Digitale Sprachressourcen für die angewandte Linguistik: Chancen, Herausforderungen, Desiderate

GAL Sektionentagung 2019 in Halle
19. September 2019

Prof. Dr. Angelika Storrer

UNIVERSITÄT MANNHEIM

Was sind digitale Sprachressourcen?

Digital gespeicherte Datenbestände, die für die sprachwissenschaftliche Nutzung aufbereitet sind, sowie digitale Tools, die auf ihnen operieren.

Dazu gehören u.a.:

- Digitale Korpora (Referenzkorpora, Spezialkorpora)
- Digitale Wörterbücher und lexikalische Datenbanken
- Grammatische Informationssysteme
- Werkzeuge zur Recherche und Auswertung digitaler Datenbestände
- Werkzeuge und Umgebungen zur automatischen und manuellen Annotation von Datenbeständen
- Plattformen, die verschiedene Ressourcentypen integrieren.

DWDS

Der deutsche Wortschatz von 1600 bis heute.

https://www.dwds.de

Schutzmann

berlin-brandenburgische
AKADEMIE DER WISSENSCHAFTEN

Schutzmann · *Substantiv*

Wörterbücher

465 000 Einträge in verschiedenen Wörterbüchern.

Konfekt, das

Bedeutungen

1. Pralinen o. Ä.

Textkorpora

Dreizehn Milliarden Belege aus historischen und gegenwartssprachlichen Textkorpora.

r Himel am Himmel sein wahres	Gold	Fell/ oder seine Gold Wolle/ auf Erd
sein wahres Gold Fell/ oder seine	Gold	Wolle/ auf Erden herab.
stund stäts der Aehren schwares	Gold	.
Du streust	Gold	von deinen Hohes/ unsre Wesen an
Wen sollte/ der Rubin in diesem	Gold	/ nicht laben?
der andre macht verlebt/ ist von	Gold	/ und scharf gespitzet.
ester; lässt/ das mehr als goldne	Gold	ihrer Haare/ ledig fliegen.
brüsten? daß mich solt/ wie nach	Gold	/ nach dir gelüsten?
Dessen Haupt von feinem	Gold	/ die Brust und Arm von Silber/ der

Statistische Auswertungen

Häufigkeiten, typische Verbindungen und zeitliche Verläufe über 400 Jahre.

Chancen (am Beispiel digitaler Korpora)

Mit digitalen Korpora kann man Sprachstrukturen und Verwendungsmuster an authentischen Sprachdaten untersuchen und Hypothesen auf der Basis großer Datenmengen überprüfen.

Digitale Korpusplattformen bieten Optionen der Auswertung, die im prädigitalen Zeitalter nicht bzw. nicht mit vertretbarem Aufwand möglich gewesen wären.

Beispiel: Man kann Textsammlungen mit mehreren Milliarden Textwörtern nach einem Suchwort durchsuchen und sich die Frequenzentwicklung als Wortverlaufskurven anzeigen lassen.

Wurde das Wort „Schutzmann“ zurecht ins „Lexikon der bedrohten Wörter“ aufgenommen?

Linguistische Annotationen: Nutzen und Grenzen

Linguistische Annotationen wie Lemmatisierung und Wortarten-Tagging (POS-Tagging) sind nützlich, um die Präzision (precision) und die Ausbeute (recall) von Suchanfragen zu verbessern.

Z.B. kann man mit POS-Tags die Präzision von Suchanfragen zu Homographen verbessern, die unterschiedlichen Wortarten angehören.

Ich schicke dir die Datei. ⇔ Er trägt immer *schicke Hemden.*

Bei großen Korpora werden die Annotationen mit sprachtechnologischen Werkzeugen **automatisch erzeugt**. Diese arbeiten zwar gut, aber **nicht fehlerfrei**.

