Originalveröffentlichung in: Wolf-Farré, Patrick/Löff Machado, Lucas/Prediger, Angélica/Kürschner, Sebastian (Hrsg.): Deutsche und weitere germanische Sprachminderheiten in Lateinamerika. Grundlagen, Methoden, Fallstudien. – Berlin [u.a.]: Peter Lang, 2023. S. 103-147. (MinGLA – Minderheiten germanischer Sprachen in Lateinamerika / Minorías de lenguas germánicas en Latinoamérica 1)
DOI: *https://doi.org/10.3726/b19898*

Göz Kaufmann, Jan Gorisch, Thomas Schmidt

Das MEND-Korpus im Archiv für Gesprochenes Deutsch: Entstehung, Möglichkeiten, Grenzen

Abstract: For many reasons, Mennonite Low German is a language whose documentation and investigation is of great importance for linguistics. To date, most research projects that deal with this language and/or its speakers have had a relatively narrow focus, with many of the data cited being of limited relevance beyond the projects for which they were collected. In order to create a resource for a broad range of researchers, especially those working on Mennonite Low German, the dataset presented here has been transformed into a structured and searchable corpus that is accessible online. The translations of 46 English, Spanish, or Portuguese stimulus sentences into Mennonite Low German by 321 consultants form the core of the MEND-corpus (Mennonite Low German in North and South America) in the Archive for Spoken German. In addition to describing the origin of this corpus and discussing possibilities and limitations for further research, we discuss the technical structure and search possibilities of the Database for Spoken German. Among other things, this database allows for a structured search of metadata, a context-sensitive token search, and the generation of virtual corpora that can be shared with others. Moreover, thanks to its text-sound alignment, one can easily switch from a particular text section of the corpus to the corresponding audio section. Aside from the desire to equip the reader with the technical knowledge necessary to use this corpus, a further goal of this paper is to demonstrate that the corpus still offers many possibilities for future research.

Keywords: Mennonite Low German • language variation • methods of data elicitation • corpus treatment and editing • functionalities of the Database for Spoken German

1. Einleitung

Oft stehen am Anfang von Forschungsprojekten Enttäuschungen. Als ich[1] Ende 1993 von meiner ersten Feldforschungsreise zu den Mennonitenkolonien in Ciudad Cuauhtémoc, Chihuahua, Mexiko nach Austin, Texas, USA zurückkehrte,

1 Göz Kaufmann schildert in diesem Abschnitt, wie es zu der hier besprochenen Datensammlung kam. Ab dem folgenden Abschnitt zeichnen alle Autoren verantwortlich. Wir alle bedanken uns für die sehr hilfreichen Kommentare von zwei anonymen Gutachtern.

Publikationsserver des Leibniz-Instituts für Deutsche Sprache
URN: http://nbn-resolving.de/urn:nbn:de:bsz:mh39-121961

wurde ich dort „enttäuscht". Alles war perfekt gelaufen, die Kontaktaufnahme mit den Mennonit/inn/en war einfacher gewesen als gedacht, und ich hatte sogar schon erste Interviews geführt. Neben allen soziolinguistischen Einsichten war mir dabei aufgefallen, dass die Informant/inn/en ihre (finiten) Verben scheinbar wahllos über den Satz verteilten. Erst dachte ich, dass es sich wohl um ein Lernerproblem im Hüagdietschen handelte, also in der selten wirklich kommunikativ beherrschten H-Varietät, die die Mennonit/inn/en mit mir, dem neugierigen Dietschländer, zumeist verwendeten. Dann aber, als sich mein Ohr ein wenig ans Plautdietsche gewöhnt hatte, bemerkte ich, dass diese Verbstellungen wohl eher den Verhältnissen, von Regeln konnte ich (noch) nicht sprechen, des Plautdietschen geschuldet waren. Ich war elektrisiert, dachte sogar darüber nach, ob dies nicht ein viel spannenderes Thema für meine Doktorarbeit wäre als der plötzlich schnöde erscheinende Spracherhalt einer Minderheitensprache.

Mit klopfendem Herzen betrat ich am Tag nach meiner Rückkehr das Büro von Mark Louden, dem Betreuer meines Forschungsaufenthaltes an der *University of Texas*. Dieser hörte sich meinen Bericht zur soziolinguistischen Situation in Mexiko zufrieden an, um dann mit einem Satz meinen inzwischen übervollen Syntaxballon zum Platzen zu bringen. Während ich von meinen bahnbrechenden Entdeckungen berichtete, bemerkte er knapp: „Ja, ein Fall von *verb raising*."[2] „Ein was?", dachte ich und erkannte, dass andere mir zuvorgekommen waren. Ernüchtert und ein wenig beschämt ob meiner Ignoranz verließ ich Marks Büro und beschloss, dem Spracherhalt treu zu bleiben. Darauf hatte mich die Germanistik in Heidelberg ja auch wesentlich besser vorbereitet als auf komplexe syntaktische Fragen.

Ganz los ließ mich die Verbsyntax aber nicht. Nicht zuletzt auf Marks Anregung hin las ich schon bei meinem nächsten Feldforschungsaufenthalt, diesmal bei den Mennonit/inn/en in Seminole, Texas, den Gewährspersonen sechs englische Stimulussätze vor und bat sie, diese ins Plautdietsche zu übersetzen. Die

2 *Verb raising* beschreibt eine rechtsverzweigende Abfolge verbaler Elemente in satzfinalen Verbclustern, also z.B. die unterstrichene Abfolge in *Du weißt doch, dass ich morgen meiner Mutter **werde** helfen müssen*. Hier steht das finite Verb *werde* vor den beiden nicht-finiten Verben *helfen* und *müssen*. Beim *verb projection raising* können nicht-verbale Elemente in solchen rechtsverzweigenden Strukturen vorkommen, also z.B.: *Du weißt doch, dass ich morgen **werde** MEINER MUTTER helfen müssen*. In vielen kontinentalwestgermanischen Varietäten findet diese Umstellung auch bei Clustern mit zwei Verben statt. Plautdietsch gehört zu diesen Varietäten (vgl. Wurmbrand 2017 für allgemeine Informationen zum *verb (projection) raising* und Kaufmann 1997: 185–195, 2003, 2007 und 2016 für diese Phänomene im Plautdietschen).

Übersetzungen notierte ich per Hand (*on-the-spot transcription*) und behandelte zwei davon in einem Unterkapitel meiner Dissertation (vgl. Kaufmann 1997: 185–195 für die Sätze (a) und (c) in Abbildung 1). Abbildung 1 zeigt einen Ausschnitt eines Fragebogens aus Seminole:

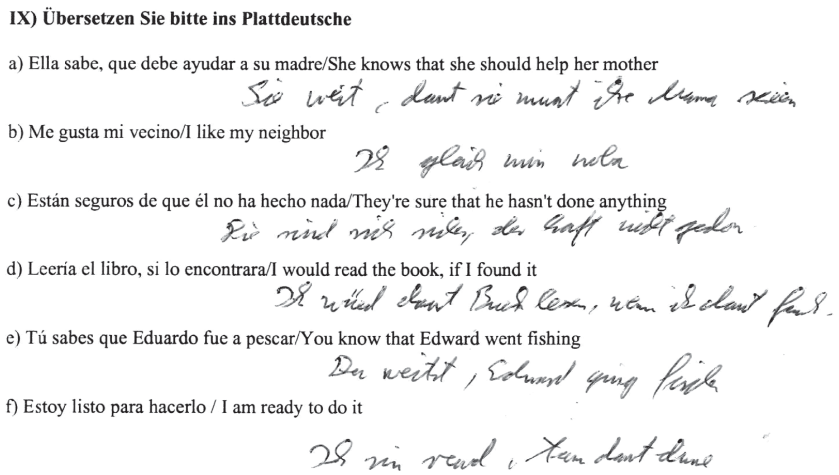


Abbildung 1: Übersetzung von sechs Sätzen durch einen texanischen Mennoniten (USA; M/56/Sp>Pl-86 %)[3]

Die Übersetzung des englischen Stimulussatzes (a) (und zugegebenerweise auch die Übersetzung meiner Handschrift) *Sie weit, daut sie muat* IHRE MAMA *sehen* ist ein Fall von *verb projection raising* (vgl. Fußnote 2), wobei *help* abweichend als *sehen* übersetzt wird. Dieser noch recht naive Versuch der Datenerhebung stellt den Keim des hier vorzustellenden Forschungsprojekts zur Verbsyntax des Plautdietschen dar. Als ich nämlich ab 1997 als DAAD-Lektor an der *Universidade*

3 Da dieser Informant nicht Teil des MEND-Korpus ist, erhält er keine Nummerierung. Angegeben sind seine Herkunft, sein Geschlecht (M = Mann; F = Frau), sein Alter in Jahren und seine Sprachkompetenz(en). Wenn Plautdietsch die dominante Sprache ist, steht hier einfach Pl. Für einen ambilingualen Sprecher aus den USA steht Pl+ E. Ist Plautdietsch nicht die dominante Sprache, wird der Grad seiner Beherrschung präzisiert. Bei dem hier vorgestellten, erst einige Jahre vorher von Mexiko nach Texas ausgewanderten Informanten bedeutet das Label *Sp>Pl-86 %*, dass Spanisch seine dominante Sprache ist und dass er 12 aus 14 möglichen Kompetenzpunkten im Plautdietschen erreicht (vgl. Fußnote 6 für Details dieser Bepunktungsmethode). Weitere Abkürzungen für Sprachen sind Port (Portugiesisch) und HD (Hüagdietsch).

Federal do Rio Grande do Sul in Porto Alegre tätig war und bei einem ersten Besuch der dort lebenden Mennonit/inn/en bemerkte, dass auch diese die Verben scheinbar wild im Satz herumwirbelten, gab es kein Halten mehr. Bei einem weiteren Aufenthalt in Austin im April 1999 entstanden die 46 englischen Stimulussätze.[4] Diese wendete ich sofort in Seminole, Texas an, wobei nun alle Übersetzungen aufgenommen[5] und später digitalisiert wurden (zuerst von mir und dann später noch einmal vom Archiv für Gesprochenes Deutsch (AGD)).

