Originalveröffentlichung in: Deppermann, Arnulf/Fandrych, Christian/Kupietz, Marc/Schmidt, Thomas (Hrsg.): Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial. – Berlin/Boston: de Gruyter, 2023. S. 181-198. (Jahrbuch / Leibniz-Institut für Deutsche Sprache 2022)
DOI: *https://doi.org/10.1515/9783111085708*

Carolin Odebrecht/Malte Belz (Berlin)

# Akustisches Signal, Mehrebenenannotation und Aufgabendesign: flexible Korpusarchitektur als Voraussetzung für die Wiederverwendung gesprochener Korpora

## Zur /e:/-Aussprache polnischer Deutschlerner/-innen

**Abstract:** Die erfolgreiche Wiederverwendung gesprochener Korpora muss fachspezifischen Evaluationskritierien genügen und erfordert daher eine flexible Korpusarchitektur, die durch multirepräsentationale (Verfügbarkeit eines akustischen Signals und einer Transliteration) und multisituationale Daten (Variabilität von Situationen bzw. Aufgaben) gekennzeichnet ist. Diese Kriterien werden in einer Fallstudie zur /e:/-Diphthongisierung polnischer Deutschlerner/-innen angewendet und diskutiert. Die Fallstudie repliziert die Ergebnisse der /e:/-Diphthongisierung bei Bildbenennungen von Nimz (2016). Vor der Wiederverwendung werden weitere fachspezifische Evaluationskriterien überprüft, wie Multisituationalität, Aufnahmequalitäten, Erweiterbarkeit, vorhandene Metadaten und vorhandene Dokumentation. Nach der Replikationsstudie werden die Herausforderungen für eine Umsetzung der Wiederverwendung bezüglich Datenmanagement, Workflows und Data Literacy in Forschungs- und Lehrkontexten diskutiert.

## 1 Forschungsfragen

Die Wiederverwendung von Daten ist ein zunehmend integraler Forschungsbestandteil, um einerseits Forschungsergebnisse nachvollziehen, reproduzieren und replizieren zu können und um andererseits wissenschaftliche Zeit- und Kostenressourcen effizient einzusetzen. Die Wiederverwendung ist nur möglich, wenn erstens Daten vorhanden sind und zweitens diese zur intendierten Forschungsfrage passen beziehungsweise daraufhin evaluiert werden können. Jede Wiederverwendung erfordert die Auseinandersetzung mit dem Design und der Modellierung der Ursprungsdaten. Eine flexible Architektur der Ursprungsdaten ist daher wichtig, da erst damit ermöglicht wird, neue Aspekte und Modellierungen einzubeziehen, welche eine neue Forschungsfrage an die Ursprungsdaten richtet. In diesem Beitrag machen wir deutlich, dass zur erfolgreichen Wiederverwendung

https://doi.org/10.1515/9783111085708-009

im Bereich der korpusbasierten Phonetik drei Dinge notwendig sind, die wir zusammen als flexible Korpusarchitektur definieren: die komplette Verfügbarkeit eines akustischen Signals, die Transliteration und Annotation mithilfe einer Mehrebenenannotation, und ein Korpusdesign, welches mindestens zwei verschiedene Aufgaben für die Sprecher/-innen enthält. Dieser Beitrag exemplifiziert diese Notwendigkeiten anhand einer phonetischen Forschungsfrage zum Fremdsprachenakzent als Fallstudie und diskutiert parallel dazu die einhergehenden Fragestellungen an das Datenmanagement.

## 1.1 Phonetische Forschungsfrage: Realisierung von /e:/ in der Aussprache polnischer Lerner/-innen

Bei polnischen Lerner/-innen des Deutschen wurde anekdotisch beobachtet, dass sie den zielsprachlichen langen obermittelhohen vorderen ungerundeten Vokal /e:/ phonetisch als Diphthong [ei] realisieren (Hirschfeld 1998). Nimz (2016) erbrachte erstmals auch akustische Evidenz für dieses Phänomen mittels einer Bildbenennungsstudie. Von 18 polnischen Lerner/-innen des Deutschen des Niveaus B2/C1 nach dem Gemeinsamen Europäischen Referenzrahmen wurden folgende acht Stimuluswörter für /e:/ elizitiert: *Fehler*, *Lehrer*, *Zehn*, *Mehl*, *Nebel*, *geben*, *Weg*, *Keks* (Nimz 2016, S. 130). Die Vokalformanten F1 und F2 wurden jeweils an zwei Punkten der Trajektorie gemessen, nämlich an der 25%-Position und an der 75%-Position. Die akustische Analyse zeigte deutliche Bewegungen von einer niedrigen zu einer höheren und vordereren Position im Vokaltrapez, allerdings nicht so hoch wie das [i]. Nimz schlägt abschließend die Repräsentation der /e:/-Diphthongisierung als [ɛe] vor. Abbildung 1 zeigt eine solche diphthongische Realisierung beispielhaft im von uns herangezogenen Korpus WroDiaCo (Wrocław Dialogue Corpus), welches in Kapitel 2 eingeführt wird. Das Beispiel macht die mehrfache Repräsentation der Daten deutlich, da es im oberen Bereich Oszillogramm und Sonagramm als Derivate aus dem akustischen Signal und im unteren Bereich die Mehrebenenannotation (Transliteration und Phone) enthält.

Wir möchten die Ergebnisse von Nimz (2016) nun korpusbasiert in spontansprachlichen Dialogen ohne konkrete, vorher festgelegte Stimuli replizieren (wir wissen noch nicht, ob und in welcher Anzahl im Korpus Wörter mit /e:/ vorhanden sind) und verwenden dazu ein Korpus polnischer Deutschlerner/-innen, welches in Kapitel 2 vorgestellt wird.

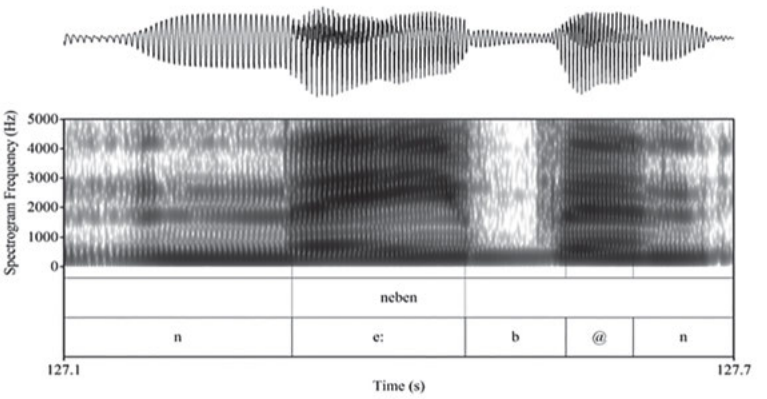


**Abb. 1:** Mit dem akustischen Signal alignierter und diphthongisch realisierter Vokal /e:/ im Wort *neben* in WroDiaCo v.2 (diapix_a_a1f_ch1, mit korrigierter Alignierung)

## 1.2 Voraussetzungen für die Wiederverwendung

Um die phonetische Forschungsfrage mit vorhandenen Daten zu beantworten, müssen zunächst die Voraussetzungen dafür geprüft werden. Gerade die Wiederverwendung gesprochener Daten stellt aufgrund ihrer mehrfachen Repräsentation (Audiodaten zusammen mit Textdaten) hohe Anforderung an das Datenmanagement, mit folgenden beispielhaften Fragen: Wie muss die Datenarchitektur eines Korpus konzipiert und erstellt werden, damit es für weitere, ursprünglich nicht intendierte Forschungsfragen verwendet werden kann? Welche technischen und intellektuellen Zugangsvoraussetzungen müssen erfüllt sein? Welche Kriterien können zur Evaluation der vorhandenen Daten herangezogen werden?

