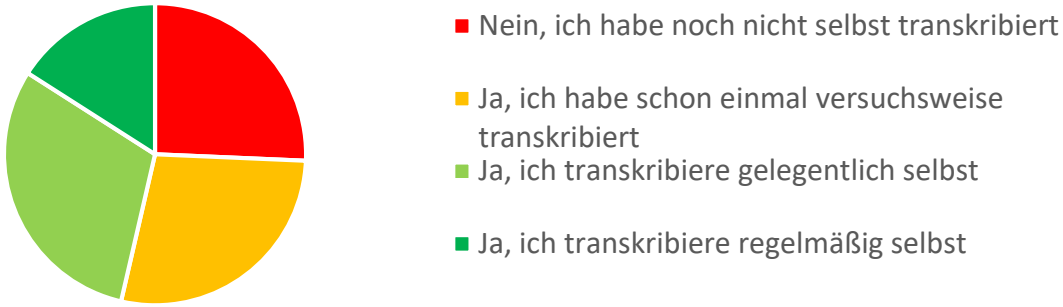


Erfahrungen

Frage 11 – „Welche der folgenden Suchabfragesprachen kennen/nutzen Sie?“

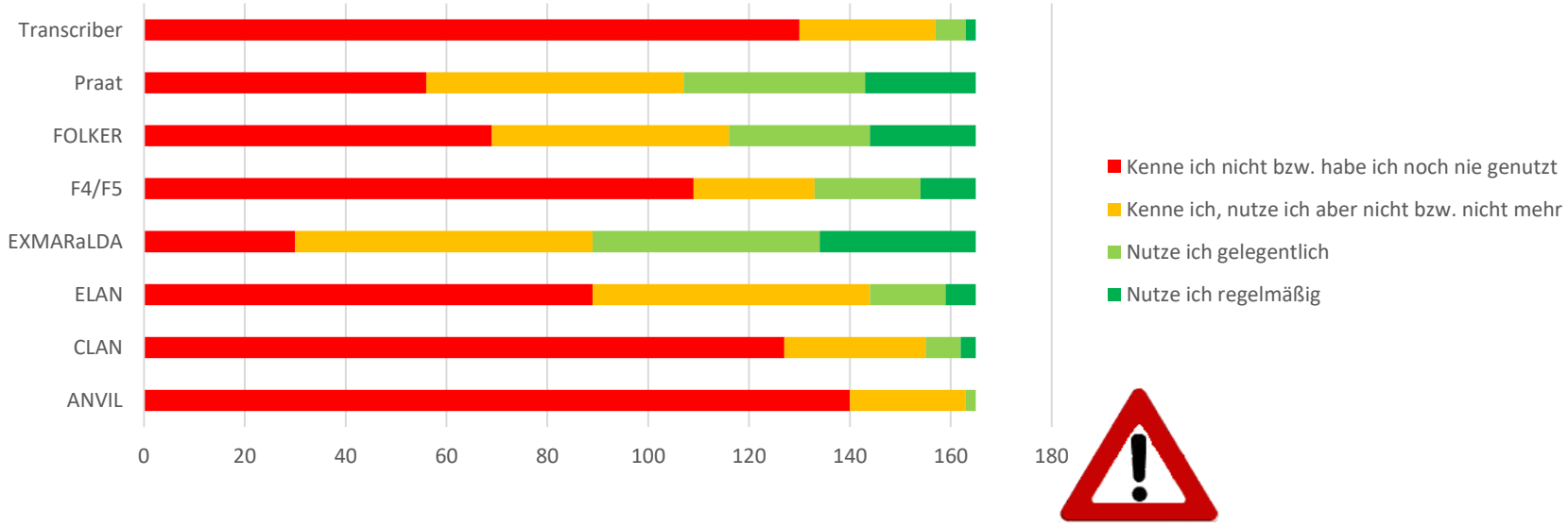


Frage 13 – „Verfügen Sie über eigene Transkriptionserfahrung?“