Zurück in Brasilien wurden die Stimulussätze ins Portugiesische übertragen und kamen im Oktober 1999 zum ersten Mal in Colônia Nova, Rio Grande do Sul zum Einsatz. Mark Loudens Reaktion, als ich im März 2001 erste Analyseergebnisse auf einer Sprachinselkonferenz an der *University of Kansas at Lawrence* vorstellte (vgl. Kaufmann 2003), „enttäuschte" mich nicht mehr. Nach meinem Vortrag sagte er beim Kaffee: „*This is a goldmine!*"

Nach vier weiteren Feldforschungsreisen zwischen April 2001 und Oktober 2002 hatten insgesamt 321 Mennonit/inn/en die 46 Stimuli übersetzt (103 in Mexiko; 81 in Paraguay (42 aus der Kolonie Menno; 37 aus der Kolonie Fernheim; 2 aus der Kolonie Neuland); 73 in den USA (darunter 6 Informant/inn/en, die zum Aufnahmezeitpunkt weniger als 5 Jahre in Seminole gelebt hatten); 56 in Brasilien; 8 in Santa Cruz de la Sierra, Bolivien). Nach vielen Jahren der Arbeit mit dieser Datensammlung, die sich in unzähligen Veröffentlichungen (vgl. die Literaturliste) und in meiner Habilitationsschrift (vgl. Kaufmann 2016) niederschlug, fragte ich Thomas Schmidt bei der IDS-Jahrestagung 2016, ob dieses Korpus für das AGD von Interesse sein könnte. Zu meiner großen Freude bejahte er dies sofort, und nach knapp drei Jahren der intensiven Zusammenarbeit mit ihm und mit Jan Gorisch, Danijel Lokas, Ulf Michael Stift, Philipp Breitenreicher, Matthias Kloft und Aaron Schmidt-Riese kam es am 30.11.2018 zum ersten Release. Das Kernstück des MEND-Korpus (vgl. Kaufmann 2018a) bilden die etwa 14.500 Übersetzungen der 46 Stimulussätze. Daneben beinhaltet die Datensammlung grundlegende Sozialdaten und die von den Informant/inn/en selbst evaluierten Kompetenzniveaus und Gebrauchshäufigkeiten in den relevanten Kontaktsprachen.[6]

---

4 Die englischen, spanischen und portugiesischen Stimulussätze wurden jeweils von Muttersprachler/inne/n auf ihre Korrektheit und „mündliche" Adäquatheit hin überprüft.

5 Verwendet wurde dabei ein Sony Walkman Professional D6C mit zwei an der Kleidung angebrachten Sony-Kragenmikrophonen.

6 Die Sprachkompetenz wurde anhand von zwei Fragen eruiert. Zum einen sollten die Gewährspersonen angeben, welche Sprache sie am besten, am zweitbesten, am drittbesten usw. beherrschen. Zum anderen sollten sie die Kompetenz in jeder Sprache absolut als sehr gut, gut, mittelmäßig oder schlecht evaluieren. Die höchste Punktezahl pro Frage betrug 7

Die folgende Beschreibung gliedert sich wie folgt: In Abschnitt 2 diskutieren wir die Vor- und Nachteile der verwendeten Erhebungsmethode und vergleichen das MEND-Korpus mit anderen auf Übersetzungen basierenden Korpora. Abschnitt 3 präsentiert dann exemplarisch einige Dimensionen der Variation, die entweder bereits untersucht wurden oder noch untersucht werden können. Während Abschnitt 4 die Übernahme und Aufbereitung des Korpus beschreibt, bespricht Abschnitt 5 technische Details und erklärt beispielhaft, wie man die Datensammlung mit verschiedenen Suchoptionen durchforsten kann. Abschnitt 6 bietet eine kurze Zusammenfassung und einen Ausblick.

2. Übersetzungen als Methode der Datenerhebung

Sicherlich wird sich manch/e Leser/in fragen, warum Übersetzungen als Grundlage dieser Datensammlung gewählt wurden. Der Grund hierfür war, dass insbesondere der Einfluss des Nebensatztyps auf das *verb (projection) raising* quantitativ untersucht werden sollte (vgl. Fußnote 2), und der Versuch, hierfür genügend vergleichbare Daten in freien Gesprächen zu finden, unrealistisch erschien. Der Nachteil der gewählten Methode ist natürlich, dass viele Forschende die Qualität und Validität von Übersetzungsdaten prinzipiell als geringer einschätzen als die von freien Gesprächen. Schon ein oberflächlicher Blick auf manches Forschungsprojekt, das mit freien Gesprächen arbeitet, reicht allerdings aus, um dieses Bild zu relativieren. Oft ist nämlich der/die Forschende Teil dieser Gespräche und seine/ihre Anwesenheit kann sogar dann zum Problem werden, wenn er/sie die zu untersuchende Varietät muttersprachlich beherrscht (*observer's paradox*).[7] Deppermann (2008: 25; vgl. auch Milroy / Gordon

Punkte. Wurde eine Sprache gar nicht beherrscht, gab es 0 Punkte, bei sehr guter Beherrschung der am besten beherrschten Sprache 14 Punkte (vgl. Kaufmann 1997: 135–138 für weitere Details). Beim Sprachgebrauch wurden die Informant/inn/en unter anderem nach der/den Sprache/n gefragt, die sie normalerweise mit den Großeltern, den Eltern, den Geschwistern, den mennonitischen Freund/inn/en und bei der Arbeit verwendeten. Daneben wurden sie gefragt, welche Sprache sie insgesamt am häufigsten benutzten. Für jede Einzelfrage gab es 5 Punkte, wurden zwei Sprachen gleichberechtigt genannt, gab es jeweils 2,5 Punkte. So entstand eine Skala von 0 bis 30 Punkten (vgl. Kaufmann 1997: 153–155 für weitere Details und für reale Werte einer Gewährsperson Abbildung 10).

7 Huffines (1991: 45), die in ihrem Forschungsprojekt alle Interviews selbst durchführte, schreibt: „*Although the context can easily be described as informal, all informants were aware that the conversation was being taped and that the investigator was interested in hearing Pennsylvania German and learning about what it means to be Pennsylvania German.*" Um solch einen Einfluss zu vermeiden, wurden bei den Mennonit/inn/en nur solche Gespräche aufgenommen, bei denen Göz Kaufmann nicht anwesend

2003: 49) macht in jedem Fall deutlich, dass jeglicher naive Vergleich zwischen verschiedenen Datentypen grundsätzlich fehlgeleitet ist:

> Jedes Datum hat vielmehr seine eigene Art von „Natürlichkeit", in Bezug auf die es adäquat untersucht werden kann. Statt generell „natürliche Daten" zu fordern, ist es deshalb zutreffender, wenn man verlangt, dass das Datenmaterial und die Art seiner Erhebung und Auswertung geeignet sein müssen, die Forschungsfragen in bestmöglicher Weise zu beantworten.

Trotz einer wie auch immer zu definierenden Natürlichkeit von Übersetzungsdaten bieten sie zweifelslos Antworten auf Fragen, die freie Gespräche nicht geben können. Insbesondere können anhand konstruierter Stimulussätze Phänomene in den Blick genommen werden, die in freien Gesprächen selten vorkommen. Der Einwand, dass gerade das häufige Vorkommen solcher Phänomene zur mangelnden Natürlichkeit dieses Datentyps beiträgt, ist zwar nicht von der Hand zu weisen, die Alternative zu solch einem Vorgehen wäre aber häufig, solche Phänomene überhaupt nicht (statistisch) untersuchen zu können. Übersetzungsdaten garantieren damit nicht nur das vorhersagbare und ausreichende Vorkommen des zu untersuchenden Phänomens, sondern ermöglichen es auch, einen (statistischen) Vergleich zwischen sprachlichen Kontexten und zwischen Informant/inn/en einer oder mehrerer Sprachgemeinschaften vorzunehmen. Denn selbst wenn wir ein Phänomen untersuchen, das in ausreichendem Maß in freien Gesprächen elizitiert werden kann, heißt das weder, dass dieses Phänomen jeweils in vergleichbaren Kontexten vorkommt, noch heißt es, dass alle diese Kontexte bei allen Gewährspersonen vorkommen. Dorian (1981: 118) schreibt über den diesbezüglichen Vorteil von englisch-keltischen Übersetzungen: „*[…] because of the comprehensiveness of the [translation] tests and because of the direct comparisons they made possible, the results were interesting and generally interpretable.*"

Es ist sicherlich kein Zufall, dass Übersetzungsdaten häufig in der Erforschung von Minderheitensprachen zur Anwendung kommen. Neben dem MEND-Korpus, Huffines (1991) und Dorian (1981) lassen auch Hill / Hill (1977) ihre Gewährspersonen aus dem Spanischen ins Nahuatl übersetzen. In jüngerer Zeit wurden Übersetzungen in einem Projekt zum Unserdeutsch verwendet (vgl. Maitz et al. 2016). Schließlich wurden auch im Rahmen des AGD-Korpus *Deutsch heute* (2006–2009) die Teilnehmer/innen gebeten, zehn englische Sätze

war. Dass dies anscheinend ein bemerkenswertes Vorgehen ist, kann man an Milroy / Gordon (2003: 68) ablesen: „*In some cases, researchers have taken the study of groups even further by removing the investigator from the scene.*" Bei den Mennonit/inn/en entstanden so viele Stunden an Aufnahmen wirklich freier Gespräche, die aber (noch) nicht Teil des MEND-Korpus sind.

ins Deutsche zu übersetzen. Bei all diesen Projekten besteht also ein (relativ) großer linguistischer Abstand zwischen der (Mehrheitssprache als) Stimulussprache und der Zielsprache. Je größer dieser Abstand ist, desto geringer dürfte die Gefahr einer kognitiv-strukturellen Beeinflussung (z.B. Priming) durch die Stimulussprache sein (Bangalore et al. 2016 legen diesen Schluss nahe).[8] Dies stellt einen großen Vorteil gegenüber Forschungsprojekten dar, bei denen dieser Abstand geringer ausfällt. Solche Projekte sind z.B. der *Syntactic Atlas of the Dutch Dialects* (SAND; *Een Syntactische Atlas van de Nederlandse Dialecten*; vgl. Barbiers et al. 2005), der *Syntaktische Atlas der deutschen Schweiz* (SADS; vgl. Bucheli / Glaser 2002) oder das Forschungsprojekt *Syntax hessischer Dialekte* (SyHD; vgl. Fleischer et al. 2015). Übersetzungen von zumeist schriftlich präsentierten standardniederländischen bzw. standarddeutschen Stimulussätzen stellen einen wichtigen Teil dieser Projekte dar. Da hierbei der sprachliche Abstand der Stimulussprache geringer ausfällt als bei MEND, ist ein kognitiv-struktureller Einfluss wahrscheinlicher (vgl. neben Bangalore et al. 2016 auch die kritischen projektbezogenen Kommentare von Bucheli / Glaser 2002: 44, 60 und Fleischer et al. 2015: 264).[9] Natürlich finden auch wir Einflüsse aus den Stimulussprachen, aber diese Einflüsse bleiben fast ausnahmslos auf oberflächliche (Syntax)Phänomene beschränkt (vgl. Kaufmann 2005 und die Diskussion zu (13) und zu Tabelle 1).