Datenarchitekturen emergieren aus einem Zusammenspiel von Datenmodell, Datenaufbereitung und -realisierungen mithilfe verschiedener Software- und IT-Services. Eine Mehrebenenarchitektur für verschiedene Konzepte von Annotationen hat sich als *best practice* in der Korpuslinguistik etabliert (Zeldes 2019) und wird zunehmend auch für gesprochene Korpora eingesetzt (z. B. BeDiaCo; Belz et al. 2021; BeMeCo; Zöllner et al. 2021; RUEG; Wiese et al. 2019; GECO; Schweitzer/Lewandowski 2013). Wir stellen in Kapitel 2 WroDiaCo vor, das diesen Mehrebenenansatz mit der Integration des akustischen Signals verbindet, somit für die phonetische Analyse nutzbar macht und für die wissenschaftliche Wiederverwendung zur Verfügung steht.

Wichtige Leit- und Richtlinien[1] sind beispielsweise die *FAIR Guiding Principles* (Wilkinson et al. 2016), die fach- und datenunabhängig vier Qualitätsmerkmale definieren: *Findability*, *Accessibility*, *Interoperability*, und *Reusability* (Auffindbarkeit, Zugänglichkeit, Interoperabilität und Wiederverwendbarkeit). Diese verwenden wir als Evaluationskriterien für die Rahmenbedingungen der Forschungsdatenwiederverwendung. Die Herausforderung im Bereich des Datenmanagements ist es, diese abstrakten Kriterien im fachlichen Kontext anzuwenden und umzusetzen (vgl. Kap. 3). In einem größeren Rahmen stellen zudem die Richtlinien der guten wissenschaftlichen Praxis Bezugspunkte dar, die auch auf datenbasierte Forschungsvorhaben Anwendung finden. Beide Richtlinien verknüpfen wir mit unserer Fallstudie (siehe Kap. 3 und Kap. 6.2).

## 2 Gesprochenes Korpus

Um die Forschungsfrage zu beantworten, benötigen wir Daten polnischer Deutschlerner/-innen. Da wir zur Analyse der Vokalqualität die Formanten berechnen müssen, benötigen wir a) das akustische Signal zum vollständigen Download und b) eine mit dem Signal alignierte textuelle Repräsentation des Gesprochenen, worauf weitere Annotationen aufgebaut sein können. Zusätzlich ist es sinnvoll, wenn die Daten c) in einem mehrdimensionalen Aufgabendesign vorliegen, was eine größere Variabilität für die Sprachverwendung und eine bessere Vergleichsbasis für sprachliches Verhalten schafft. Wir verwenden das Wrocław Dialogue Corpus (WroDiaCo; Wesolek et al. 2021). Es enthält akustische Aufnahmen und Annotationen spontansprachlicher freier und aufgabenbasierter Dialoge von 16 Sprecher/-innen mit Polnisch als Erst- und Deutsch als Zweitsprache, wobei wir in unserer Fallstudie nur ein Subkorpus von acht Sprecher/-innen verwenden (siehe Kap. 4). Der freie Dialog enthält als Gesprächsanlass die Frage nach dem letzten Wochenende; der aufgabenbasierte Dialog nutzt Diapixe (Baker/

**1** Es gibt eine Reihe von Richtlinien und Regelungen für das Datenmanagement mit verschiedenen Schwerpunkten auf Ebene der Länder/des Bundes (z. B. DSGVO), der Hochschulen (z. B. HU-Forschungsdatenpolicy www.cms.hu-berlin.de/de/dl/dataman/hu-fdt-policy/view) und der Förderer (z. B. die Empfehlungen des DFG-Fachkollegium 104 „Sprachwissenschaften“ 2019 www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf oder auch von EU Horizon 2020 https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm). Diese variieren hinsichtlich adressierter Themenfelder, Granularität der Empfehlungen und Verbindlichkeit. Dies gilt im Übrigen auch immer mehr für Software (vgl. Chue Hong et al. 2021).

Hazan 2011), eine Suchbildaufgabe, bei der die Sprecher/-innen auf leicht unterschiedlichen Bildern zehn Unterschiede finden müssen, ohne dass sie diese gegenseitig einsehen können. Die Sprecher/-innen beherrschen Deutsch auf dem Niveau B1–C1 nach dem Gemeinsamen Europäischen Referenzrahmen für Sprachen (Hilpisch 2012), was anhand abgelegter Prüfungen und/oder durch Selbsteinschätzung der Sprecher/-innen ermittelt wurde (für weitere Metadaten bspw. zu Auslandsaufenthalten siehe Wesolek 2021). Für diese Studie wird ein Subkorpus von acht Sprecher/-innen gebildet, da eine Annotationsebene (*KORphon*) in Version 2 nicht für alle Sprecher/-innen annotiert ist. Von diesen sind vier auf B1- und vier auf C1-Niveau. Das Subkorpus umfasst 2,4 Stunden Aufnahmen (11.987 Token). Das Korpus wurde im Rahmen der Masterarbeit von Sarah Wesolek am Institut für deutsche Sprache und Linguistik der Humboldt-Universität zu Berlin erstellt und für die wissenschaftliche Forschung verfügbar gemacht.

# 3 Datenmanagement

Forschungsdatenmanagement ist ein expliziter Prozess, der die Erstellung und Verwaltung von Forschungsmaterialien umfasst, um deren Nutzung zu ermöglichen (übersetzt aus Whyte/Rans 2022). Die Erstellung von Daten setzt ein konkretes Design (Zusammenstellung und Komposition) und eine adäquate Konzeption von Annotation (Festlegung der Ebenen und Werte) voraus und ist damit eine Modellierungsaufgabe, die damit eine fachliche Dimension besitzt. Dies wird typischerweise als Datenlebenszyklen im Datenmanagement beschrieben (vgl. Dierkes 2021). Wir nehmen diesen Grundgedanken auf und erweitern ihn mit dem Begriff des Workflows, der die Interaktion von Daten, Tools und methodischen Zielsetzungen stärker in den Fokus rückt.

Datenmanagement verstehen wir folglich als Teil der wissenschaftlichen Methode und somit auch als Bestandteil der guten wissenschaftlichen Praxis (vgl. Deutsche Forschungsgemeinschaft 2019). Die DFG versteht darunter unter anderem den Einsatz von fachlich angemessenen Methoden und Dokumentationen, die Berücksichtigung des Forschungskontexts und zudem die Reliabilität, Integrität, Überprüfbarkeit und Nachvollziehbarkeit von Forschungsergebnissen. Diese Aspekte müssen demnach auch im digitalen Paradigma – in unserem Fall der korpusbasierten Phonetik – berücksichtigt werden. Als zusätzliche Evaluationskriterien wenden wir die FAIR-Prinzipien an (siehe Kap. 1) und erweitern diese um die Prinzipien Referenzierbarkeit und Zitierbarkeit. Beide Richtlinien sind Voraussetzung für die nach den Regeln der guten wissenschaftlichen Praxis aus-

zuweisende Wiederverwendung der Daten, welche das erklärte Ziel in vielen Fachbereichen und von vielen weiteren Förderern ist.