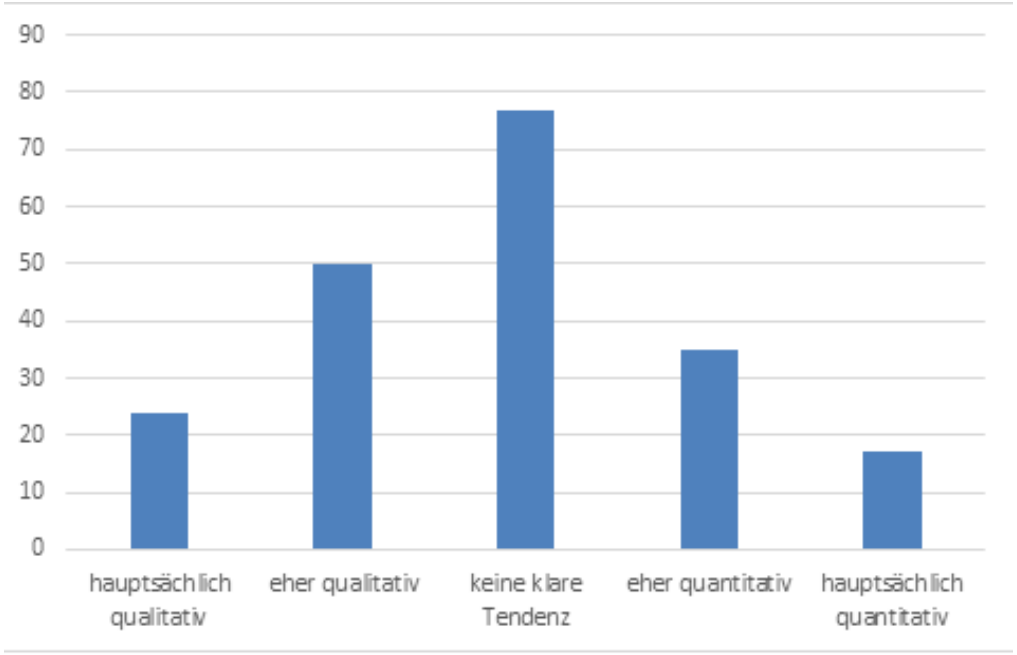
56% - Transkription mit Textverarbeitung / 55% - Transkription mit spezialisierten Editoren

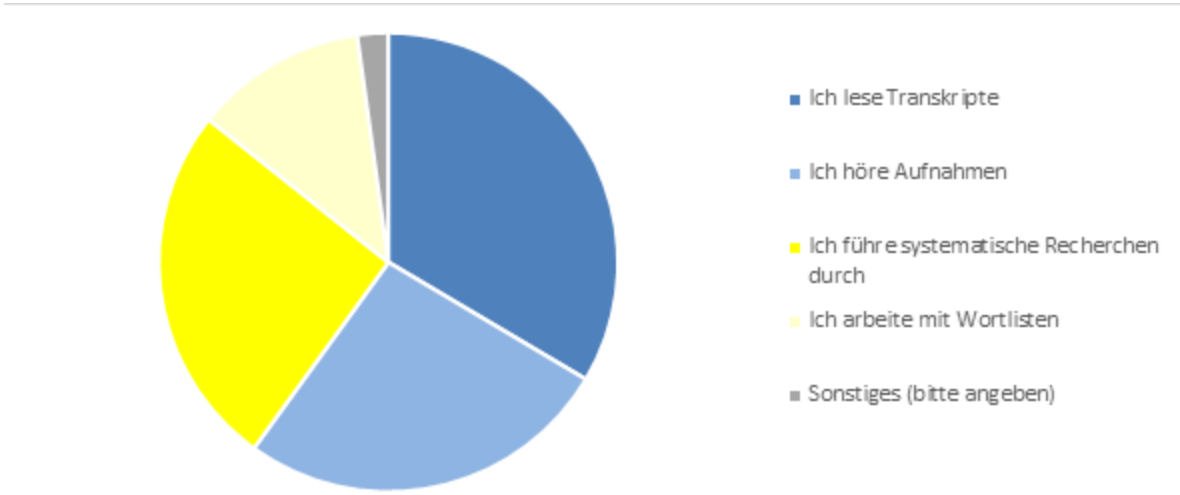
Frage 16 – „Mit welchem/en spezialisierten Transkriptionseditor/en arbeiten Sie?“