Bedeutung

umgangssprachlich

1. **drückt Zustimmung aus** jawohl; es ist recht

   BEISPIELE:

   »wollen wir gehen?« »**Okay**!«

   »konntest du alles erledigen?« »**Okay**!«

2. ⟨etw. ist **okay**⟩ etw. ist gut, ist in Ordnung

   BEISPIELE:

   wenn ich erst das Examen habe, ist alles **okay**

   bei den Kindern war alles **okay**

   ⟨jmd. ist **okay**⟩ jmd. ist tüchtig, gesund

   BEISPIELE:

   der neue Mitschüler ist **okay**

   er war zwei Tage krank, aber jetzt ist er wieder **okay**, fühlt er sich wieder **okay**

Antwortpartikel (Responsiv)

POS-Tag: PTKANT

Prädikativ bzw. adverbial gebrauchtes Adjektiv

POS-Tag: ADJD

@okay with $p=ADJD

Korpus: Referenz- und Zeitungsk

Start: 1900 **Ende:** 2016

Textklassen: ☑ Belletristik ☐ Wissenschaft ☐ Gebrauchsliteratur ☐ Zeitung

Anzeige: ○ KWIC ◉ voll ○ maximal

Sortierung: zufällig

Anzahl Treffer pro Seite: 50

1–50 von 118 Treffern (173 insgesamt) Treffer exportieren

⇤ -10 -5 ← 1 2 3 → +5 +10 ⇥

1: Goosen, Frank: Liegen lernen, Frankfurt am Main: Eichborn AG 2000, S. 111
Sie sagte **okay**, und wir gingen hoch.

2: Regener, Sven: Herr Lehmann, Frankfurt am Main: Eichborn AG 2006 [2001], S. 46
»Das sind Spiegeleier mit Speck und Bratkartoffeln, das ist doch **okay**.«

3: Krausser, Helmut: Eros, Köln: DuMont 2006, S. 308
Ist das **okay** für dich?«

4: Duckers, Tanja: Spielzone, Berlin: Aufbau-Taschenbuch-Verl. 2002 [1999], S. 179
»Okay, **okay**«, murmelt Laura.

5: Goosen, Frank: Liegen lernen, Frankfurt am Main: Eichborn AG 2000, S. 36
Auch Dylan war **okay**, schon wegen »With God on our side«, wo er gegen die Deutschen wetterte.

DWDS

PTKANT	Antwortpartikel	ja, nein, danke, bitte

Korpusbelege (Referenz- und Zeitungskorpora (aggregiert))

@okay with $p=PTKANT

Korpus: Referenz- und Zeitungsk

Start: 1900 **Ende:** 2016

Textklassen: ☑ Belletristik ☐ Wissenschaft ☐ Gebrauchsliteratur ☐ Zeitung

Anzeige: ○ KWIC ◉ voll ○ maximal

Sortierung: zufällig

Anzahl Treffer pro Seite: 50

Es wurden keine Treffer in diesem Korpus gefunden.

DWDS

@okay

Korpus: Referenz- und Zeitungsk

Start: 1900 **Ende:** 2016

Textklassen: ☑ Belletristik ☐ Wissenschaft ☐ Gebrauchsliteratur ☐ Zeitung

Anzeige: ○ KWIC ◉ voll ○ maximal

Sortierung: zufällig

Anzahl Treffer pro Seite: 50

1–50 von 118 Treffern (173 insgesamt) Treffer exportieren

⇤ -10 -5 ← 1 2 3 → +5 +10 ⇥

1: Goosen, Frank: Liegen lernen, Frankfurt am Main: Eichborn AG 2000, S. 130
Ich fragte, was das Theater mache, und sie sagte, das sei auch **okay**.

2: Schulze, Ingo: Simple Storys, Berlin: Berlin-Verl. 1998, S. 259
»Sag ihnen morgen früh, daß alles **okay** ist, ja?

3: Goosen, Frank: Liegen lernen, Frankfurt am Main: Eichborn AG 2000, S. 111
Sie sagte **okay**, und wir gingen hoch.

4: Merian, Svende: Der Tod des Märchenprinzen, Hamburg: Buntbuch-Verl. 1980 [1980], S. 111
Ich habe nach dem Dienstag abend, wo du mir das so beiläufig gesagt hast und nachher sehr hartnäckig darauf bestanden hast, daß das ganz unwesentlich für eine Beziehung sei, erstmal gedacht, **okay**, er hat Recht, wichtig ist, wie wir unsere Beziehung gestalten.

5: Noll, Ingrid: Ladylike, Zürich: Diogenes 2006, S. 283
»Ist es **okay**, wenn wir uns eine Odenwälder Wildrahmsuppe mit Waldpilzen bestellen?«

Grenzen: Semantische Blindheit (vgl. Storrer 2011)

Die großen deutschen Referenzkorpora sind bislang nicht semantisch annotiert. Man kann zwar nach (annotierten) Formeinheiten, aber nicht nach Wortbedeutungen suchen.