Um den fachlichen Kontext umfassend mit zu berücksichtigen, stellen wir ebenfalls fachspezifische Evaluationskriterien an die Daten, die die DFG-Richtlinien fachlich kontextualisieren: Die Daten müssen multirepräsentational,[2] multisituational[3] und erweiterbar[4] hinsichtlich ihrer Annotationen sein, in hoher Aufnahmequalität vorliegen, anonymisiert und pseudonymisiert sein, umfangreiche Sprecher- und Korpusmetadaten enthalten und forschungsorientiert dokumentiert sein. Somit evaluieren wir die Daten aus zwei Perspektiven: datenmanagementfokussiert und fachlich. Die Evaluation mit Blick auf das Datenmanagement ist dabei die Voraussetzung für die fachliche Evaluation.

Tabelle 1 zeigt eine – nicht notwendigerweise die einzige – mögliche Zuordnung unserer Evaluationskriterien zur phonetischen Domäne und dem gesprochenen Korpus. Ausgehend von den sehr allgemeinen FAIR-Prinzipien können die ausgewählten DFG-Leitlinien grob zugeordnet werden. In der Spalte fachliche Domäne versuchen wir konkretere Zuordnungen zu Evaluationskriterien unserer Fallstudie. In der Spalte WroDiaCo zeigen wir die Realisierungen dieser drei Ebenen.

WroDiaCo ist auf dem Medienrepositorium der Humboldt-Universität zu Berlin[5] langfristig gespeichert, mit fachlich spezifizierten Metadaten ausgewiesen und für wissenschaftliche Zwecke zugänglich. Zu jeder Version der Daten ist auch ein umfangreiches Korpushandbuch (für Version 2 Wesolek/Belz 2021) publiziert. Damit ist das Korpus *findable* sowie technisch und intellektuell *accessible*, da relevante Informationen zu Verantwortlichen, Zugangsregelungen und eine umfassende Korpusdokumentation für Nutzer/-innen zur Verfügung stehen. Das akustische Signal und die Annotationen liegen als TextGrid-Daten vor, ein Format des Tools Praat (Boersma/Weenink 2022). Dieses Format kann in eine Emu-Datenbank (Winkelmann et al. 2017) konvertiert und in R mit *emuR* (Winkelmann et al.

**2** Unterschiedliche Repräsentationen, in denen die Daten vorliegen – mindestens müssen das Audiosignal und eine erste signalalignierte Annotation (z. B. Transliteration) vorliegen.

**3** Multisituationale Daten enthalten unterschiedliche Situationen oder Aufgaben (z. B. eine freie Kommunikation und eine aufgabenbasierte Kommunikation) und sind für die Erfassung der Variabilität innerhalb einer Domäne sowie für die Registerforschung von besonderer Bedeutung.

**4** Die Erweiterbarkeit gilt neben den Annotationen im Prinzip auch für das Sprachdatenmaterial und setzt eine gute Dokumentation voraus. Dann können die Sprachdaten mit neuen Daten im gleichen Korpusdesign oder mit einer begründeten Veränderung des Korpusdesigns erhoben werden.

**5** https://medien.hu-berlin.de/phon. Das Medienrepositorium ist ein Basisdienst des Computer- und Medienservice der Humboldt-Universität zu Berlin.

2020) analysiert werden. Damit ist eine hohe Interoperabilität zu weiteren Workflows gewährleistet. Einen Beleg hierfür stellt die erste Wiederverwendung dieser Daten in Belz/Odebrecht (2022) dar.

**Tab. 1:** Assoziationen zwischen FAIR- und DFG-Evaluationskriterien bezogen auf die fachliche Domäne und deren Realisierung in WroDiaCo. Wir verwenden in dieser Übersicht einen Auszug der DFG-Richtlinien, die besonders relevant für unser Beispiel erscheinen

| FAIR | DFG | Fachliche Domäne | Korpus (WroDiaCo) |
|---|---|---|---|
| Findability | Referenzierbarkeit, Zitierbarkeit | Fachlich spezifizierte Korpusmetadaten | Medienrepositorium |
| Accessibility | Dokumentation | Sprechermetadaten, forschungsorientierte Dokumentation | Medienrepositorium, Korpushandbuch |
| Interoperability | Methoden, Forschungskontext | Multirepräsentational, Aufnahmequalität | Audiosignal, offene Formate (Praat, Emu-DB), Mehrebenenannotation |
| Reusability | Überprüfbarkeit, Nachvollziehbarkeit | Anonymisiert, pseudonymisiert, erweiterbar | Wissenschaftlicher Zugang |

Die weitere fachliche Datenevaluation überprüft die Passgenauigkeit und Verarbeitungsmöglichkeit (Workflow) für die eigene Forschungsfrage (siehe Kap. 1). Das Korpus enthält spontansprachliche Dialoge polnischer Deutschlerner/-innen. Diese sind zwar relativ kurz (ca. 4 min) und enthalten teils lange Pausen, dies ist aber kein großer Nachteil – ebenfalls kein Nachteil ist, dass das Korpus ursprünglich für eine andere Forschungsfrage erhoben wurde (siehe Wesolek/Belz 2021). Das Korpus enthält aufgrund der automatischen Alignierung von akustischem Signal und Transliteration stellenweise Alignierungsungenauigkeiten, welche bei besonderem Interesse an einer bestimmten Stelle bzw. deren phonetischen Segmenten dann dort manuell korrigiert werden können. Von Vorteil ist insbesondere die Möglichkeit zur Wiederverwendung sowohl der akustischen als auch der Annotations- und Metadaten für wissenschaftliche Dritte und das Korpus- bzw. Aufgabendesign, welches zwei verschiedene Register abdeckt (eine freie Konversation und eine Suchbildaufgabe). Als mögliches Thema der freien Konversation wurde beispielhaft das vergangene oder kommende Wochenende genannt, was von den Versuchspersonen aufgegriffen wurde (für Details siehe die Dokumentation in Wesolek/Belz 2021). Insgesamt sind die Daten für die Beantwortung dieser Forschungsfrage gut geeignet.

Der Workflow geht von der aktuellen Korpusversion v.2 aus, die auf der GitLab-Instanz der Humboldt-Universität zu Berlin liegt. Die Daten (die identisch mit Version 2 sind, die auf dem Medienrepositorium veröffentlich ist) können so direkt mithilfe von Git bearbeitet und neue Annotationen oder Korrekturen in das Korpus integriert werden. Nach Abschluss der Datenanalyse wurde das Korpus in einer neuen Version v2.1 veröffentlicht (Wesolek/Belz 2022; Wesolek et al. 2022).