Methodische Herangehensweisen

Frage 33 – „Wie lässt sich Ihre methodische Herangehensweise am besten beschreiben, wenn Sie mit der Datenbank für gesprochenes Deutsch (DGD) arbeiten?“

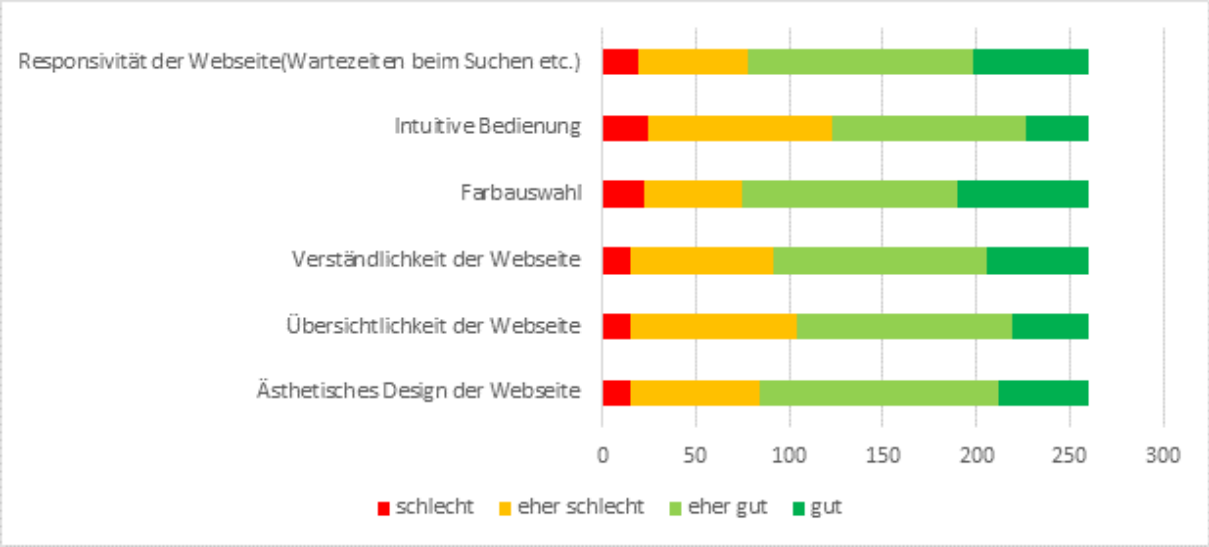




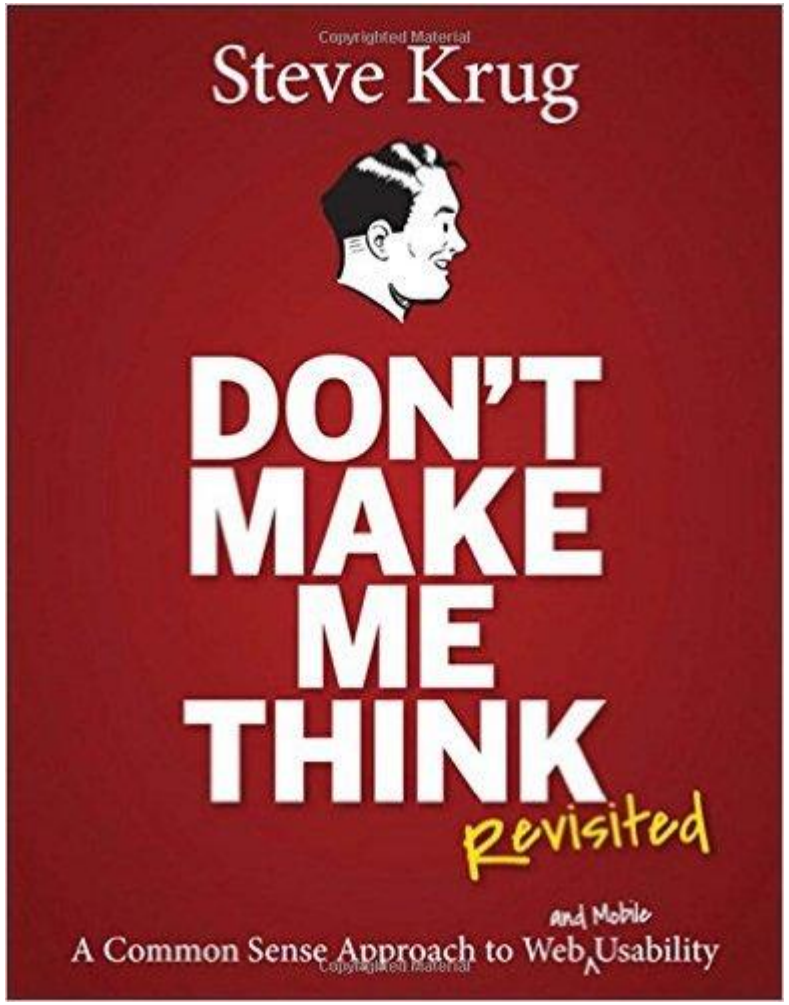
Frage 34 – „Was ist Ihre Haupttätigkeit, wenn Sie mit der Datenbank für gesprochenes Deutsch (DGD) arbeiten?