Großprojekte haben natürlich auch Vorteile, insbesondere die Tatsache, dass sie über weit größere Personal- und Geldressourcen verfügen. Beim Schweizer Projekt z.B. wurde der erste Fragebogen von 2.672 Gewährspersonen aus 344 Erhebungsorten ausgefüllt (vgl. Bucheli / Glaser 2002: 53). Solch eine Zahl übertrifft die Möglichkeiten eines Einzelprojektes bei weitem, selbst wenn es sich um eine große, mehrere Länder umfassende Feldstudie handelt. Aber Einzelprojekte haben auch Vorteile. Zuallererst muss hier die Tatsache genannt werden, dass es immer dieselbe Person ist, die die Daten erhebt. Dies schließt Fehler und Verzerrungen zwar nicht grundsätzlich aus, aber Verzerrungen, die auf unterschiedliche Verhaltensformen im Interview bzw. auf unterschiedliche Transkriptions- und Interpretationsweisen zurückzuführen sind (vgl. Mathussek 2016 und Milroy / Gordon 2003: 52), kommen weit seltener vor. Daneben gibt

8 Allerdings ergibt sich aus der Tatsache, dass das MEND-Korpus mit drei Stimulussprachen arbeitet, in dieser Hinsicht ein gewisses Problem, da diese Sprachen einen jeweils unterschiedlichen linguistischen Abstand zum Plautdietschen aufweisen.

9 Sowohl in Bezug auf die schriftlichen Stimuli als auch in Bezug auf den geringen linguistischen Abstand kann man hier auch den Deutschen Sprachatlas nennen, also die Übersetzungen der Wenker-Sätze.

es im Rahmen von MEND nur sechs Erhebungsorte, wodurch die Anzahl der Gewährspersonen pro Ortspunkt deutlich höher ausfällt als bei Projekten, die Hunderte von Erhebungsorten abdecken. Milroy / Gordon (2003: 21) beschreiben diesen Unterschied folgendermaßen: „*In favoring breadth across communities over depth within communities, this project* [Telsur[10]] *has much in common with traditional dialect geography.*"

Ein Problem von Datensammlungen wie MEND besteht darin, dass nur Personen teilnehmen konnten, die sowohl die Stimulussprache als auch die nicht eng verwandte Zielsprache auf hohem Niveau beherrschen. Dies schloss z.B. viele (ältere) monolinguale Frauen in Mexiko und Bolivien aus, da sie kaum Spanisch beherrschten. Außerdem schloss es jüngere Mennoniten in den USA und Brasilien aus, wenn diese das Plautdietsche nicht gut genug beherrschten. Während der Ausschluss der zweiten Gruppe kein prinzipielles Problem darstellt, wenn es um die Erforschung der plautdietschen Syntax geht, liegt im Ausschluss der ersten Gruppe ein nicht zu unterschätzendes Analyserisiko.

3. Struktur der Stimulussätze

Bevor wir den strukturellen Zuschnitt einzelner Stimulussätze beschreiben (vgl. auch Kaufmann 2005), führen wir die leicht gekürzte Beschreibung von MEND auf der IDS-Homepage[11] an. Hier wird der allgemeine Zuschnitt der Stimulussätze beschrieben (vgl. Abschnitt 4 zu Details der Übernahme und Aufbereitung ins AGD):

> Das Korpus MEND, das im Zeitraum von 1999 bis 2002 von Göz Kaufmann erhoben wurde, besteht aus den plautdietschen Übersetzungen von 46 Stimulussätzen durch 321 mennonitische Informant(inn)en. Insgesamt handelt es sich also um etwa 14.500 verwertbare Satzübersetzungen mit einer Gesamtaufnahmedauer von etwa 40 Stunden. Dieses Korpus wurde vom AGD in Zusammenarbeit mit Göz Kaufmann und Aaron Schmidt-Riese aufbereitet. Im Regelfall wurden in Mexiko, in Paraguay und in Bolivien spanische Stimuli verwendet, während in Brasilien portugiesische und in den USA englische Stimuli zum Einsatz kamen. […] Die 46 Stimulussätze […] decken verschiedene Satztypen ab. Neben sechs Hauptsätzen (Sätze 41–46) werden zehn nachgestellte Komplementsätze (Sätze 1–10), zehn vorangestellte Konditionalsätze (Sätze 11–20), zehn nachgestellte Kausalsätze (Sätze 21–30) und zehn satzmediale oder satzfinale Relativsätze (Sätze 31–40) abgefragt. Alle Nebensätze bedingen die zusätzliche Übersetzung

10 Telsur ist das Forschungsprojekt, das dem *Atlas of North American English* zugrunde liegt (vgl. Labov et al. 2005).

11 Vgl. http://agd.ids-mannheim.de/MEND_extern.shtml; zuletzt aufgerufen am 14.06.2021.

> eines Matrixsatzes, Hauptsatz 42 beinhaltet einen vorangestellten Temporalsatz. Achtzehn Stimulussätze zielen auf Übersetzungen mit einem (Partikel)Verb (Sätze 1–4, 11–14, 21–24, 31–34, 41+42), achtzehn Stimulussätze auf die Übersetzung mit zwei Verben (Modalverb+Infinitiv bzw. Tempusauxiliar+Partizip; Sätze 5–8, 15–18, 25–28, 35–38, 43+44) und zehn Stimulussätze auf die Übersetzung von drei Verben (9 Sätze mit kontrafaktischer Proposition und Modalverb und 1 Satz (Satz 9) mit einem epistemischen Modalverb mit Infinitiv Perfekt; Sätze 9+10, 19+20, 29+30, 39+40, 45+46). […] Die Hauptverben der Sätze regieren fast immer ein direktes Objekt, mit dessen Hilfe die Positionen der finiten und infiniten Verbelemente besser bestimmt werden können. Daneben wurden in einigen Sätzen Adverbien oder Negationsmarker eingebaut. Die Sätze wurden dem jeweiligen Informanten/der jeweiligen Informantin einzeln vorgelesen[12] und dann sofort – ohne Hilfe einer schriftlichen Version – übersetzt. […] Bei Übersetzungsproblemen wurde der Stimulus entweder gleich oder am Ende des Interviews noch einmal vorgelesen.

Die folgende detailliertere Beschreibung dient dazu, die grundsätzlichen Ziele der Datenerhebung exemplarisch zu konturieren. Daneben werden wir auch auf einige Probleme der Stimulussätze und auf weitere, ursprünglich nicht im Fokus stehende Analysemöglichkeiten hinweisen. In dieser Hinsicht teilen wir die Methodenkritik von Bucheli / Glaser (2002: 61) nicht:

> However, translation carries the danger that too many unintended variants appear. Even if these unintended variants inspire the linguist to conduct further research, all these useless answers, which come up to 10–15% of the whole, clearly show the disadvantage of translation: the control over the elicitation is minimal because the informant has too much freedom in answering.

Wir schätzen die Möglichkeit, anhand von nicht beabsichtigter Variation weitere Forschungen durchführen zu können, als deutlich gewichtiger ein als die Gefahr des Datenverlustes, selbst wenn dieser 10–15 % beträgt. In vielen Fällen führen Variationen, die nicht im Fokus der eigentlichen Untersuchung standen, also

---

12 Die Sätze wurden immer in derselben Reihenfolge vorgelesen, wobei mit weniger komplexen Sätzen begonnen wurde. Ob diese Reihenfolge zu Verzerrungen durch Ermüdung oder eine mögliche Beeinflussung früherer auf spätere Sätze geführt hat, wurde noch nicht untersucht. Im lexikalischen Bereich kam dies in seltenen Fällen vor. Es wurde aber immer darauf geachtet, dass Sätze mit ähnlichen Charakteristika (Satztyp, Verbanzahl etc.) nicht direkt hintereinander vorgelesen wurden. Die verwendete Satzreihenfolge und alle Stimulusvarianten können in der Rubrik *Zusatzmaterial* (MEND_Z_01_Stimulussätze – *real sequence*) eingesehen werden. Für die englische Version beschreibt diese Datei auch die *structural sequence*, die die Sätze nach ähnlichen Charakteristika ordnet. Hieraus erschließt sich die Nummerierung von <1> bis <46>.

z.B. lexikalische Varianten oder Aussprachevarianten, sowieso nicht dazu, dass Übersetzungen unbrauchbar werden.

3.1 (Isolierte) Hauptsätze

Stimulussatz <45> ist ein kontrafaktischer Hauptsatz:[13]

Stimulus <45>	Englisch	Yesterday I could have sold the ring.
(1)	Spanisch	Ayer podría haber vendido el anillo.
	Portugiesisch	Ontem eu poderia ter vendido o anel.

An den Übersetzungen dieses Satzes lässt sich wegen des vorangestellten Adverbs zum einen die V2-Qualität des Plautdietschen überprüfen. Diese ist fast durchgängig gegeben (eine Ausnahme ist (2c)). Interessant ist nun aber auch, wie die drei Verben dieses Satzes übersetzt werden. Entscheidend für die Analyse von Verbclustern ist die Tatsache, dass diese hohe Zahl an Verbformen auch in einem V2-Hauptsatz ein Cluster aus mindestens zwei Verben am Satzende garantiert. Die Mennoniten produzierten dabei meistens drei (vgl. (2a+c)) bzw. vier Verben (vgl. (2b)), hier immer mit *verb projection raising* (vgl. die Abfolge des Modalverbs und des Hauptverbs und Fußnote 2).

13 Bei einem Pilotprojekt, bei dem einige wenige Pommer/inne/n in Rio Grande do Sul die 46 Stimulussätze übersetzten, sorgten die an syntaktischen *rara* reichen Übersetzungen von Stimulussatz <45> dafür, dass weitere fünfzehn Stimulussätze erstellt wurden und dass zwischen 2017 und 2021 291 Pommer/inne/n aus Brasilien die jetzt 61 Stimulussätze übersetzten (vgl. für eine erste Analyse dieser *rara* Kaufmann 2022). Wie die Übersetzungen von 24 brasilianischen Hunsrücker/inne/n aus Rio Grande do Sul stehen diese pommerschen Daten aus Rio Grande do Sul, Santa Catarina, Espírito Santo und Rondônia der Öffentlichkeit (noch) nicht zur Verfügung.