# 4 Korpusstudie

Aus WroDiaCo v.2 verwenden wir die *ORTword*- und die *KORphon*-Ebene. *ORTword* enthält eine orthografische Transliteration. *KORphon* enthält Phone, die automatisiert aus der Transliteration heraus mittels eines BAS Web Service (Kisler et al. 2017) erstellt wurden und die wahrscheinlichste Aussprache (ausgehend von der kanonischen Aussprache) darstellen. Dies ist eine erste Annäherung an die realisierte phonetische Form und gibt nicht immer die tatsächliche Realisierung wieder, bspw. werden Diphthongisierungen von /e:/ nicht gesondert repräsentiert. Die Segmentgrenzen dieser Ebene wurden für /e:/ manuell korrigiert. Die *KORphon*-Ebene ist in Version 2 nicht für alle Sprecher/-innen annotiert, weswegen wir nur ein Subkorpus von acht Sprecher/-innen untersuchen können. Die akustischen Daten und die beiden Ebenen werden mit R (R Core Team 2022) in eine Emu-Datenbank konvertiert. Anschließend suchen wir zunächst nach allen Vorkommen von /e:/ auf *KORphon*. Abbildung 2 enthält den Suchbefehl und die 556 auf *ORTword* enthaltenen Token.

```
> e <- query(wrodiaco, "[KORphon == e: ^ #ORTword =~ .*]")
> table(e$labels)

    achtzehn    Aktivität    angesehen     Apotheke    ausgesehen     aussehen            b           bl
           1            1            1           11            1            1            2            2
  Blaubeeren            d          dem          den         denen          der     dreizehn            e
           2            3           61            7            1           85            2            2
          eh         eher          ehm     entweder       erste       ersten           fe      Fernseh
           1            1            1            1            5            3            1            1
    fünfzehn            g           ge       gehen       gehend         gehn         geht        gehts
           1            2           10            9            1            2            2            2
     gelesen      gesehen      gewesen          heh         hehe       hehehe          hey         Idee
           1            4            1            3            7            2            1            4
italienische         jede        jeden       jemand      Kollege     Kollegen        leer        leere
           1            2            4            1            1            1            6            1
     nächste      nächsten        neben          neh       nehmen           ok         okay      Problem
           5            1           16            1            1            5          121            3
         sch      sechsehn          seh         sehe        sehen         sehn         sehr     siebsehn
           1            1            9           36            8            1           28            1
     sleeves       später       stehen        steht       stehts      T-shirt     Tableten      versteh
           1            3            1           25            2            2            1            1
    verstehe     vierzehn       vorher            w          weg          Weg        wegen         zehn
           2            1            1            3            1            1            1            4
      zehten        zehtn
           1            1
```

**Abb. 2:** Alle Token auf der diplomatischen Transliterationsebene in WroDiaCo, für die auf *KORphon* automatisiert ein realisiertes /e:/ geschätzt wurde

Um die Studie zu der von Nimz (2016) vergleichbar zu halten, werden keine Wortabbrüche verwendet. Token, bei denen die automatisierte Transkription auf *KORphon* fälschlicherweise aus der Orthografie abgeleitet wird, werden ausgeschlossen (bspw. *sleeves*). Zusätzlich soll /e:/ nur in phonetischen Kontexten stehen, in denen es zielsprachlich monophthongisch verwendet wird, von keinen Sekundärdiphthongen gefolgt wird (bspw. in *sehr*) und keinen Hiat mit der Folgesilbe aufweist.

Dabei bilden wir aus ähnlichen Token Gruppen (Typen), um bei einem Wortvergleich (jetzt: Typenvergleich) mit weniger Kategorien arbeiten zu können. Die Token *achtzehn, dreizehn, fünfzehn, sechsehn* (sic), *siebsehn* (sic), *vierzehn,* und *zehn* werden im Type *NUM-zehn*, die Token *dem* und *den* im Typ *dem|den*, *jede* und *jeden* im Type *JED, steht* und *stehts* im Typ *steht* zusammengefasst. Tabelle 2 fasst die Anzahl der Typen für die beiden Aufgabensituationen zusammen. Im Vergleich zu Nimz (2016) werden also, geleitet von den Vorkommen der tatsächlichen Token in WroDiaCo, keine Nomen ausgewertet. Hingegen können wir nun aufgrund der Beschaffenheit der Korpusdaten die Replikation des Diphthongisierungseffektes auf andere Wortarten ausdehnen, wie definite Artikel, Präpositionen, Zahladverbien, Indefinitpronomina und Verben. Obwohl ein direkter Vergleich mit der Nimz-Studie *prima facie* für *NUM-zehn* (hier) mit *zehn* (bei Nimz) und *steht* (hier) mit *geben* (bei Nimz) möglich scheint, wird sich zeigen, dass eine Diphthongisierung von diesen beiden Fällen nur für das flektierte Verb *steht* repliziert wird (vgl. die Diskussion in Kap. 6).

**Tab. 2:** Analysierte Typen je Register in WroDiaCo

| Typ | Diapix | Freier Dialog |
|---|---|---|
| dem\|den | 65 | 3 |
| *JED* | 0 | 6 |
| *neben* | 16 | 0 |
| *NUM-zehn* | 9 | 1 |
| *steht* | 27 | 0 |
| Token | 117 | 10 |

Im nächsten Schritt wurden die Formanten für das komplette Korpus berechnet und der Datenbank hinzugefügt. Für alle 127 Token in Tabelle 2 wurden sowohl die zeitliche Alignierung des Signals (orientiert am Oszillo- und Sonagramm) als auch der erste und zweite Vokalformant (in den berechneten Trajektorien, die über das Sonagramm gelegt wurden) manuell in der Emu-Datenbank auf der Ebene *KORphon* korrigiert.

Die Formanten werden vokal-extrinsisch, formant-intrinsisch und sprecherintrinsisch normalisiert (Lobanov 1971) und anschließend zurück auf die Hertz-Skala skaliert (Thomas/Kendall 2007). Die Überlappung zweier Vokalverteilungen zu einem bestimmten Zeitpunkt wird mithilfe des Pillai-Wertes gemessen (Nycz/Hall-Lew 2013). Dieser beruht auf einer multifaktoriellen Varianzanalyse. Je höher der Pillai-Wert, desto größer ist die gemeinsame Distanz von F1 und F2 zwischen zwei Vokalen.

Nach Abschluss der Analyse wurde die veränderte Ebene *KORphon* aus der Emu-Datenbank mithilfe von *emuR* zurück in das TextGrid-Format exportiert. Hierbei entstehen aufgrund unterschiedlicher zeitlicher Repräsentationen kleine Ungenauigkeiten unterhalb einer Millisekunde. Damit die nicht-korrigierten Intervallgrenzen mit den restlichen Daten weiterhin übereinstimmen, werden diese mit einem selbsterstellten Skript[6] unter Verwendung von *rPraat* v.1.3.2.1 (Boril/Skarnitzl 2016) anhand der Ebene *KORphon* aus v.2 auf ihre ursprüngliche Position gesetzt. Unsere neuen Annotationen für diese Studie werden als WroDiaCo Version 2.1 im Medienrepositorium der Humboldt-Universität zu Berlin zur Verfügung gestellt.