- „Download first“ – Ausdrucken oder Bearbeitung mit Office Software (Word, Excel) weit verbreitet

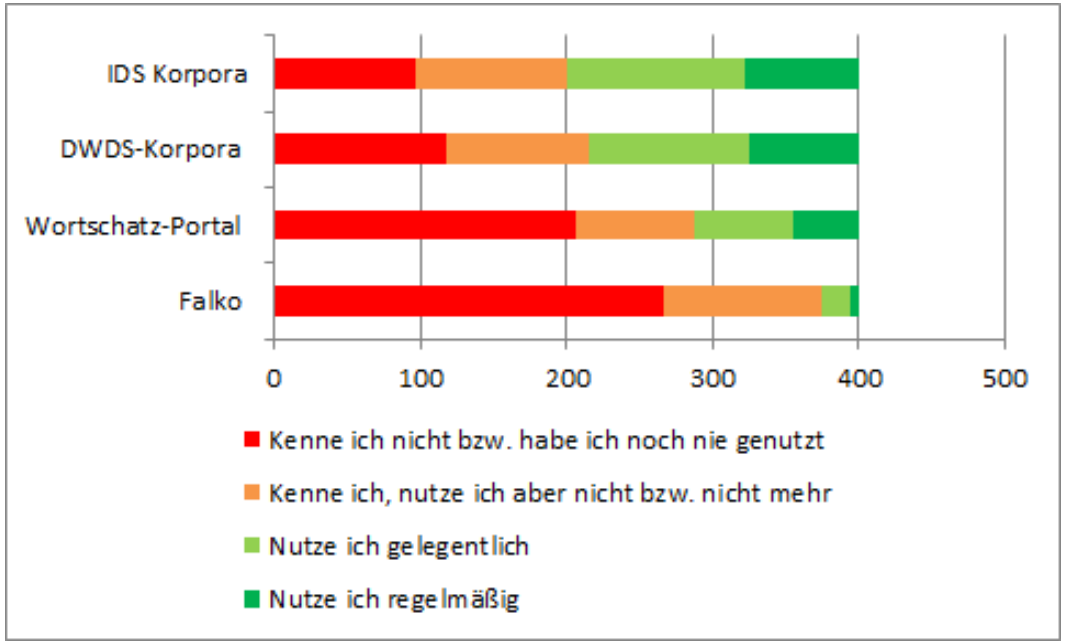
Usability



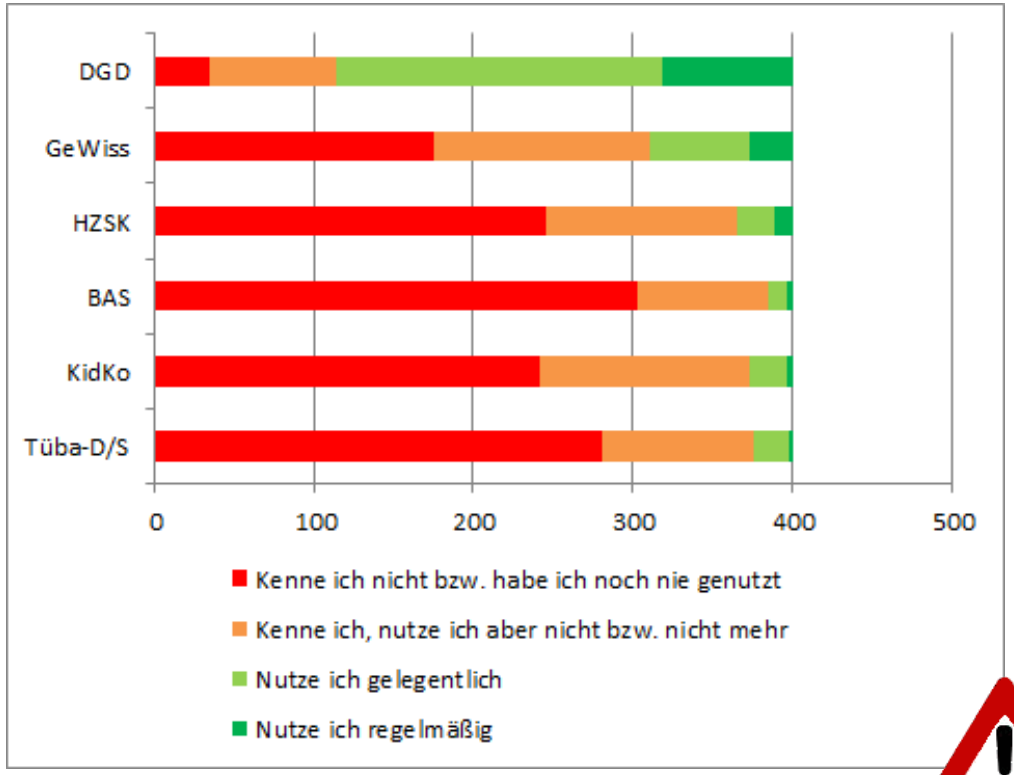
- „unklar“, „verwirrend“, „nicht intuitiv“, „mühsam “
- „autodidaktisch“, „so schnell wie möglich“, „learning by doing“
- viel komplexe Funktionalität vorhanden und gewünscht!
- einige „Wege zum Erfolg“ nicht bekannt