Stimulus <45>		Portugiesisch	Ontem eu poderia ter vendido o anel						
		Englisch	Yesterday I could have sold the ring						
(2)	a.	Gestere	hat	ik	könnt	den	Fingerring	verköpe[14]	
		gestern	*hatte.1SG.PRT*	*ich.1SG.NOM*	*gekonnt.PART*	*den*	*Ring*	*verkaufen.INF*[15]	
								(MEND_S_00281/Bra-6; F/23/Pl)[16]	
	b.	Gestere	hat	ik	könnt	min	F- (0.3)	Ring	
		gestern	*hatte.1SG.PRT*	*ich.1SG.NOM*	*gekonnt.PART*	*meinen*	*~~F~~- (0.3)*	*Ring*	
		verkauft	habe.						
		verkauft.PART	*haben.INF*						
								(MEND_S_00282/Bra-7; F/47/Pl+Port)	
	c.	Gesteren	ik	hat	könnt	den	Ring	verköpen	
		gestern	*ich.1SG.NOM*	*hatte.1SG.PRT*	*gekonnt.PART*	*den*	*Ring*	*verkaufen.INF*	
								(MEND_S_00041/USA-42; F/47/Pl)	

Neben der Frage der Verbserialisierung und der Frage, warum einige Informant/inn/en drei Verbformen produzieren, während andere vier Verbformen

14 Das Plautdietsche wird in einer an das Standarddeutsche angelehnten Orthographie wiedergegeben. Es handelt sich also um eine einfache, durchaus noch fehlerbehaftete Transliteration und nicht um eine genaue phonetische oder konversationsanalytische Transkription. Die vorhandenen „Transkripte" sollten bei Analysen also immer überprüft werden. Die Stimulussätze werden zuerst in der jeweils verwendeten Stimulussprache angeführt. Ist dies nicht Englisch, folgt noch die englische Version. Während wir nicht-sprachliche Ereignisse mit doppelten Klammern anzeigen (z.B. *((lacht))* in (23)), werden ungefüllte Pausen mit einfachen Klammern markiert für den Fall, dass sie länger als 0.25 Sekunden dauern (z.B. *(0.3)* für das Intervall von 0.25 bis 0.34 Sekunden). Gefüllte Pausen werden als *äh(m)* transliteriert. Abbrüche und Reparaturen werden mit einem Bindestrich markiert, ein Doppelpunkt repräsentiert eine deutlich hörbare Segmentdehnung.

15 Die folgenden Abkürzungen werden in den standarddeutschen Glossen verwendet (diese Glossen sind nicht Teil des MEND-Korpus): SG (Singular); PL (Plural); NOM (Nominativ); PRS (Präsens); PRT (Präteritum); PART (Partizip II); INF (Infinitiv); KOMPL (Komplementierer); ADV (Adverb(iale Fügung)). Unterstrichene Elemente markieren Abweichungen vom Stimulussatz; ein ø zeigt an, dass ein Element des Stimulussatzes nicht übersetzt wurde. Außer bei zusätzlichen verbalen Elementen, insbesondere bei Auxiliaren wie *dune* (‚tun') in (4b), zeigen durchgestrichene Elemente an, dass die Gewährsperson dem Stimulussatz etwas hinzufügte oder eine Reparatur stattfand.

16 MEND_S_00281 ist die Nummerierung im MEND-Korpus. Bra-6 ist die Nummerierung in den Publikationen von Göz Kaufmann. Alle anderen Informationen in dieser Zeile sind in Fußnote 3 aufgeschlüsselt.

verwenden,[17] gibt es noch mindestens drei weitere Punkte, die in Satz <45> untersucht werden könnten. Zum einen bietet das Subjektpronomen *ik* eine von vielen Möglichkeiten, die unterschiedlichen Grade der Palatalisierung von /k/ im Kontext von Vordervokalen zu analysieren (vgl. Siemens 2012: 92–98, der *etj* statt *ik* schreibt, und das Ende von Abschnitt 5.3). Zum anderen ist die Reparatur in (2b) interessant. Informantin MEND_S_00282/Bra-7 will *anel* zuerst als *F(ingerring)* übersetzen, entscheidet sich dann aber nach einer kurzen Pause für *Ring*. Auch hier könnte es lohnend sein, Gründe für diese Variation zu eruieren. Schließlich fällt in derselben Übersetzung der Possessivdeterminierer *min* auf. Der Stimulussatz verwendet hier einen definiten Artikel. Zu analysieren, warum einige Informant/inn/en eine Possessivkongruenz zwischen Subjektpronomen und Determinierer produzieren, während die meisten der Vorgabe des Stimulus folgen, könnte im Rahmen kognitiver Untersuchungen von Interesse sein. Für diese Fragestellung könnten mehrere Stimulussätze untersucht werden, neben <45> z.B. auch Stimulussatz <42>, ein Hauptsatz mit einem vorangestellten Temporalsatz:

Stimulus <42>	Englisch	Before leaving the house I always turn off the lights.
(3)	Spanisch	Antes de irme de casa siempre apago las luces.
	Portugiesisch	Antes de sair de casa eu sempre apago as luzes.

Der Grund für die Verwendung dieses Satzes war nicht die Frage, ob und warum einige Informant/inn/en dem definiten Artikel bei *house* und manchmal auch bei *lights* einen Possessivdeterminierer vorziehen. Vielmehr ging es darum, wie die Mennonit/inn/en Partikelverben verwenden, in diesem Fall also, wie sie *turn off*/*apago* übersetzen. Daneben war wichtig zu sehen, ob das finite Verb im Plautdietschen direkt nach einem vorangestellten Nebensatz serialisiert wird. Obwohl die Mehrzahl der Sätze integrierte Temporalsätze wie in (4b) aufwiesen, gab es doch relativ viele desintegrierte Temporalsätze wie in (4a), eine Tatsache, die dem Bild bei vorangestellten Konditionalsätzen entspricht (vgl. (11)).

17 Quantifizierende Analysen bedürfen natürlich der statistischen Aufbereitung des Korpus. Göz Kaufmann hat dies für viele (morpho)syntaktische Charakteristika der 14.500 Übersetzungen durchgeführt. Diese SPSS-Datei ist zwar nicht Teil des MEND-Korpus, allerdings wurden vor kurzem (DGD-Release 2.16) Prototypen aus dem ZuMult-Projekt integriert (vgl. https://zumult.org/; zuletzt aufgerufen am 14.06.2021), im speziellen ZuRecht (vgl. Frick / Schmidt 2020), mit dessen Hilfe nun eine Suchanfrage anhand der Satznummern der MEND-Sätze möglich ist. Es können jetzt also alle Übersetzungen eines bestimmten Satzes in einer eigenen Datei in der DGD generiert werden.

Stimulus <42>		Spanisch	Antes de irme de casa siempre apago las luces
		Englisch	Before leaving the house I always turn off the lights
(4)	a.	Ehe ik von T'Hus fohr ik switch immer dat Lich üt	
		ehe ich von Zuhause fahre ich mache immer das Licht aus	
			(MEND_S_00125/Mex-51; M/22/Pl)
	b.	Ehe ik äh von Hüs fahren du du ik immer die Lichter ütmeaken	
		ehe ich äh von Haus fahren tue tue ich immer die Lichter ausmachen	
			(MEND_S_00084/Mex-10; F/15/Pl)

Auch hier zeigen sich weitere Analysemöglichkeiten. Interessant ist zum einen das relativ häufige Vorkommen der artikellosen Konstruktion *von Hüs* wie in (4b) (statt *von T'Hus* in (4a) oder *von det/vom Hüs* (nicht angeführt)). Daneben stellt sich die Frage, wie die Variation zwischen hohem Vorder- und Hintervokal in *Hüs* und *T'Hus* und bei vielen anderen Lexemen zu erklären ist (vgl. Siemens 2012: 62). Auch verwendet (4b), anders als (4a), sowohl im Temporalsatz als auch im Matrixsatz habituelles *dune*, das von *immer* getriggered wird. Schließlich ist es sicherlich eine Diskussion wert, warum viele Informant/inn/en in Mexiko englische Lehnwörter wie *switch […] üt* in (4a) verwenden. Aufgrund der Benennung von Einzelbildern hat Kaufmann (1997: 167–185) diese Frage bereits untersucht. Für den Großteil des MEND-Korpus stehen Untersuchungen zu Lehnwörtern aber noch aus. Kaufmann (2017) beschränkt sich auf Übersetzungen aus Brasilien, also auf das Plautdietische der Mennonit/inn/en, das Hünsrückische und das Pomerano (vgl. Fußnote 13).

Nur in Stimulussatz <42> kommt ein Temporalsatz vor. Trotzdem liegt das Hauptaugenmerk von MEND auf Nebensätzen, und zwar auf Komplement-, Konditional-, Kausal- und Relativsätzen. Wir stellen nun einige Stimuli dieser Nebensatztypen vor, wobei wir mit Komplementsätzen beginnen.

3.2 Nachgestellte Komplementsätze

Die Stimulussätze <2> und <8> weisen nachgestellte Komplementsätze auf:

Stimulus <2>	Englisch	John doesn't think that you know your friends well.
(5)	Spanisch	Juan no cree que conozcas bien a tus amigos.
	Portugiesisch	O João não acha que tu conheces bem os teus amigos.

Stimulus <8>	Englisch	Are you sure that he has repaired the chair?
(6)	Spanisch	¿Estás seguro que él arregló la silla?
	Portugiesisch	Tu tens certeza que ele consertou a cadeira?