# 5 Ergebnisse

Wir konnten den Diphthongisierungseffekt im Vokal /e/ für die Typen *neben* und *steht* replizieren. Abbildung 3 zeigt die Formantverläufe je Typ. Um Koartikulation auszuschließen, interpretieren wir die Verläufe hier nur zwischen 25% und 75% der normalisierten Dauer. Die Daten lassen sich für F1 visuell in zwei Gruppen teilen, nämlich fallende Verläufe (*NUM-zehn, steht, neben*) und steigende Verläufe (*dem*|*den, JED*). Fallende Verläufe kennzeichnen eine Bewegung von einer geschlosseneren hin zu einer offeneren Position im Vokaltrapez. Für F2 zeigen *NUM-zehn, steht* und *neben* einen steigenden Verlauf, *dem*|*den* und *JED* einen fallenden, was eine Bewegung von einer hintereren zu einer vordereren Position im Vokaltrapez kennzeichnet.

Tabelle 3 enthält die Pillai-Werte für die Differenz zwischen der 25%- und der 75%-Position der Formanten F1 und F2 im Vokal /e/ in den fünf Typen und den zwei Sprachstufen. Für *NUM-zehn*, *dem*|*den* und *JED* sind die Pillai-Werte nicht besonders unterschiedlich und auch nicht signifikant, was bedeutet, dass mit diesen Daten keine Diphthongisierung belegt werden kann. Hingegen unterscheiden sich die Pillai-Werte für *neben* und *steht* signifikant.

**6** Skript *move-boundaries-a-little.R*, verfügbar unter https://hu.berlin/mb-skripte (Stand: 3.5.2022).

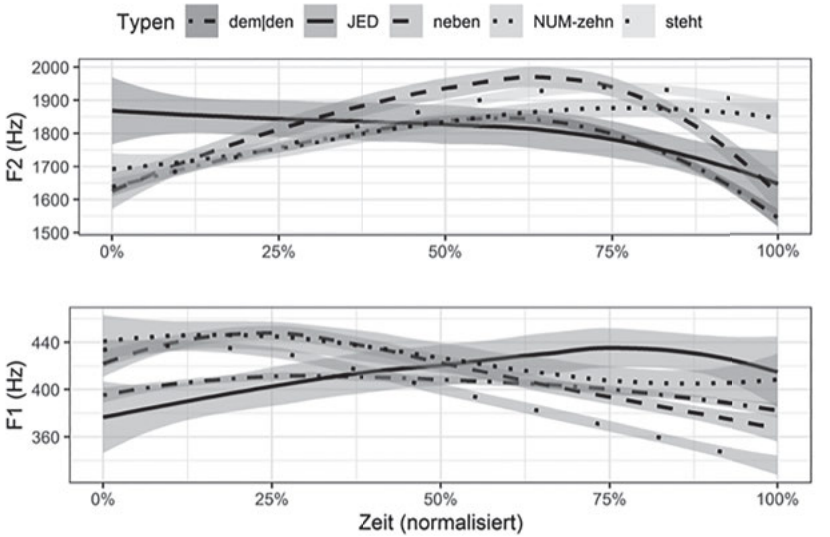


**Abb. 3:** Normalisierte Formanttrajektorien für den /e/-Vokal in den fünf Typen für F1 (unten) und F2 (oben)

**Tab. 3:** P-Werte für die gemeinsame Überlappung von F1 und F2 zu 25% im Vergleich zu 75% (gemessen mit dem Distanzmaß Pillai – je größer Pillai, desto größer der Abstand zwischen 25% und 75%) der Vokaltrajektorie je Typ und Niveau

| Typ | Niveau | Pillai | p |
|---|---|---|---|
| dem\|den | B1 | 0,11 | 0,1 |
| dem\|den | C1 | 0,04 | 0,2 |
| JED | B1 | 0,93 | 0,3 |
| JED | C1 | 0,31 | 0,4 |
| neben | B1 | 0,74 | < 0,001 |
| neben | C1 | 0,64 | < 0,001 |
| NUM-zehn | B1 | 0,2 | 0,6 |
| NUM-zehn | C1 | 0,32 | 0,2 |
| steht | B1 | 0,84 | < 0,001 |
| steht | C1 | 0,56 | < 0,001 |

In den Vokaltrapezen in Abbildung 4 ist deutlich zu sehen, dass der Vokal in *neben* und *steht* als Diphthong realisiert wird, mit einer Bewegung von einer hinteren-offeneren zu einer vorderen-geschlosseneren Position. Vom Niveau B1 hin

zu C1 nimmt Pillai für beide Typen ab, was ein Indikator dafür ist, dass die Sprecher/-innen mit höherem Sprachniveau zumindest akustisch und im Mittel eine monophthongischere Aussprache des Vokals /e:/ erreichen.

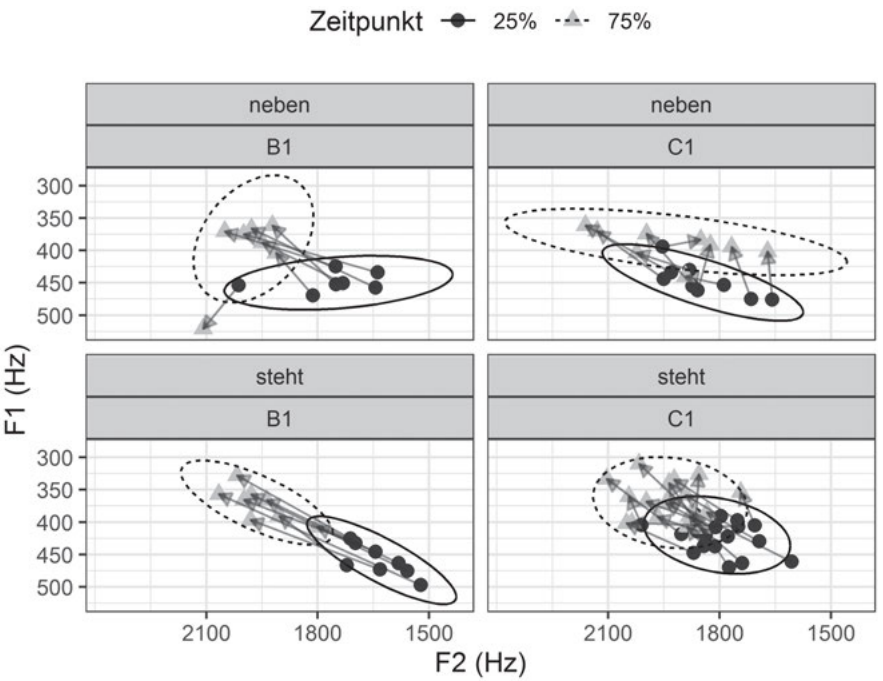


**Abb. 4:** Positionen zu 25% und 75% der Formanttrajektorien für den Vokal /e/ je Typ, Pfeile deuten den Pfad vom ersten zum zweiten Zeitpunkt an

# 6 Diskussion

## 6.1 Phonetische Diskussion

Insgesamt konnten wir für die Präposition *neben* und das flektierte Verb *steht* in der jeweils betonten Silbe zeigen, dass Diphthongisierung bei der Aussprache von /e:/ von den polnischen Sprecher/-innen des Deutschen produziert wird. Zudem produzieren die Sprecher/-innen mit höherem Sprachniveau eine akustisch weniger ausgeprägte Diphthongisierung. Die Ergebnisse von Nimz (2016) konnten also korpusbasiert repliziert und für einen zusätzlichen Typ belegt werden.