Wer an semantischen Einheiten interessiert ist, muss unter Umständen viel Zeit für die intellektuelle Nachbereitung der Ergebnisse investieren.

Auch die statistischen Auswertungen operieren auf formalen und nicht auf semantischen Einheiten.

Stollen[1] Stollen[2]

DWDS

Stollen, der

Grammatik Substantiv (Maskulinum) · Genitiv Singular: **Stollens** · Nominativ Plural: **Stollen**

Aussprache 🔊

Worttrennung Stol-len (computergeneriert)

Wortbildung mit ›Stollen‹ als Erstglied: ↗Stollenbau ... 3 weitere · mit ›Stollen‹ als Letztglied: ↗Bruchstollen ... 8 weitere

Bedeutungen | Etymologie | Thesaurus | Typische Verbindungen | Verwendungsbeispiele

Bedeutungen

eWDG, 1976

I. **Bergmannssprache** am Berghang meist leicht ansteigender, ins Gebirge getriebener, tunnelartiger Gang

II. Zapfen am Hufeisen der Pferde oder an den Schuhen der Fußballspieler, der ein Ausgleiten verhindern soll

III. **Literaturwissenschaft** eine der beiden gleichgebauten Strophen im Lied des Meistergesanges

DWDS

Anmelden

Stollen

Stollen[1] Stollen[2]

Stollen, der

Grammatik Substantiv (Maskulinum) · Genitiv Singular: **Stollens** · Nominativ Plural: **Stollen**

Worttrennung Stol-len (computergeneriert)

Wortbildung mit ›Stollen‹ als Letztglied: Christstollen ... 2 weitere

Bedeutung | Etymologie | Thesaurus | Typische Verbindungen | Verwendungsbeispiele

Bedeutung

eWDG, 1976

(besonders) süddeutsch, (besonders) österreichisch, (besonders) schweizerisch zu Weihnachten gebackener länglich-ovaler Hefekuchen von fester Beschaffenheit und mit vielen Zutaten