Kontrastive und kombinierte Nutzung



Schriftsprachliche Korpora des Deutschen

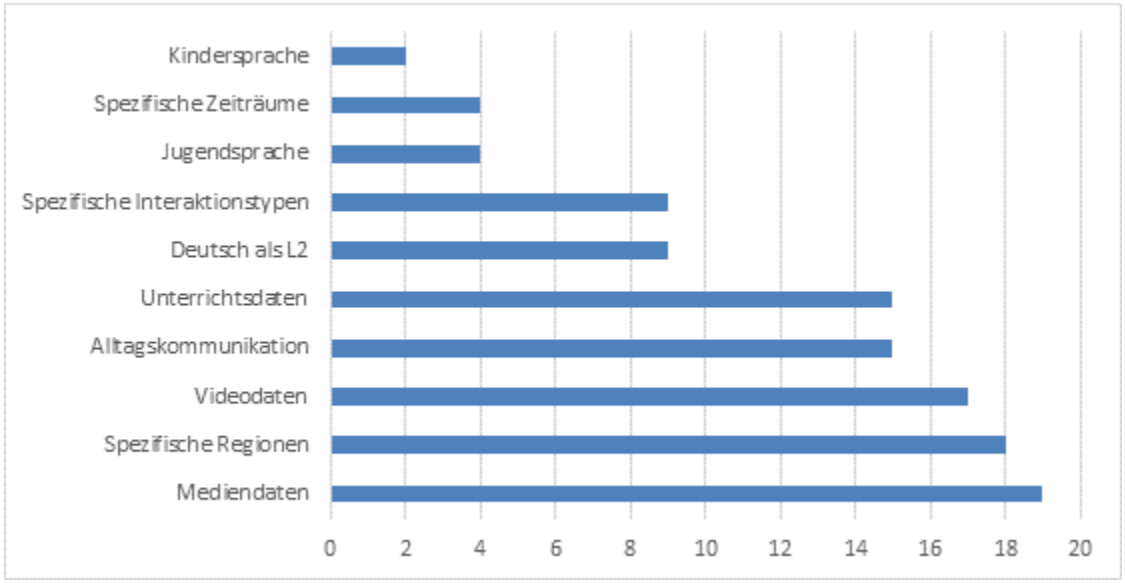


Mündliche Korpora des Deutschen

Kontrastive und kombinierte Nutzung

- Kontrastive Nutzung (= Auswertungen auf mehr als einem Korpus): 30% aller Nutzer
- Für DGD: davon 46% mit geschriebenen, 39% mit anderen gesprochenen Korpora, 33% mit „eigenen Daten“, 28% innerhalb der jeweiligen Plattform
- Für GeWiss: 83% innerhalb der Plattform, 43% außerhalb
- Beispiele:
  - Korpora aus DGD mit Korpora aus HZSK / GeWiss
  - Vergleich mit deutschen schriftsprachlichen Referenzkorpora: DeReKo und DWDS
  - Sprachübergreifend: SBCSAE, BNC / Spokes / C-ORAL_ROM
  - eigene Daten: Arzt-Patienten-Interaktion, L2 Lernerdaten
  - IBK-Daten (Chat etc.)

Nutzerwünsche: Daten



Mehr Vielfalt:

- Interaktionstypen: Arzt-Patienten-Interaktion, Konfliktgespräche, andere akademische Disziplinen (GeWiss)
- Regional: ehemalige DDR, Schweiz, Norddeutschland
- Sprecherypen: Kinder, Jugendliche, L2-Lerner
- Zeitlich: nach der Wende, “älter”
- Sprachlich: weitere (osteuropäische) Sprachen (GeWiss)

Mehr:

- 11% (DGD) bzw. 19% (GeWiss) unzufrieden mit “Quantität der Daten”
- “Mehr Daten sind immer besser”
- “Auswertung nach Methode X” nicht möglich, weil Datenmenge nicht ausreichend
- “Auswertung nach Phänomen Y” nicht möglich, weil nicht ausreichend Belege

Erkenntnisse (1)

- Heterogene Nutzerschaft
  - Heterogene Vorkenntnisse
  - Heterogene Annäherungen an die Daten
  - Daten > Methoden

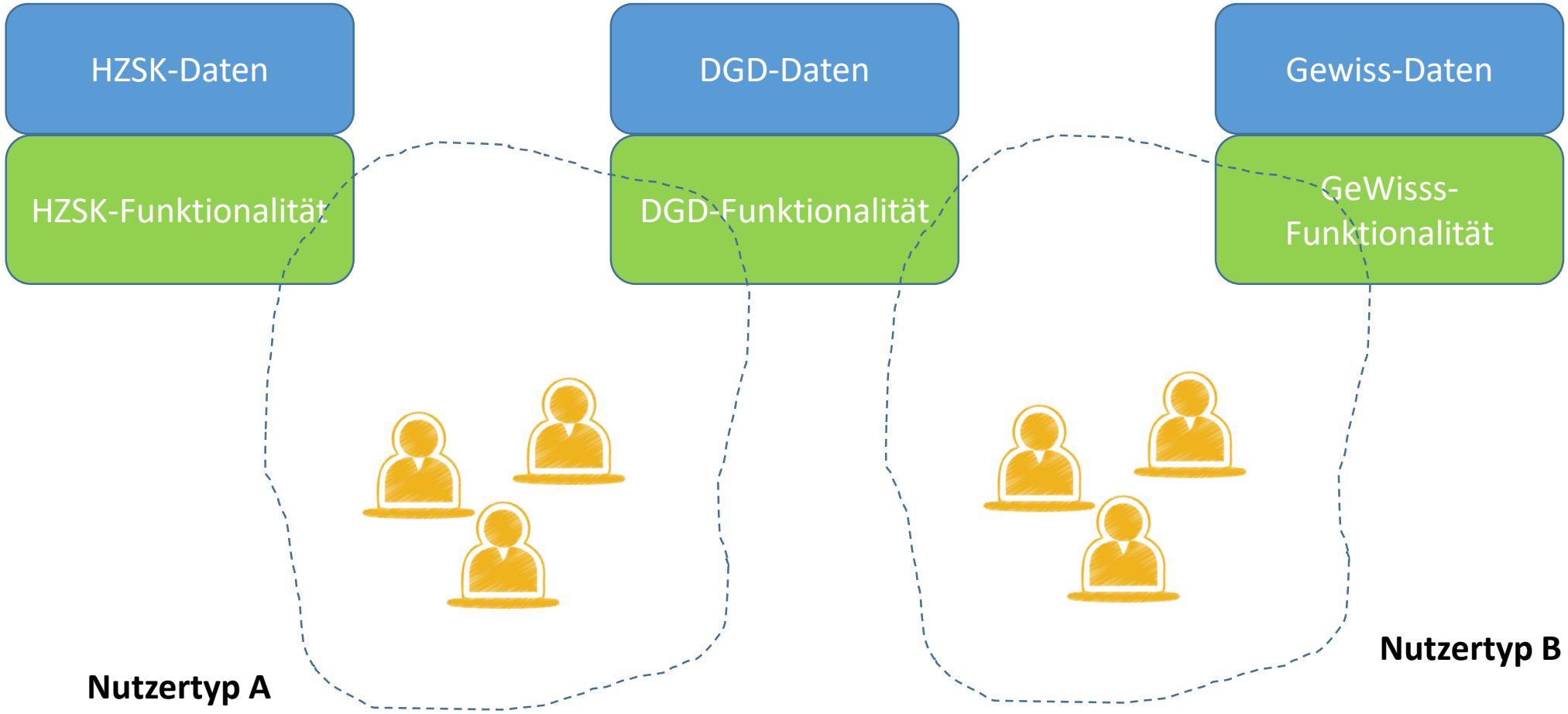
- Systematisches Retrieval nicht die dominante Herangehensweise
- „Download first“-Paradigma nicht obsolet