Dass nur für zwei Typen eine Diphthongisierung festzustellen ist, könnte an den Betonungsmustern und möglicher Koartikulation bei den restlichen Typen liegen. Definite Artikel sind meistens unbetont und der Vokal daher kürzer als in Typen, in denen der /e/-Vokal den Hauptakzent erhält. Ein ähnlicher Grund könnte für den Typ *NUM-zehn* vorliegen, da hier die Hauptbetonung auf der ersten Silben liegt; die unbetonte Silbe *zehn* wird daher auch eher kürzer sein und weniger Zeit zur Diphthongisierung lassen als wenn das Wort *zehn* wie bei Nimz (2016) als Einzelwort ausgesprochen wird.

*JED* trägt zwar den Hauptakzent auf der ersten Silbe, zu der /e/ gehört, hat jedoch im linken Kontext einen palatalen Approximanten /j/. Dieser hat besonders ausgeprägte Formanten (tiefer F1, hoher F2), so dass zumindest für F2 keine ausladende Bewegung in der Trajektorie zu erwarten ist (von einer hohen Position in /j/ zu einer hohen Position in /e/).

Ob die weniger ausgeprägte Diphthongisierung der Sprecher/-innen mit C1-Niveau von deutschen Muttersprachler/-innen auch so perzipiert wird, also tatsächlich ein weniger stark ausgeprägter Fremdsprachenakzent wahrgenommen wird, muss in einem Perzeptionsexperiment untersucht werden. Nicht unerwähnt darf bleiben, dass für diesen Beitrag nur acht Sprecher/-innen untersucht wurden (bzw. vier je Sprachniveau). Dennoch ist der Effekt für die beiden Typen *neben* und *steht* in den Abbildungen gut zu erkennen. Für eine größere Datengrundlage und die mögliche Einbeziehung weiterer Typen können in zukünftigen Arbeiten die acht weiteren Sprecher/-innen aus WroDiaCo untersucht werden.

## 6.2 Methodische Diskussion

Typischerweise würden fachliche Beiträge mit dem letzten Kapitel 6.1 enden. Meist können die Interaktion von Datenmanagement und Forschung sowie die wesentlichen und hohen Anforderungen für die in Kapitel 5 gezeigten und in Kapitel 6.1 diskutierten Ergebnisse nicht ausreichend dargelegt und diskutiert werden, weil das Datenmanagement nicht immer als Teil des Forschungsbeitrags verstanden wird und noch immer große Herausforderungen mit häufig offenen Fragen im Fachbereich stellt. Daher möchten wir in diesem Beitrag drei Schwerpunkte nachgelagert diskutieren: Datenmanagement als Teil des wissenschaftlichen Arbeitens, Datenworkflows und *Data Literacy* (im Sinne der Einheit von Forschung und Lehre).

**Datenmanagement**: Wir verstehen die Wiederverwendung von Daten als Normalfall. Datenmanagement ist folglich ein wesentlicher Teil des Forschungsprozesses und ein Baustein der Forschungsmethode, was mindestens vier Konsequen-

zen für das wissenschaftliche Arbeiten hat: 1) Im Bereich des Forschungsdesigns wird es mit der Evaluation auf Ebene des Datenmanagements und auf fachlicher Ebene in den ersten Schritten des Forschungsvorhabens eingebunden. 2) Die Beachtung der Richt- und Leitlinien von Förderern und Fachgemeinschaften setzt auch in der Planung und Durchführung einzelner Forschungsvorhaben an.[7] 3) Die enge Verbindung und gegenseitige Befruchtung von Forschung und Lehre sind auch im methodischen Bereich des Datenmanagements elementar und können sowohl die *Data Literacy* der Studierenden als auch die Forschung selbst fördern. 4) Die Re-Integration der eigenen Daten in bestehende Datenpublikationen (z. B. mittels einer neuen Korpusversion) oder die erneute eigenständige Publikation ist Teil des wissenschaftlichen Workflows und ist durch die vorangegangenen Regelungen des Datenmanagements bedingt. Die Umsetzung dieser vier Konsequenzen lässt Forschung zwangsläufig kollaborativ und interdisziplinär werden. Eine letzte Konsequenz ergibt sich damit automatisch: 5) Datenmanagement ist wie die Forschung selbst lebendig in dem Sinne, dass jede Regelung, jede Modellierungsfrage und jeder Workflow neu verhandelt, evaluiert und umgesetzt werden muss, wobei dabei *best practices* helfen.

**Workflow**: Was bedeutet das für unsere Fallstudie und das Korpus WroDiaCo? Mit den Kriterien des Datenmanagements können vorhandene Möglichkeiten der Verarbeitung und Analyse betrachtet werden. Abbildung 5 zeigt den gesamten Datenworkflow im Kontext der Forschungsvorhaben und der Datenpublikation. In Bezug auf die Zeit ist der Workflow linear, dennoch sind die Interaktionen in alle Richtungen denkbar und es muss nicht zwangsläufig ein einziger Bearbeitungs- und Publikationsstrang entstehen.[8] Jede Bearbeitung und Analyse der Daten erfolgt zwangsläufig mit der Hilfe von Tools und Services, womit Softwaremanagement in unserem Fall auch Teil des Datenmanagements ist. Dafür muss im Übrigen auch die Software den FAIR-Kriterien genügen (Chue Hong et al. 2021). Dieser Aspekt limitiert – wie bei Daten auch – die Möglichkeiten der Wiederverwendung in Bezug auf mögliche Workflows. Die eingesetzte Software und Pakete sind frei verfügbar und flexibel einsetzbar.

**7** Diese werden sonst typischerweise bei der Konzeption und Einreichung von Drittmittelprojekten konsultiert und beachtet. Wir zeigen mit unserem Beitrag, dass auch korpusbasierte Forschung mit umfassendem Datenmanagement ohne diesen Hintergrund möglich und auch erforderlich ist.

**8** Dieser Strang würde sich beispielsweise aufteilen, wenn vorhandene Daten nachgenutzt, aber nicht wieder in die bestehenden Infrastrukturen eingespeist werden können oder sollen.