DWDS

DWDS-Wortprofil

Typische Verbindungen

computergeneriert

Bergleute

Bergwerks Einsturz

Fußballschuhe

Salzbergwerks Schächte

Sohle

Standseilbahn

angeschnitten

aufgeschlitzt

backen

eingelagert

eingestürzt

eingestürzten

gebackenen

gebohrt

geflutet

genehmigten

getriebenen

graben

kilometerlangen

stillgelegten

unterirdischen

verfüllt

verlaufenden

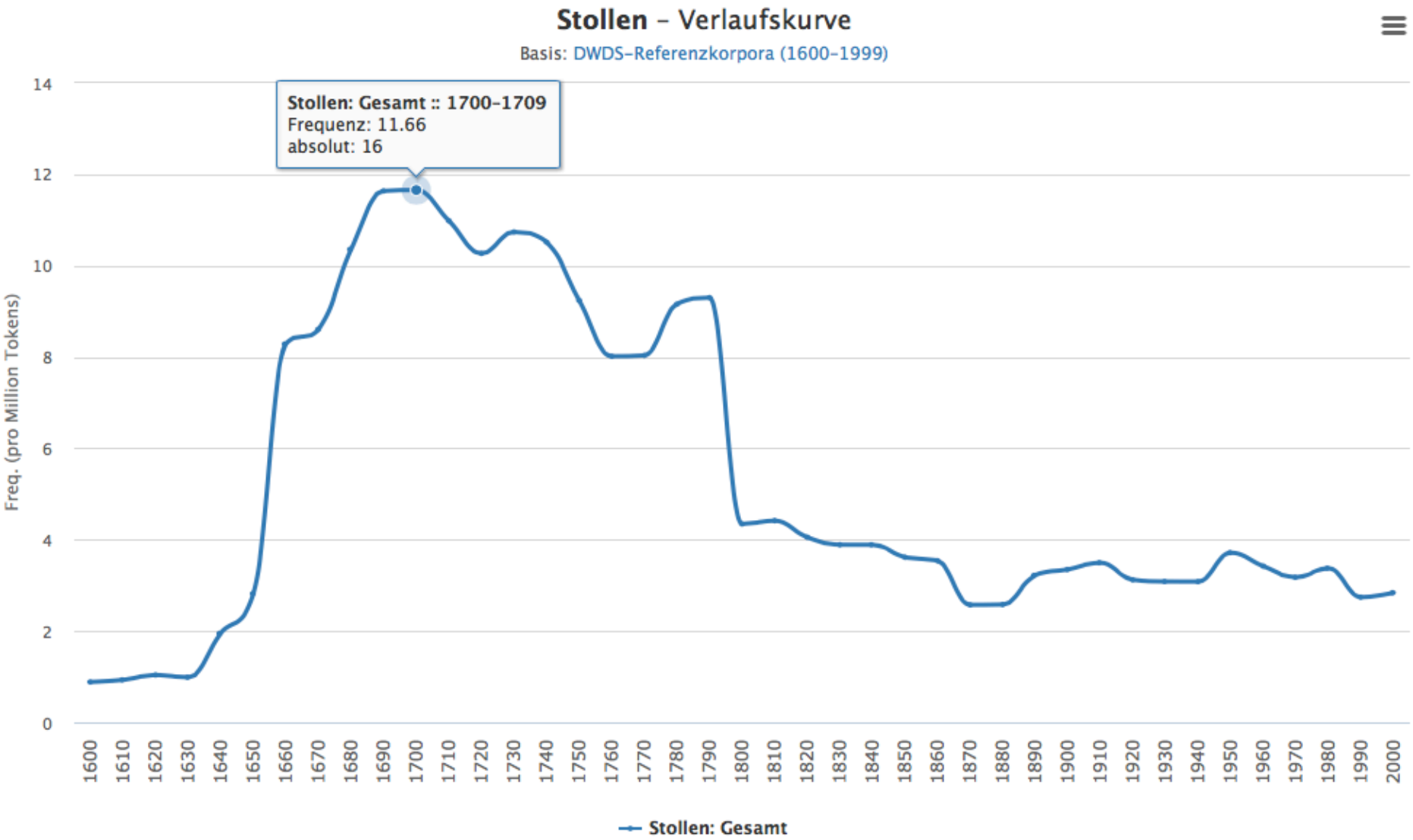
verschüttet

verschütteten

vorgetrieben

vorgetriebenen

zugeschüttet



Hinweis: Klicken Sie in der Legende auf einzelne Textklassen, um diese aus- bzw. einzublenden.

Grenzen und Herausforderungen

In großen deutschen Textkorpora kann man **sprachliche Zeichen** (= Verbindungen von Form und Inhalt) nicht direkt suchen oder auswerten.

Auch die statistischen Werkzeuge operieren über Wortformen und nicht über sprachlichen Zeichen. Dies muss man bei der Interpretation der Ergebnisse berücksichtigen.

Es empfiehlt sich, die Qualität der Annotationen, die Qualität der Metadaten und die Qualität der Digitalisierung von älteren Texten zumindest in Stichproben zu prüfen.

Fallbeispiel Google n-gram Viewer

Google books Ngram Viewer

Graph these comma-separated phrases: Troll ☑ case-insensitive

between 1800 and 2000 from the corpus German (2009) with smoothing of 1 Search lots of books

Search in Google Books:

1800 - 1858 | 1859 - 1948 | 1949 - 1956 | 1957 - 1990 | 1991 - 2000 | troll | German

Rezension schreiben

Sämmtliche Werke. - Erlangen, Heyder 1826-
von Martin Luther

"troll" Suche

Über dieses Buch

Meine Bücher

Mein Verlauf

Bücher bei Google Play

Allgemeine Nutzungsbedingungen

Sämmtliche Werke. - Erlangen, Heyder 1826- - Seite 53

books.google.com/books?id=MzRXAAAAcAAJ

Martin Luther - 1826 - Vollansicht - Mehr Ausgaben

Diefee T'röller nun heißt auch „ein Geilt derWahrheit": denn er tröllet nicht wie die Weiß da 'ein Befland bei illn fondern fein **Troll** währet ewige lil!)- und kann niemand betrugen. "4 Uber da llößec fich's wieder; denn das Gewilfen fpricht: Du ...