In beiden Stimuli ist das Bemühen erkennbar, komplementiererlose Übersetzungen zu vermeiden (vgl. die *dat*-losen Übersetzungen der Sätze (c) und (e) in Abbildung 1). In Satz <2> geschieht das mithilfe der Negation des Matrixsatzes, in Satz <8> mit einem interrogativen Matrixsatz in Verbindung mit einem Prädikatsadjektiv. Der Frage, warum trotz aller Bemühungen viele Sätze ohne Komplementierer übersetzt wurden, widmet sich ein Kapitel zur Desintegration von Komplementsätzen in Kaufmann (2016: 308–359). In Satz <2> hängt dies auch mit dem Matrixverb *think* zusammen, dessen Verneinung die Folge einer Anhebung der Negationspartikel *not* sein könnte. Viele Informant/inn/en bevorzugten aber die bedeutungsneutrale Verschiebung der Negation vom Matrixsatz in den Komplementsatz, in gewissem Sinne eine Umkehrung des angenommenen *negative raising*.[18] Wurde das Matrixverb nicht verneint, konnte der Komplementierer wie in (7) problemlos weggelassen werden (vgl. für Sonderzeichen und Unter/Durchstreichungen die Fußnoten 14 und 15):

Stimulus <2>	Englisch	John doesn't think that you know your friends well
(7)	Johann gleuft	dü (0.6) kenns die Fr- dine Frend nich sehr gut
	Johann glaubt	*Ø Ø du (0.6) kennst ~~die~~ ~~Fr-~~ deine Freunde nicht ~~sehr~~ gut*

(MEND_S_00038/USA-39; M/46/Pl)

Die Übersetzung in (7) macht deutlich, dass bei der Erstellung der Stimulussätze nicht alle Ziele erreicht wurden. Einige Stimulussätze weisen solche oder ähnliche Unzulänglichkeiten auf. In dieser Hinsicht unterscheidet sich MEND nicht von anderen Projekten, die auf Übersetzungen aufbauen. Wie beschreibt es Dorian (1981: 118) so treffend, „*[…] the test sentences proved not always to be perfectly designed.*“

18 *Negative raising* beschreibt die bedeutungsneutrale Anhebung einer Negationspartikel aus einem eingebetteten Satz in einen Matrixsatz. Semantisch gibt es z.B. keinen Unterschied zwischen (i) und (ii) (vgl. Collins / Postal 2014 für eine ausführliche Diskussion).

(i) Er glaubt, dass Stan **nicht** nach Gelsenkirchen gefahren ist.
(ii) Er glaubt **nicht**, dass Stan nach Gelsenkirchen gefahren ist.

Übersetzung (8) zeigt einen weiteren interessanten Fall, der in Kaufmann (2016) behandelt wird, die Verwendung der Relativpartikel *wat* in Komplementsätzen. Konvergenzen zwischen Komplement- und Relativsätzen gibt es natürlich in vielen Sprachen, genannt seien hier nur die (teilweise) übereinstimmenden Einleiter im Spanischen, Portugiesischen (jeweils *que*) und Englischen (*that*).

Stimulus <8>	Spanisch	Estás seguro que arregló la silla
	Englisch	Are you sure that he has repaired the chair
(8)	Bis dü sicher wat	her dat- (0.6) äh den Stuhl haf fertiggemeakt
	bist du sicher was.KOMPL	*er ~~das~~- (0.6) äh den Stuhl hat fertiggemacht*

(MEND_S_00087/Mex-13; M/28/Pl)

An der in (8) vorhandenen Reparatur zeigt sich ein weiteres, bisher noch nicht konsequent untersuchtes Variationsphänomen. Der Informant scheint sich des Genus von *Stuhl* nicht sicher zu sein. Dieses Phänomen betrifft hauptsächlich Lehnwörter, findet sich aber auch bei einigen, zum Teil frequenten plautdietschen Erbwörtern wie *Stuhl*. Nur für die US-amerikanischen Daten liegt mit Kaufmann (2008) eine erste Analyse solcher Abweichungen vor.

3.3 Vorangestellte Konditionalsätze

Die folgenden Stimuli beinhalten jeweils einen vorangestellten Konditionalsatz mit zwei Verben und einem Adverb. Stimulus <15> weist ein finites Modalverb und ein temporales Satzadverb auf, Stimulus <17> ein finites Tempusauxiliar und ein epistemisches Adverb.

Stimulus <15>	Englisch	If he has to sell the house now, he will be very sorry.
(9)	Spanisch	Si tiene que vender la casa ahora, se va a poner muy triste.
	Portugiesisch	Se ele tiver que vender a casa agora, ele vai ficar muito triste.
Stimulus <17>	Englisch	If he really killed the man, nobody can help him.
(10)	Spanisch	Si realmente mató al hombre, nadie lo puede ayudar.
	Portugiesisch	Se ele realmente matou o homem, ninguém pode ajudar ele.

Kaufmann (2007 und 2016) behandeln die Position der Adverbien in beiden Sätzen, Kaufmann (2016: 430–448) das Auftreten deiktischer Anaphern wie *der* statt nicht-deiktischem *her* (vgl. (11)). Auch die Desintegration des Konditionalsatzes (vgl. ebenfalls (11)) und die Verwendung von *dann* (vgl. (12)) werden dort ausführlich im Rahmen der (Des)Integration von Nebensätzen diskutiert. Was dabei allerdings noch fehlt, ist eine Analyse der Intonation, die mit unterschiedlichen Graden der (Des)Integration korrelieren dürfte.

Stimulus <15>	Englisch	If he has to sell the house now he'll be very sorry
(11)	Wann der dat Hüs nü mut verköpen der wird sehr (0.7) sorry sein	
	wenn der das Haus nun muss verkaufen der wird sehr (0.7) traurig sein	
		(MEND_S_00008/USA-8; F/14/E>PL-Ø[19])

Stimulus <17>	Spanisch	Si realmente mató al hombre nadie lo puede ayudar
	Englisch	If he really killed the man nobody can help him
(12)	Wann der wirklich den Omtje todgemeakt haf dann kann ihm keiner helpen	
	wenn der wirklich den Mann umgebracht hat ~~dann~~ kann ihm keiner helfen	
		(MEND_S_00096/Mex-22; M/17/Pl)

Wie viele andere lexikalische Elemente kommen auch *sorry* in <15> und *hombre* (‚Mann') in<17> in mehreren Stimulussätzen vor. Hier könnte es sich lohnen, zu untersuchen, ob die Verwendung von Lehnwörtern wie *sad*, *sorry* oder *triste* (alle ‚traurig') wie in (11) nur von den Charakteristika der Gewährspersonen abhängt, oder ob auch unterschiedliche Satzkontexte eine Rolle spielen. Daneben gibt es eine faszinierende Variation bei *hombre* in (12), die unter Umständen davon abhängen könnte, ob dieses Lexem im Singular oder Plural vorkommt. Es kommen neben *Omtje* (etymologisch *Onkelchen*) in (12) auch *Mann*, *Mensch* und in Brasilien *Onkel* vor, wohl eine standarddeutsche Übertragung von *Omtje*.

3.4 Nachgestellte Kausalsätze

Die Stimulusversionen in (13) präsentieren nachgestellte Kausalsätze:

19 Bei 31 Informant/inn/en in den USA und vierzehn Informant/inn/en in Brasilien liegen nur Angaben über die dominante Sprache vor, aber kein exakter Wert für die Kompetenz in jeder einzelnen Sprache (vgl. Fußnote 6).

Stimulus <21>	Englisch	He is not coming, because he doesn't have any time.
(13)	Spanisch	No va a venir porque no tiene tiempo.
	Portugiesisch	Ele não vem porque não tem tempo.

Hier mag die jeweilige Stimulussprache einen Einfluss auf die Übersetzungen gehabt haben. Man könnte sich z.B. fragen, ob der strukturelle Unterschied zwischen komplexem *not […] any* im Englischen und simplerem *no*/*não* im Spanischen und Portugiesischen die Verteilung von *nich* wie in (14a) und *keine* wie in (14b) beeinflusst hat.

Stimulus <21> Englisch He- he is not coming because he doesn't have any time[20]
Spanisch No va a venir porque no tiene tiempo

(14) a. Der wird nich kumme wegens der nich Tied haf
der wird nicht kommen weil der nicht Zeit hat
(MEND_S_00269/Bol-3; M/31/Pl)

b. Hei kemmt nich weils hei haft keine Tied
er kommt nicht weil er hat keine Zeit
(MEND_S_00266/Men-47; F/60/Pl)

Neben der Frage der Negation im Kausalsatz von (14a+b) ist noch unklar, warum manche Gewährspersonen das Tempusauxiliar *wird* im Matrixsatz benutzen und andere nicht. Daneben könnte untersucht werden, wovon die in allen Sprachgemeinschaften vorkommende Variation zwischen *wegen(s)* und *weil(s)*/*weil(s)* abhängt (vgl. Abbildung 12 und die darauffolgende Diskussion). Einzig die Frage, ob der Kausalsatz als V2-Satz oder als Nebensatz mit finalem Finitum übersetzt wird, wird in Kaufmann (2003 und 2016) ausführlich diskutiert. Diese Variation hängt nicht von der Wahl des einleitenden Elements ab.

3.5 Relativsätze

Stimulussatz <32> beinhaltet einen restriktiven Relativsatz, in dem der Relativmarker das direkte Objekt des Satzes ist. Stimulussatz <31> dagegen enthält einen Relativmarker in Subjektfunktion. Bei diesem Satz ist es aufgrund der kontextlosen Präsentation nicht möglich zu entscheiden, ob der Relativsatz von den Gewährspersonen als restriktiv oder nicht-restriktiv interpretiert

20 Die Doppelung des Personalpronomens *he* ergibt sich daraus, dass die Stimulusversionen konkreter Übersetzungen so aufgeführt werden, wie sie vorgelesen wurden (vgl. auch Fußnote 24).

wurde. Stimulussatz <32> zeigt daneben noch einmal, dass viele Lexeme mehrfach vorkommen. *Sad* kann mit *sorry* in Satz <15> (vgl. (9)) verglichen werden – die romanischen Versionen verwenden in beiden Fällen *triste* – und *men* kommt jetzt anders als in Satz <17> (vgl. (10)) im Plural vor.

Stimulus <32>	Englisch	The stories that he is telling the men are very sad.
(15)	Spanisch	Las historias que les está contando a los hombres son muy tristes.
	Portugiesisch	As estorias que ele está contando para os homens são muito tristes.

Stimulus <31>	Englisch	I don't like people who make a lot of noise.
(16)	Spanisch	No me gustan las personas que hacen mucho ruido.
	Portugiesisch	Eu não gosto de pessoas que fazem muito barulho.