- „Standard“-Methoden für Großteil der Nutzer nicht verfügbar
- Nicht alle Funktionen relevant für alle Nutzergruppen
- Gesamte Funktionalität (schon) zu umfangreich

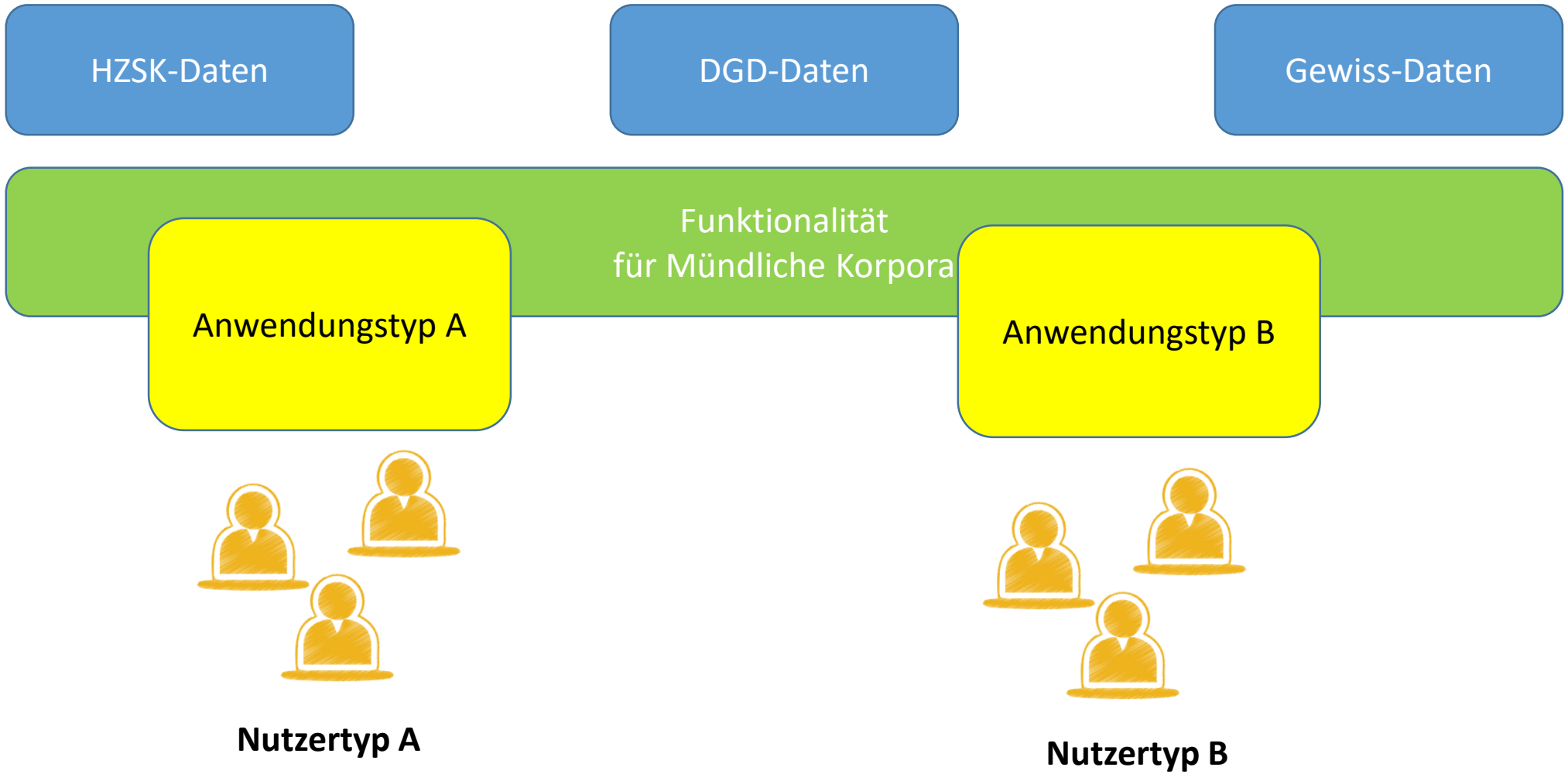
Erkenntnisse (2)