‹ Zurück Weiter › - Alle anzeigen

trübniß ist, da ist der heilige Geist, der Tröster, nicht daheim. Dieser Tröster nun heißt auch „ein Geist der Wahrheit": denn er tröstet nicht wie die Welt, da kein Bestand bei ist, sondern sein Trost währet ewiglich, und kann niemand betrügen.

Aber da stößet sich's wieder; denn das Gewissen spricht: Du sagest mir wohl von einem Trost; aber ich fühle ihn nicht, ja das Widerspiel fühle ich, daß die Welt Freud' und Trost hat, da dagegen die Christen sich leiden müssen. Johannes der Täufer muß seinen Kopf hergeben; Herodes und seine Hure panketieren dieweil mit einander, und haben einen guten Muth. Mit uns gehet's auch also; die Welt gönnet uns nicht einen Bissen Brods, und läßt sich jedermann dünken, was er einem Christen Uebels thue, das sey wohl gethan. Dagegen Papst, Cardinäle, Bischöfe, und alles,

Desiderate zur Diskussion

In der Lehre: Corpus literacy aufbauen

Digitale Sprachressourcen und insbesondere Korpora lassen sich sehr gut in der forschungsorientierten Lehre einsetzen.

Neben technischen Fertigkeiten muss auch **Methodenwissen** vermittelt und eine **Bewusstheit für die Möglichkeiten und Grenzen aktueller Korpustechnologie** geschaffen werden.

Corpus Literacy = "the ability to use the technology of corpus linguistics to investigate language and enhance the language development of students” (Heather & Helt 2012: 417)

Anbieter von Sprachressourcen halten dafür schon Informationen bereit.

grammis
Grammatisches Informationssystem

https://grammis.ids-mannheim.de/korpusgrammatik/4652

Korpusgestützte Grammatik

- **Grundlagen**
  - Datenbasis für Untersuchungen zur grammatischen Variabilität im Standarddeutschen
  - Statistische Methoden der Korpusgrammatik
  - **Verlässlichkeit grammatischer Annotationen**
    - Anlass und Rahmen der Annotationsprüfung
    - Annotationsüberprüfung aus grammatischer Sicht
    - Fazit der Tagger-Evaluation
- Pilotstudien
- Ressourcen

Verlässlichkeit grammatischer Annotationen

Untersuchungen zu Varianz und Variation grammatischer Strukturen, die auf sehr großen Korpora basieren, sind ohne maschinelle Unterstützung von Taggern, die automatisiert grammatische Klassifikationen des gesamten Wortmaterials vornehmen, kaum zu denken. Laut Studien zum Wortarten-Tagging (engl. *POS tagging*) für deutsche Texte genügen die besten Tagger bereits hohen Ansprüchen. Von 93- bis 98-prozentiger Verlässlichkeit ist die Rede. Zu bedenken ist aber, dass das Gros deutscher Wörter einigen wenigen Klassen (Wortarten) zugeordnet werden kann, die vergleichsweise eindeutig zu fassen sind. Erweisen sich die Tagger in diesem Bereich als leistungsfähig, kann also die sehr gute Durchschnittsverlässlichkeit manche Schwäche bei Wortklassen mit wenig Elementen verdecken.

Die Verlässlichkeit und Brauchbarkeit der Tagger sollen daher exemplarisch im Hinblick auf Wortformen *meine/Meine*, die Präpositionen im Allgemeinen und die Wortformen *bar, eingangs, fernab, halber, jenseits, laut* überprüft werden. Auf dem Prüfstein stehen die Werkzeuge TreeTagger und Connexor Machinese Phrase Tagger, mit denen die Korpora des Instituts für Deutsche Sprache annotiert sind. Kritisch betrachtet werden diese Tagger u. a. im Hinblick auf falsche Zuordnungen für ein gesuchtes

Welche Probleme kann es mit der Wortverlaufskurve geben?

Mangelnde Ausgewogenheit der Korpora: Die Korpusgrößen sind weder zeitlich noch über die Textsorten hinweg gleichverteilt. Das DWDS bemüht sich um urheberrechtliche Vereinbarungen, um eine größere Ausgewogenheit der Korpora zu erreichen.