Die Übersetzungen von Satz <31> in (17a+b) zeigen noch einmal nachdrücklich, welche Analysemöglichkeiten MEND bietet:

Stimulus <31>		Spanisch	No	me	gustan	las	personas	que	hacen	mucho	ruido	
		Englisch	I	don't	like	people	who	make	a	lot	of	noise
(17)	a.		Ik	gleich	die	Menschen	nich	wat	da	sehr	lüt	sin
			Ich	*mag*	*die*	*Menschen*	*nicht*	*die*	*da*	*sehr*	*laut*	*sind*
												(MEND_S_00086/Mex-12; M/28/Pl)
	b.		äh:	(0.3)	Mi	gefolle	die	Mensche	nich	die	viel	Krach moake
			äh:	*(0.3)*	*mir*	*gefallen*	*die*	*Menschen*	*nicht*	*die*	*viel*	*Krach machen*
												(MEND_S_00207/Fern-26; F/39/HD>Pl-64 %)

Statt unmarkiertem *wat* verwendet Informant MEND_S_00086/Mex-12 in (17a) *wat da*, eine komplexere Form der Relativpartikel, während Informantin MEND_S_00207/Fern-26 das Relativpronomen *die* verwendet. Diese Varianten wurden in Kaufmann (2018b) untersucht. Nicht untersucht wurde allerdings die Verteilung der unterschiedlichen Übersetzungsweisen des Matrixverbs (*gleiche*/*gefolle*) und die genaue Verteilung der *n*-Apokope in *Mensche(n)* und vielen anderen Lexemen, die weitgehend, aber nicht ausschließlich auf die jeweils dominante Varietät des Plautdietschen zurückzuführen ist (Chortitza oder

Molotschna;[21] vgl. Siemens 2012 für viele hier bestehende Unterschiede und für die *n*-Apokope speziell Siemens 2012: 13 und 64).

3.6 Kontext- und Häsitationsphänomene bei den Übersetzungen

Zum Abschluss dieses Abschnittes sollen noch (einmal) der außersprachliche Kontext, Reparaturen und ungefüllte Pausen angesprochen werden, da sie bei Übersetzungen zum Teil eine andere Rolle spielen als bei freien Gesprächen. Die ersten beiden Phänomene können an zwei Übersetzungen von Stimulussatz <4> illustriert werden:

Stimulus <4>	Englisch	Can't you see that I am wearing a new dress?
(18)	Spanisch	¿No ves que estoy usando un vestido nuevo?
	Portugiesisch	Não estás vendo que eu estou usando um vestido novo?

Dieser Satz wurde von zwei US-amerikanischen Gewährspersonen folgendermaßen übersetzt:

Stimulus <4>		Englisch	Can't you see that I'm wearing a new dress											
(19)	a.	Kos	dat	nich	sehen	kiek	ik	ha	ja	en	nüet	Kleid		
		kannst	*Ø*	~~*es/das*~~	*nicht*	*sehen*	~~*schau*~~	*ich*	*habe*	~~*ja*~~	*ein*	*neues*	*Kleid*	
													(MEND_S_00068/USA-81; M/48/Pl)	
	b.	Kos	nich	sehen	dat	ik	ha	en-	äh	dat	ik	en	nüet	Kleid anha
		kannst	*Ø*	*nicht*	*sehen*	~~*dass*~~	~~*ich*~~	~~*habe*~~	~~*ein-*~~	*äh*	*dass*	*ich*	*ein*	*neues* *Kleid* *anhabe*
													(MEND_S_00019/USA-20; F/14/E>Pl-Ø)	

Wie viele andere männliche Gewährspersonen ist sich Informant MEND_S_00068/USA-81 der Seltsamkeit bewusst, dass er jemanden „fragen" soll, ob er/sie „sein" neues Kleid sieht. Ihm ist der wahrscheinliche Kontext eines solchen

21 Die Gewährspersonen aus den USA, aus Mexiko und aus Bolivien sprechen die Chortitza-Varietät, während die Gewährspersonen aus Brasilien und aus den Kolonien Fernheim und Neuland in Paraguay die Molotschna-Varietät verwenden. Die Kolonie Menno in Paraguay gehört historisch zur ersten Gruppe, allerdings ist das Bild heute nicht mehr eindeutig, da ihre Bewohner/innen einem starken, auch sprachlichen Einfluss aus Fernheim und Neuland ausgesetzt waren.

Satzes, also eines von einer Frau geäußerten Vorwurfs einer an modischen Fragen uninteressierten Person gegenüber, klar, und er inszeniert diesen Kontext in (19a) mithilfe einer in der Transliteration nicht repräsentierten starken Betonung auf *sehen*, einer steigenden finalen Intonation auf *Kleid*, eines aufmerksamkeitssichernden *kiek* (‚schau') und der Modalpartikel *ja*. Gerade die relativ seltene Verwendung von in den Stimulussätzen nicht vorkommenden (Modal)Partikeln wie *ja*, *doch* und *ook* zeugt davon, dass in vielen Fällen ein außersprachlicher Kontext für die kontextlos präsentierten Stimulussätze konstruiert wurde.

In (19b) repariert Informantin MEND_S_00019/USA-20 die scheinbare V2-Position des Vollverbs *ha*, indem sie es in die Endposition des Komplementsatzes bringt. Dass die erste Verbposition wirklich eine scheinbare und keine echte V2-Position ist, wird deutlich, wenn man sich alle Nebensätze mit nur einer Verbform ansieht. Insbesondere bei Komplementsätzen in den USA zeigt sich dabei, dass in einigen Übersetzungen wie in (20) das Verb vor dem Objekt, aber nach einem eventuell vorhandenen Adverb serialisiert wird und dass es sich dabei nicht um Übersetzungsfehler, sondern um eine Art von *verb projection raising* mit nur einem Verb handelt (vgl. Kaufmann 2015 und Fußnote 2).

Stimulus <2>	Spanisch	Juan no cree que conozcas bien a tus amigos								
	Englisch	John doesn't think that you know your friends well								
(20)	Johann	gleuf	nich	dat	dü	gut	kenns	sine	Frend	
	Johann	*glaubt*	*nicht*	*dass*	*du*	*gut*	*kennst*	*seine*	*Freunde*	

(MEND_S_00100/Mex-26; M/34/Pl)

Reparaturen wie in (21) sind gleichfalls aufschlussreich (vgl. die drei Stimulusvarianten in (10)):

Stimulus <17>	Englisch	If he really killed the man nobody can help him										
(21)	If	der	ap ierns	den	Mensch	todgemeak	haf	kann-	keiner	kann	den	helpen
	wenn	*der*	*wirklich*	*den*	*Mensch*	*umgebracht*	*hat*	~~*kann-*~~	*keiner*	*kann*	*den*	*helfen*

(MEND_S_00065/USA-78; M/20/E>Pl-71 %)

Der Übersetzer von (21) beginnt den Matrixsatz mit dem finiten Verb, integriert also den Konditionalsatz auf erwartbare Weise. Dann aber bricht er ab und beginnt mit dem indefiniten Subjektpronomen *keiner*. Die Präferenz einer Desintegration, die stark mit dem phonetischen Gewicht des Subjekts zusammenhängt (*keiner* ist schwerer als z.B. die Personalpronomen *her* oder *der*), ist ein deutlicher Hinweis darauf, dass die Desintegration von Konditionalsätzen im Plautdietschen nicht dem Einfluss der Konstituentenabfolge des (englischen) Stimulussatzes geschuldet ist. Daneben erscheint *den* im Matrixsatz von (21) im Akkusativ und nicht im Dativ. Auch hier gibt es in allen Erhebungsorten ein großes Maß an Variation in beide Richtungen, das bisher nur ansatzweise untersucht wurde (vgl. Kaufmann 2004 und 2011).

Der dritte Punkt, der noch erwähnt werden soll, sind ungefüllte Pausen. Wie gefüllte Pausen (vgl. (4b), (8), (17b) und (19b)) können ungefüllte Pausen ein Zeichen von Übersetzungsproblemen sein. Wir illustrieren ihre Präsenz mit einer Übersetzung von Stimulussatz <12>:

Stimulus <12>	Englisch	If he does his homework, he can have some ice cream.
(22)	Spanisch	Si hace sus deberes, puede tomar helado.
	Portugiesisch	Se ele fizer o tema, ele pode comer sorvete.

Informantin MEND_S_00294/Bra-19 produziert folgendes Token:

Stimulus <12>	Portugiesisch	Se ele fizer o tema ele pode comer sorvete
	Englisch	If he does his homework he can have some ice cream
(23)	Wann her de (0.9) Hüsapgabe maakt kann her en Sorvet ete da- (0.6)	
	wenn er die (0.9) Hausaufgabe macht ~~kann~~ ~~er~~ ~~ein~~ ~~Eis~~ ~~essen~~ ~~da-~~ (0.6)	
	((lacht)) daut's nich- ((lacht)) kann her en Ies habe	
	((lacht)) ~~das ist~~ ~~nicht~~ ((lacht)) kann er ein Eis haben	

(MEND_S_00294/Bra-19; F/50/Pl)

Die Informantin stockt schon vor *Hüsapgabe*, was in Brasilien vielfach mit einem portugiesischen Lehnwort übersetzt wird, verwendet dann aber im Matrixsatz selbst das Lehnwort *Sorvet* für *Eis*. Nach einer weiteren, relativ langen Pause lacht die Informantin und beginnt einen Metakommentar. Bevor sie diesen beendet – wahrscheinlich will sie sagen *daut's nich Plautdietsch* –, fällt ihr

das plautdietsche Wort *Ies* ein, und sie übersetzt den Matrixsatz erneut. Viele (un)gefüllte Pausen, Segmentdehnungen und Metakommentare sind das Resultat einer Vermeidungsstrategie der Gewährspersonen, die offensichtlich keine deutlich erkennbaren Lehnwörter in plautdietschen Übersetzungen produzieren wollten (vgl. Kaufmann 2017). Andere hängen vermutlich mit der Komplexität einzelner Stimulussätze zusammen. Eine vollständige Analyse dieser Phänomene fehlt noch.