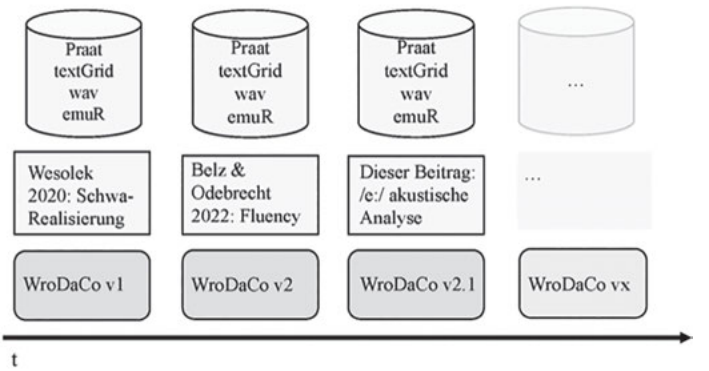


**Abb. 5:** Datenmanagement, Workflow und Forschungsbeiträge von und für WroDiaCo. Die Korpusversionen sind jeweils das Ergebnis der angegebenen Studien pro Spalte. Jede weitere Studie basiert auf der vorherigen Korpusversion und passt das Korpus für die eigene Forschungsfrage an

**Data Literacy**: *Data Literacy* meint die Fähigkeit, planvoll mit Daten umgehen zu können (Heidrich et al. 2018).[9] Dieser dritte Aspekt ist gerade für die vorliegende Fallstudie besonders relevant, weil sie zeigt, dass studentische Arbeiten sehr wertvoll für die weitere Forschung sein können, wenn das Datenmanagement, der Workflow und die fachliche Kontextualisierung ebenso integrale Bestandteile der forschungsorientierten Lehre und Abschlussarbeiten sind. Der übergeordnete Begriff *Data Literacy* fasst dies für alle Bereiche und professionelle Ebenen zusammen. Dabei verfolgen wir einen ganzheitlichen Ansatz für forschungsorientierte Lehre, die die Ebene des Datenmanagements direkt mit der des fachlichen Wissens verbindet: Die Konzeption, der Aufbau und die Architektur von Daten werden als genuine Bestandteile der Lehrinhalte und der Abschlussarbeiten verstanden. Uns gilt das Projekt *Register in Diachronic German Science*[10] als Vorbild, das seit 2011 mit Seminaren in BA- und MA-Studiengängen an der Humboldt-Universität zu Berlin und fortlaufenden Datenpublikationen (z. B. Lüdeling et al. 2022) nach den oben genannten Kriterien die *Data Literacy* in verschiedenen Fachbereichen fördert. Dieser Ansatz konnte erfolgreich zum Beispiel mit WroDiaCo als Teil einer studentischen Arbeit im Bereich Phonetik adaptiert werden.

Mit WroDiaCo und diesem Beitrag zeigen wir, dass Datenmanagement in allen Forschungs- und Lehrkontexten umgesetzt werden kann. Mit der Verwendung von IT-Services der Humboldt-Universität zu Berlin sowie vorhandenen

**9** Dies umfasst alle möglichen Schritte in einem Datenlebenszyklus beziehungsweise Workflow.
**10** Unter der Leitung von Prof. Dr. Anke Lüdeling: https://hu-berlin.de/ridges.

offenen Tools zur Datenbearbeitung und -analyse der Fachcommunity ist es im laufenden Forschungsbetrieb möglich, hohe Standards zu setzen und diese auch vermitteln zu können. Hierbei zeigt sich die enorme Wichtigkeit einer von Seiten der Forschungsinstitution gut aufgestellten Forschungsdatenserviceinfrastruktur, die durch die Implementierung offener Tools auf der eigenen Domäne den Forschenden ermöglicht, sensible Daten zu schützen und gleichzeitig kollaborativ zu arbeiten. Nicht zuletzt kann auf diese Weise die FAIR entstandene Forschungsarbeit von Studierenden im Sinne guter wissenschaftlicher Praxis in der Forschung und Lehre referenziert und gewürdigt werden.

## Literatur

Baker, Rachel/Hazan, Valerie (2011): DiapixUK. Task materials for the elicitation of multiple spontaneous speech dialogs. In: Behavior Research Methods 43, 3, S. 761–770. DOI: 10.3758/s13428-011-0075-y.

Belz, Malte/Odebrecht, Carolin (2022): Abschnittsweise Analyse sprachlicher Flüssigkeit in der Lernersprache. Das Ganze ist weniger informativ als seine Teile. In: Zeitschrift für germanistische Linguistik 50, 1, S. 131–158. DOI: 10.1515/zgl-2022-2051.

Belz, Malte/Mooshammer, Christine/Zöllner, Alina/Adam, Lea-Sophie (2021): Berlin Dialogue Corpus (BeDiaCo). Version 2. Berlin: Humboldt-Universität zu Berlin (Medien-Repositorium). https://rs.cms.hu-berlin.de/phon (Stand: 23.8.2022).

Boersma, Paul/Weenink, David (2022): Praat. Doing phonetics by computer. Version 6.2. www.praat.org/ (Stand: 17.8.2022).

Bořil, Tomáš/Skarnitzl, Radek (2016): Tools rPraat and mPraat. Interfacing phonetic analyses with signal processing. In: Sojka, Petr/Horák, Aleš/Kopecek, Ivan/Pala, Karel (Hg.): Text, speech and dialogue. 19th International Conference on Text, Speech and Dialogue (TSD 2016), Brno, Czech Republic, September 12–16. (= Lecture Notes in Computer Science 9924). Cham: Springer, S. 367–374.

Chue Hong, Neil P./Katz, Daniel S./Barker, Michelle/Lamprecht, Anna-Lena/Martinez, Carlos/Psomopoulos, Fotis E./Harrow, Jen/Castro, Leyla J./Gruenpeter, Morane/Martinez, Paula A./Honeyman, Tom/Struck, Alexander/Lee, Allen/Loewe, Axel/van Werkhove, Ben/Jones, Catherine/Garijo, Daniel/Plomp, Esther/Genova, Francoise/Shanahan, Hugh/Leng, Joanna/Hellström, Maggie/Sandström, Malin/Sinha, Manodeep/Kuzak, Mateusz/Herterich, Patricia/Zhang, Qian/Islam, Sharif/Sansone, Susanna-Assunta/Pollard, Tom/Atmojo, Udayan-to Dwi/Williams, Alan/Czerniak, Andreas/Niehues, Anna/Fouilloux, Anne Claire/Desinghu, Bala/Goble, Carole/Richard, Céline/Gray, Charles/Erdmann, Chris/Nüst, Daniel/Tartarini, Daniele/Ranguelova, Elena/Anzt, Hartwig/Todorov, Ilian/McNally, James/Moldon, Javier/Burnett, Jessica/Garrido-Sánchez, Julián/Belhajjame, Khalid/Sesink, Laurents/Hwang, Lorraine/Tovani-Palone, Marcos R./Wilkinson, Mark D./Servillat, Mathieu/Liffers, Matthias/Fox, Merc/Miljković, Nadica/Lynch, Nick/Martinez Lavanchy, Paula/Gesing, Sandra/Stevens, Sarah/Martinez Cuesta, Sergio/Peroni, Silvio/Soiland-Reyes, Stian/Bakker, Tom/Rabemanantsoa, Tovo/Sochat, Vanessa/Yehudi, Yo (2021): FAIR principles for research software (FAIR4RS Principles).

Deutsche Forschungsgemeinschaft (2019): Guidelines for safeguarding good research practice. Code of conduct. Bonn: Deutsche Forschungsgemeinschaft. DOI: 10.5281/ZENODO.3923601.

DFG-Fachkollegium 104 „Sprachwissenschaften“ (2019): Handreichung: Empfehlungen zu datentechnischen Standards und Tools bei der Erhebung von Sprachkorpora. (= Empfehlungen des DFG-Fachkollegiums 104 “Sprachwissenschaften“). www.dfg.de/download/pdf/foerderung/grundlagen_dfg_foerderung/informationen_fachwissenschaften/geisteswissenschaften/standards_sprachkorpora.pdf (Stand: 23.8.2022).