Absolute Häufigkeiten: Mit aggregierten Korpora lassen sich Wortverlaufskurven von hochfrequenten Wörtern einigermaßen verlässlich darstellen. Anders ist dies bei mittel- und niedrigfrequenten Wörtern. Hier können Ausreißer in Einzelwerken dazu führen, dass die Wortverläufe verzerrt dargestellt werden. Auch hier empfiehlt sich der Wechsel zur Ansicht „Rohfrequenzen".

https://www.dwds.de/d/plot

Desiderate zur Diskussion

In der Lehre: Corpus literacy aufbauen

Digitale Sprachressourcen und insbesondere Korpora lassen sich sehr gut in der forschungsorientierten Lehre einsetzen.

Neben technischen Fertigkeiten muss auch Methodenwissen vermittelt und eine Bewusstheit für die Möglichkeiten und Grenzen aktueller Korpustechnologie geschaffen werden.

Corpus Literacy = "the ability to use the technology of corpus linguistics to investigate language and enhance the language development of students” (Heather & Helt 2012: 417)

Anbieter von Korpusressourcen bieten hierfür schon Informationen an.

Es wären wünschenswert, didaktische Materialien für verschiedene Zielgruppen als Open Educational Resources zur Verfügung zu stellen.

Desiderate zur Diskussion

Bei Anbietern und Nutzern:

„Kultur des Teilens“: Ressourcen, die in kleineren Projekten aufgebaut werden, für andere verfügbar machen.

Gute Beispiele: https://fussballlinguistik.de/korpora/ (Meier 2017) oder https://db.mocoda2.de/ (Beißwenger et al. 2019)

Nachhaltigkeit: Projektbezogene Ressourcen in das Angebot großer Plattformen integrieren.

Dialog zwischen den Entwicklungsteams für Plattformen und den NutzerInnen fördern: Workshops; Befragungen, digitale Angebote zur Diskussion und Rückmeldung.

Desiderate zur Diskussion

Für die Weiterentwicklung der Korpusplattformen

Verbesserung der linguistischen Basisannotation, insbesondere auch in Korpora mit einem hohem Anteil von nicht standardgerechten Schreibungen.

Aufbau multimodaler Korpora, insbesondere auch Berücksichtigung von multimodalen Ressourcen in genuin digitalen Genres (Websites, mobile Kommunikation, Social Media).

Mobile Nutzung / Apps ?

Erwähnte Ressourcen und Literatur

DEREKO: Deutsches Referenzkorpus am Leibniz-Institut für deutsche Sprache Mannheim. http://www1.ids-mannheim.de/kl/projekte/korpora/ ; Recherche über COSMAS II: http://www.ids-mannheim.de/cosmas2/

DWDS: DWDS Digitales Wörterbuch der deutschen Sprache. Berlin-Brandenburgische Akademie der Wissenschaften. http://www.dwds.de/.

FOLK: Forschungs- und Lehrkorpus gesprochenes Deutsch. Leibniz-Institut für deutsche Sprache Mannheim. http://agd.ids-mannheim.de/folk.shtml.

GRAMMIS: Grammatisches Informationssystem am Leibniz-Institut für deutsche Sprache Mannheim. https://grammis.ids-mannheim.de

Beißwenger, Michael; Imo, Wolfgang; Fladrich, Marcel; Ziegler, Evelyn (2019): *https://www.mocoda2.de: a database and web-based editing environment for collecting and refining a corpus of mobile messaging interactions.* In: European Journal of Applied Linguistics 7 (2).

Heather, Julian/Helt, Marie, 2012: *Evaluating corpus literacy training for pre-service language teachers: Six case studies.* In: Journal of Technology and Teacher Education 20/4, 415-440.

Herzberg, Laura; Storrer, Angelika. (2019) *OKAY across genres, modes and languages: A corpus-based study on German and French.* Cahiers du Laboratoire de Recherche sur le Language CLRL, 8., 148-176

Meier, Simon (2017): *Korpora zur Fußballlinguistik – eine mehrsprachige Forschungsressource zur Sprache der Fußballberichterstattung.* In: Zeitschrift für germanistische Linguistik 45(2), S. 345–349.

Storrer, Angelika (2011): Korpusgestützte Sprachanalyse in Lexikographie und Phraseologie. In: Karlfried Knapp u.a. (Hrsg.): Angewandte Linguistik. Ein Lehrbuch. 3. Auflage. Tübingen: Francke Verlag. 216-239