In diesem Abschnitt haben wir die Struktur einiger Stimulussätze vorgestellt und anhand ihrer Übersetzungen auf verschiedene analysewürdige Dimensionen der Variation hingewiesen. Viele dieser Phänomene wurden noch nicht (erschöpfend) analysiert. Dies betrifft zum einen den Gebrauch von Lehnwörtern, die Kasus- und Genusmorphologie von Pronomen und Determinierern und die Möglichkeit einer Kontextkonstruktion der Gewährspersonen während des Übersetzungsprozesses. Zum anderen betrifft es die Satz- und Wortintonation, die Analyse von (un)gefüllten Pausen, Segmentdehnungen und Metakommentaren und die Aussprache von Einzellexemen. In Bezug auf die Aussprache sei dabei noch einmal darauf hingewiesen, dass diese nicht im Zentrum der Transkripte von MEND stand. Sie sollte also immer überprüft werden. Schließlich stellt sich die Frage, ob man bei Gewährspersonen mit relativ geringer Kompetenz im Plautdietschen Prozesse des Sprachverfalls feststellen kann. Immerhin 23 der 276 Gewährspersonen (8,3 %), für die genaue Informationen zur Sprachkompetenz vorliegen (vgl. Fußnote 19), erreichen weniger als 10 von 14 Kompetenzpunkten, neun sogar 8 Punkte oder weniger (3,3 %). In den beiden folgenden Abschnitten zeigen wir nun, wie die Daten von der AGD übernommen und aufbereitet wurden, über welche Informationen das MEND-Korpus verfügt und wie die vorhandenen Suchoptionen angewendet werden können.

4. Übernahme und Aufbereitung des MEND-Korpus am AGD

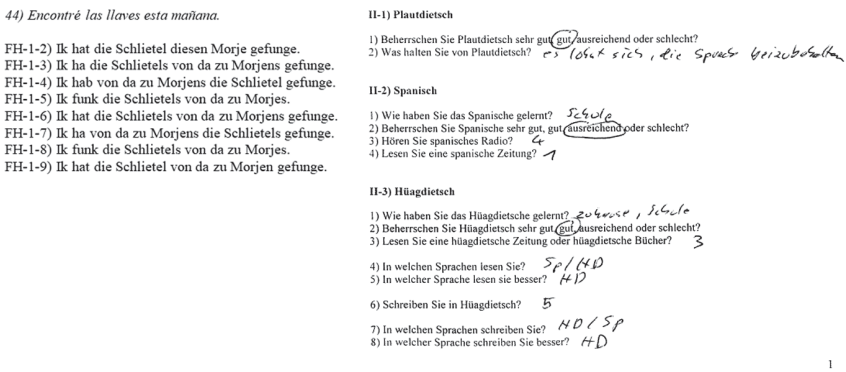
Das Archiv für Gesprochenes Deutsch (AGD) am Leibniz-Institut für Deutsche Sprache (vgl. Stift / Schmidt 2014) ist ein Forschungsdatenzentrum, das auf Korpora des gesprochenen Deutsch, insbesondere auf Gesprächs-, Variations- und Interviewkorpora spezialisiert ist. Als Mitglied im KonsortSWD[22] ist es an der Nationalen Forschungsdateninfrastruktur (NFDI) beteiligt und bringt sich in weitere Initiativen in den Bereichen des Forschungsdatenmanagements, der digitalen Forschungsdateninfrastruktur und der *Digital Humanities* ein. Die

22 Vgl. https://www.konsortswd.de/; zuletzt aufgerufen am 14.06.2021.

am AGD archivierten und bereitgestellten Daten setzen sich folgendermaßen zusammen:

- Bestandsdaten seines 1932 gegründeten Vorgängers *Deutsches Spracharchiv* (DSAv), z.B. die Korpora *Deutsche Mundarten*, auch als *Zwirner-Korpus* bekannt (vgl. Zwirner / Bethge 1958), oder *Grundstrukturen*, auch als *Freiburger Korpus* bekannt (vgl. Engel / Vogel 1975)
- Korpora, die in IDS-eigenen Projekten erstellt wurden, z.B. das Forschungs- und Lehrkorpus *Gesprochenes Deutsch* (FOLK; vgl. Schmidt 2018) oder das Korpus *Deutsch Heute* (DH; vgl. Kleiner 2015)
- Korpora, die von externen Projekten in Zusammenarbeit mit dem AGD aufgebaut wurden, z.B. die Korpora *Deutsch in Namibia* (vgl. Zimmer et al. 2020) oder *Unserdeutsch* (vgl. Götze et al. 2018)
- Daten, die von abgeschlossenen Forschungsprojekten übernommen und am AGD für eine Weitergabe und Archivierung aufbereitet werden, z.B. die Korpora *Australiendeutsch* (vgl. Clyne 1981) oder *Gesprochene Wissenschaftssprache Kontrastiv* (GeWiss, vgl. Fandrych et al. 2012)

Das MEND-Korpus fällt unter die letztgenannte Kategorie der abgeschlossenen Forschungsprojekte und reiht sich in den Kontext einer Anstrengung des AGD ein, vermehrt Daten zu extraterritorialen Varietäten des Deutschen in seine Bestände zu integrieren (vgl. dazu Gorisch et al. 2022). Nach Klärung der Übernahmemodalitäten wurden die Originaldaten des Korpus dem AGD im März 2017 übergeben. Im Wesentlichen handelte es sich um die Originale der Aufnahmen auf Kompaktkassetten, Transliterationen der übersetzten Sätze in Word-Dokumenten (vgl. Abbildung 2 links), PDF-Scans der ausgefüllten Fragebögen (vgl. Abbildung 2 rechts) sowie eine SPSS-Datei, in der Informationen zu den Sprecher/inne/n maschinenlesbar kodiert waren.



44) Encontré las llaves esta mañana.

FH-1-2) Ik hat die Schlietel diesen Morje gefunge.
FH-1-3) Ik ha die Schlietels von da zu Morjens gefunge.
FH-1-4) Ik hab von da zu Morjens die Schlietel gefunge.
FH-1-5) Ik funk die Schlietels von da zu Morjes.
FH-1-6) Ik hat die Schlietels von da zu Morjens gefunge.
FH-1-7) Ik ha von da zu Morjens die Schlietels gefunge.
FH-1-8) Ik funk die Schlietels von da zu Morjes.
FH-1-9) Ik hat die Schlietel von da zu Morjen gefunge.

II-1) Plautdietsch

1) Beherrschen Sie Plautdietsch sehr gut, gut, ausreichend oder schlecht?
2) Was halten Sie von Plautdietsch? es lohnt sich, die Sprache weiterzubehalten

II-2) Spanisch

1) Wie haben Sie das Spanische gelernt? Schule
2) Beherrschen Sie Spanische sehr gut, gut, ausreichend oder schlecht?
3) Hören Sie spanisches Radio? 4
4) Lesen Sie eine spanische Zeitung? 1

II-3) Hüagdietsch

1) Wie haben Sie das Hüagdietsche gelernt? zuhause, Schule
2) Beherrschen Sie Hüagdietsch sehr gut, gut, ausreichend oder schlecht?
3) Lesen Sie eine hüagdietsche Zeitung oder hüagdietsche Bücher? 3

4) In welchen Sprachen lesen Sie? Sp / HD
5) In welcher Sprache lesen sie besser? HD

6) Schreiben Sie in Hüagdietsch? 5

7) In welchen Sprachen schreiben Sie? HD / Sp
8) In welcher Sprache schreiben Sie besser? HD

1

Abbildung 2: Ausschnitte von transliterierten Übersetzungen (links) und einem gescannten Fragebogen (rechts)

Die Aufbereitung im AGD hatte zum Ziel, eine erste vollständig digitale Version des Korpus zu erstellen, die dem Datenmodell der Datenbank für Gesprochenes Deutsch (DGD, vgl. Abschnitt 5 und dgd.ids-mannheim.de) entspricht und über diese zur Nachnutzung bereitgestellt werden kann. Dazu wurden zunächst die Kompaktkassetten im AGD-Medienstudio professionell digitalisiert und gemäß der darauf aufgezeichneten Sprechereignisse in Einzeldateien geschnitten. Metadaten zu den Sprechereignissen und den daran beteiligten Sprecher/inne/n wurden aus den Fragebögen und der SPSS-Datei in das AGD-Metadatenschema übertragen. Aus den Word-Dokumenten wurden die von Göz Kaufmann transliterierten Einzelsätze extrahiert und aufnahmeweise in der korrekten Reihenfolge zu Transkriptdateien im Format des EXMARaLDA Partitur-Editors gebündelt.[23] Dabei wurde jedem übersetzten Satz der „kanonische", also per Erhebungsmethode vorgesehene Stimulus vorangestellt und zwischen Stimulus und Übersetzung jeweils eine Leerstelle im Transkript für die dort üblicherweise in der Aufnahme befindliche Pause eingefügt. Anschließend wurde dieses Transkript mit der zugehörigen Audiodatei verknüpft und von

23 Die Originale sind am AGD archiviert und sollen zu einem späteren Zeitpunkt in die DGD integriert werden. Eine Ansicht, die systematisch elizitierte Korpusdaten (wie sie außer in MEND auch in vielen anderen Variationskorpora erhoben wurden, beispielsweise die Wenkersätze im Zwirner-Korpus) statt nach Aufnahmen nach ihren Vorlagen (hier: den Übersetzungsstimuli) organisiert, ist ein Desiderat, das sich gleichfalls in Bearbeitung befindet.

einer studentischen Hilfskraft satzweise aligniert (vgl. Abbildung 3). Die daraus resultierenden Transkripte wurden schließlich automatisch in das von der DGD verwendete Format konvertiert und dabei tokenisiert, also in einzelne Wörter segmentiert. Audios, Metadaten und Transkripte (sowie Fragebögen, die Satzlisten und einige weitere Zusatzmaterialien) wurden als erste Fassung des Korpus im November 2018 in der Version 2.11 der DGD veröffentlicht.