Dierkes, Jens (2021): Planung, Beschreibung und Dokumentation von Forschungsdaten. In: Putnings, Markus/Neuroth, Heike/Neumann, Janna (Hg.): Praxishandbuch Forschungsdatenmanagement. Berlin u. a.: De Gruyter Saur, S. 303–326.

Heidrich, Jens/Bauer, Pascal/Krupka, Daniel (2018): Future skills. Ansätze zur Vermittlung von Data Literacy in der Hochschulbildung. (= Hochschulforum Digitalisierung Arbeitspapier 37). https://gi.de/fileadmin/GI/Hauptseite/Aktuelles/Aktionen/Data_Literacy/HFD_AP37_DALI_Studie_2018-09.pdf (Stand: 17.8.2022).

Hilpisch, Kai (2012): Gemeinsamer Europäischer Referenzrahmen für Sprachen. Der GER im Überblick. Hamburg: Diplomica.

Hirschfeld, Ursula (1998): Einige Schwerpunkte für die Arbeit an der Aussprache bei polnischen Deutschlernenden. In: Glottodidactica XXVI, S. 113–122. https://repozytorium.amu.edu.pl/bitstream/10593/2614/1/09%20Ursula%20HIRSCHFELD%2C%20Einige%20Schwerpunkte%20fur%20die%20Arbeit%20an%20der%20Aussprache%20bei%20polnischen%20Deutschlernenden.pdf (Stand: 17.8.2022).

Kisler, Thomas/Reichel, Uwe/Schiel, Florian (2017): Multilingual processing of speech via web services. In: Computer Speech & Language 45, S. 326–347. DOI: 10.1016/j.csl.2017.01.005.

Lobanov, Boris M. (1971): Classification of Russian vowels spoken by different speakers. In: The Journal of the Acoustical Society of America 49, 2B, S. 606–608.

Lüdeling, Anke/Odebrecht, Carolin/Krause, Thomas/Schnelle, Gohar/Fischer, Catharina (2022): RIDGES Herbology (Version 9.0). Berlin: Humboldt-Universität.

Nimz, Katharina (2016): Sound perception and production in a foreign language. Does orthography matter? (= Potsdam Cognitive Science Series 9). Potsdam: Universitätsverlag Potsdam. http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-88794 (Stand: 17.8.2022).

Nycz, Jennifer/Hall-Lew, Lauren (2013): Best practices in measuring vowel merger. In: Proceedings of Meetings on Acoustics 20, 1 (167th Meeting of the Acoustical Society of America). Providence, RI: Acoustical Society of America, S. 1–19.

R Core Team (2022): R: A language and environment for statistical computing. Wien: R Foundation for Statistical Computing.

Schweitzer, Antje/Lewandowski, Natalie (2013): Convergence of articulation rate in spontaneous speech. In: Proceedings of Interspeech 2013, S. 525–529.

Thomas, Erik R./Kendall, Tyler (2007): NORM. The vowel normalization and plotting suite. http://lingtools.uoregon.edu/norm/ (Stand: 17.8.2022).

Wesolek, Sarah/Belz, Malte (2021): Dokumentation und Annotationsrichtlinien für das Korpus WroDiaCo Version 2. Berlin: Humboldt-Universität zu Berlin.

Wesolek, Sarah/Belz, Malte (2022): Dokumentation und Annotationsrichtlinien für das Korpus WroDiaCo Version 2.1. Berlin: Humboldt-Universität zu Berlin.

Wesolek, Sarah/Belz, Malte/Mooshammer, Christine (2021): Wrocław Dialogue Corpus (WroDiaCo). Version 2. Berlin: Humboldt-Universität zu Berlin (Medien-Repositorium). https://rs.cms.hu-berlin.de/phon (Stand: 15.2.2021).

Wesolek, Sarah/Belz, Malte/Mooshammer, Christine (2022): Wrocław Dialogue Corpus (WroDiaCo). Version 2.1. Berlin: Humboldt-Universität zu Berlin (Medien-Repositorium). https://rs.cms.hu-berlin.de/phon (Stand: 15.2.2021).

Whyte, Angus/Rans, Jonathan (2022): Glossary: Research data management. www.dcc.ac.uk/about/digital-curation/glossary#R (Stand: 17.8.2022).

Wiese, Heike/Alexiadou, Artemis/Allen, Shanley/Bunk, Oliver/Gagarina, Natalia/Iefremenko, Kateryna/Jahns, Esther/Klotz, Martin/Krause, Thomas/Labrenz, Annika/Lüdeling, Anke/Martynova, Maria/Neuhaus, Katrin/Pashkova, Tatiana/Rizou, Vicky/Rosemarie, Tracy/Schroeder, Chris-toph/Szucsich, Luka/Tsehaye, Wintai/Zerbian, Sabine/Zuban, Yulia (2019): RUEG corpus (Version 0.2.0).

Wilkinson, Mark D./Dumontier, Michel/Aalbersberg, IJsbrand Jan/Appleton, Gabrielle/Axton, Myles/Baak, Arie/Blomberg, Niklas/Boiten Jan-Willem/da Silva Santos, Luiz Bonino/Bourne, Philip E./Bouwman, Jildau/Brookes, Anthony J./Clark, Tim/Crosas, Mercè/Dillo, Ingrid/Dumon, Olivier/Edmunds, Scott/Evelo, Chris T./Finkers, Richard/Gonzalez-Beltran, Alejandra/Gray, Alasdair J. G./Groth, Paul/Goble, Carole/Grethe, Jeffrey S./Heringa, Jaap/'t Hoen, Peter A. C./Hooft, Rob/Kuhn, Tobias/Kok, Ruben/Kok, Joost/Lusher, Scott J./Martone, Maryann E./Mons, Albert/Packer, Abel L./Persson, Bengt/Rocca-Serra, Philippe/Roos, Marco/van Schaik, Rene/Sansone, Susanna-Assunta/Schultes, Erik/Sengstag, Thierry/Slater, Ted/Strawn George/Swertz, Morris A./Thompson, Mark/van der Lei, Johan/van Mulligen, Erik/Velterop, Jan/Waagmeester, Andra/Wittenburg, Peter/Wolstencroft, Katherine/Zhao, Jun/Mons, Barend (2016): The FAIR guiding principles for scientific data management and stewardship. In: Scientific Data 3, 160018. DOI: 10.1038/sdata.2016.18.

Winkelmann, Raphael/Harrington, Jonathan/Jänsch, Klaus (2017): EMU-SDMS. Advanced speech database management and analysis in R. In: Computer Speech & Language 45, S. 392–410. DOI: 10.1016/j.csl.2017.01.002.

Winkelmann, Raphael/Jaensch, Klaus/Cassidy, Steve/Harrington, Jonathan (2020): Main package of the EMU Speech Database Management System. [R package emuR Version Version 2.1.1.].

Zeldes, Amir (2019): Multilayer corpus studies. (= Routledge Advances in Corpus Linguistics). New York City u. a.: Routledge.

Zöllner, Alina/Mooshammer, Christine/Hamann, Silke (2021): Berlin Menutask Corpus (BeMeCo). Version 1. Berlin: Humboldt-Universität zu Berlin. https://rs.cms.hu-berlin.de/phon (Stand: 17.8.2022).