- Kontrastive / Kombinierte Nutzung häufig
- Vielfältige Hinzunahme weiterer Daten
- “[Tools widely used by corpus linguists] all offer a different user-experience, because each tool is created in isolation and thus offers a different user interface, control flow, and functionality.” (Anthony 2013: A critical look at software tools in corpus linguistics. In: Linguistic Research 30, 141-161)

Konsequenzen



Konsequenzen



Ausblick

- Federated ~~Content Search~~ Access, z.B.
  - Alltagssprache (FOLK/DGD) vs. Akademische Sprache (GeWiss)
  - L2-Lernerdaten (GeWiss) vs. L2-Lernerdaten (HaMaTaC/HZSK)
  - Niederdeutsche Dialektdaten (SiN/HZSK) vs. Südwestdeutsche Dialektdaten (SW/DGD)
- Drei-Ebenen-Architektur: Daten – „Business Tier“ – Anwendung
- Basis-Architektur / Methoden-Baukasten
- Zielgruppenspezifische Ausdifferenzierungen, z.B.
  - „Didaktisches Szenario“ (Sprachvermittlung) – beispielorientiertes Arbeiten, qualitativ orientiert, am Laien orientierte Visualisierungen
  - „Variationslinguistisches Szenario“ – 100% Recall, systematische Korrelation mit Metadaten, quantitativ orientiert, am Experten orientierte Visualisierungen

hzsk hamburger zentrum für sprachkorpora

DGD DATENBANK FÜR GESPROCHENES DEUTSCH

GeWiss Gesprochene Wissenschaftssprache

21	FOLK_00070	HS		zeit auch schon festgelecht war untendrunter vielen **dank** das sin die fümf komma
22	FOLK_00069	HG		äh vielen **dank** äh das äh bedeutet
23	FOLK_00064	HG		gut äh v äh vielen **dank**
24	FOLK_00126	SF		vielen **dank**
25	FOLK_00069	HG		ja herr professor wittke vielen **dank**
26	FOLK_00021	DK		ey viel vielen **dank** subber
27	FOLK_00070	PC		vielen **dank**
28	FOLK_00173	TB		super vielen herzlichen **dank**
29	FOLK_00064	HG		gut vielen **dank** also jetzt käme glaub der herr conradi dran nich oder
30	FOLK_00021	CH		pascal vielen **dank**
31	FOLK_00007	GS		dann mal vielen **dank** bis nächste woche