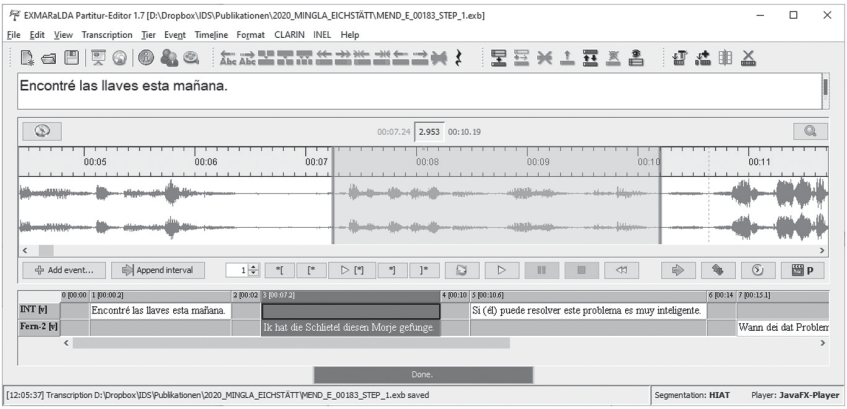


Abbildung 3: Transkript eines Interviews im EXMARaLDA Partitur-Editor

Mit den Transkripten in dieser Form lassen sich bereits systematische Recherchen auf den plautdietschen Übersetzungen ausführen, der zugehörige Ausschnitt der Audioaufnahme lässt sich gezielt ansteuern und Metadaten zum Sprechereignis und zu den Sprecher/inne/n lassen sich abrufen. Allerdings blendet diese Form der Transkription eine Reihe interessanter Phänomene noch weitestgehend aus, insbesondere:

- die genaue sprachliche Form, in der der Stimulussatz geäußert wurde (vgl. Fußnote 20)
- Performanzphänomene bei der Äußerung der Übersetzung wie Pausen, Reparaturen, Fehl- und Neustarts und Ähnliches (vgl. z.B. (19b) und (21)–(23))
- Eventuelle weitere Bestandteile des Interviews, insbesondere weitere abgebrochene oder korrigierte Übersetzungsversuche

In einem weiteren Schritt wurde daher in einer Kooperation des AGD mit Göz Kaufmann Aaron Schmidt-Riese beauftragt, die Transkripte mit Blick auf die genannten Phänomene zu verfeinern. Wie in Abbildung 4 illustriert, geben die

daraus resultierenden Transkripte den genauen Verlauf des Interviews präziser wieder: Die genaue Form der Intervieweräußerung (im gegebenen Beispiel identisch mit der Schriftform des Stimulussatzes) sind darin ebenso festgehalten wie äußerungsinterne Pausen (*((0.3))*) und abgebrochene Übersetzungsversuche. Im Beispiel wird eine standardnahe Übersetzung, *Isch hab die Schlüssel diesen Morgen*, vom Interviewer unterbrochen, der mit seiner Intervention *Plautdietsch* auf die erwartete Zielsprache aufmerksam macht.

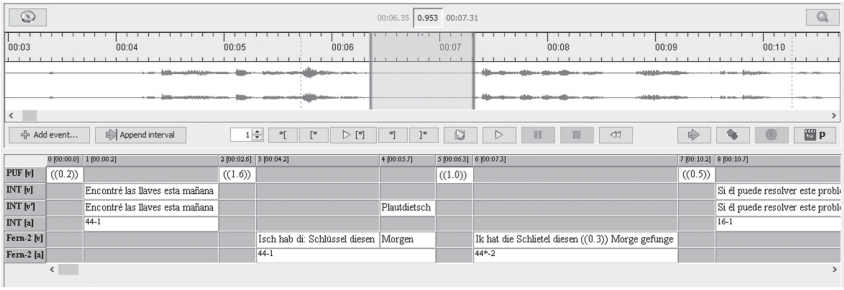
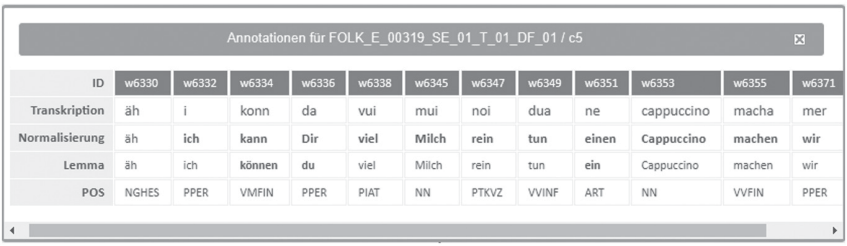


Abbildung 4: Verfeinertes Transkript[24]

Für eine systematische Korpusrecherche in gesprochensprachlichen Daten hat es sich bei Gesprächskorpora des Deutschen bewährt, literarisch transkribierte Formen (*isch*, *zwo*, *runner*) in einer zusätzlichen Annotationsebene auf ihre standardorthographischen Entsprechungen (*ich*, *zwei*, *runter*) abzubilden. Eine solche Normalisierung sorgt nicht nur dafür, dass verschieden transkribierte Realisierungen des gleichen Lexems (wie z.B. *nee*, *nö* oder *nein*) über die Suche nach einer einzigen Referenzform (eben standardorthographisches *nein*) gefunden werden können, sondern sie ermöglicht es auch, sprachtechnologische Tools, die auf Grundlage von und für schriftsprachliche Daten entwickelt wurden, zur

24 Die Zeile *INT [v]* präsentiert die intendierte Form des Stimulussatzes. *INT [v']* gibt die konkret vorgelesene Form wieder. Diese beiden Formen sind nicht immer identisch (vgl. (14)). In *INT [v']* erscheinen auch Metakommentare oder Erklärungen des Interviewers. Die Zeile *INT [a]* nennt zuerst die Nummer des jeweiligen Stimulussatzes und dann den ersten bzw. weitere Vorleseversuche (44-1). Auch die Übersetzungsversuche der jeweiligen Gewährspersonen in Zeile *Fern-2 [v]* sind in Zeile *Fern-2 [a]* nummeriert. In diesem Fall übersetzt der Informant zuerst in eine standardnahe Varietät (44-1). Der zweite Versuch, 44*-2, ist dann auf Plautdietsch, wobei der Asterisk die beste (und zumeist einzige) Übersetzung markiert. Die erste Zeile *PUF [v]* quantifiziert die Pause zwischen Vorlesen und Übersetzen und nächstem Vorlesen.

Anreicherung der Transkripte mit zusätzlichen Annotationen anzuwenden. Für die DGD-Korpora wird auf diese Weise eine Lemmatisierung (Abbildung auf Grundformen *gehst* ➔ *gehen*) und ein *Part-Of-Speech-Tagging* (POS: *gehst* ➔ VVFIN für „finites Vollverb") erstellt, durch die viele korpuslinguistische Herangehensweisen überhaupt erst ermöglicht werden (vgl. Abbildung 5).



ID	w6330	w6332	w6334	w6336	w6338	w6345	w6347	w6349	w6351	w6353	w6355	w6371
Transkription	äh	i	konn	da	vui	mui	noi	dua	ne	cappuccino	macha	mer
Normalisierung	äh	**ich**	**kann**	**Dir**	**viel**	**Milch**	**rein**	**tun**	**einen**	**Cappuccino**	**machen**	**wir**
Lemma	äh	ich	**können**	**du**	viel	Milch	rein	tun	**ein**	Cappuccino	machen	wir
POS	NGHES	PPER	VMFIN	PPER	PIAT	NN	PTKVZ	VVINF	ART	NN	VVFIN	PPER

Abbildung 5: Normalisierung, Lemmatisierung, POS-Tagging einer Äußerung aus FOLK

Eine orthographische Normalisierung birgt umso mehr Schwierigkeiten, je weiter die zu normalisierenden Daten vom Standarddeutschen entfernt sind. Exemplarisch beschreiben dies Schütte / Schmidt (2017) für das Niederdeutsche. Beim Plautdietschen, dem eher der Status einer eigenständigen germanischen (Minderheiten)Sprache denn der einer „bloßen" Varietät des Deutschen zuzuschreiben ist, sind diese Schwierigkeiten sehr ausgeprägt. Sie resultieren neben der grundsätzlichen Problematik der Verschriftlichung (vgl. Fußnote 14) aus einer Vielzahl von älteren niederländischen und russischen Lehnwörtern bzw. neueren englischen, spanischen und portugiesischen Lehnwörtern, aber auch aus zahlreichen lexikalischen Eigenheiten wie z.B. häufiges *Omtje* oder *Onkel* für *Mann* (vgl. (12)). Auch plautdietsches *wiese(n)*, das z.B. in den Übersetzungen von Stimulus <46> *I should have shown the little dog to the kids* vorkommt, ist hier problematisch, da es als *zeigen* und nicht als das standarddeutsche Kognat *weisen* normalisiert werden müsste. Auch die allgemeine Verwendung von *dat* sowohl für deiktisches *das* als auch für nicht-deiktisches *es* (vgl. (19a)) und von *wat* als Komplementierer (vgl. (8)) führen zu einer nicht eineindeutigen Normalisierung.

Die Normalisierung des MEND-Korpus ist dementsprechend eine aufwändige Aufgabe mit Pilotcharakter, weil für viele Einzelfälle, die nicht von Normalisierungsregeln für Varietäten des Deutschen (wie Winterscheid et al. 2019, das hier als Ausgangspunkt verwendet wurde) abgedeckt sind, immer aufs Neue Entscheidungen getroffen werden müssen. Im Zuge der AGD-Aufbereitung des Korpus konnte dennoch für einen Ausschnitt der Daten eine manuelle

Normalisierung durchgeführt werden, auf deren Grundlage der Rest des Korpus automatisch normalisiert wurde. Das Ergebnis ist nicht fehlerfrei (etwa 10 % der Formen der Interviewten sind nicht korrekt normalisiert), verbessert aber dennoch die Recherchierbarkeit des Korpus deutlich. Für Lemmatisierung und POS-Tagging gilt Ähnliches. Auch hier wäre zunächst die Entwicklung von Regeln und Annotationsrichtlinien für das Plautdietsche notwendig, bevor entsprechende automatische Verfahren trainiert und angewendet werden können. Dies bleibt – wie auch der Entwurf robuster Richtlinien für die Normalisierung standardferner Varietäten (vgl. Blevins 2019) – eine Aufgabe für die Zukunft. Das MEND-Korpus ist aktuell auf der Lemma- und POS-Ebene nicht annotiert (vgl. Abbildung 6).

Annotationen für MEND_E_00183_SE_01_T_01_DF_01 / c7

ID	w13	w14	w15	w16	w17	w18	w19
Transkription	Ik	hat	die	Schlietel	diesen	Morge	gefunge
Normalisierung	**ich**	hat	die	**Schlüssel**	diesen	Morge	**gefunden**
Lemma	x	x	x	x	x	x	x
POS	Y	Y	Y	Y	Y	Y	Y

Abbildung 6: Normalisierung einer Übersetzung aus MEND[25]

Die überarbeiteten Transkripte mit Normalisierungsannotationen wurden im Januar 2020 mit Version 2.13 der DGD veröffentlicht.

5. MEND in der DGD

Bleiben wir in dem von Mark Louden verwendeten Bild der Goldmine (vgl. Abschnitt 1) und wagen in diesem Abschnitt einen Blick in diese Mine. Als Werkzeug steht uns die DGD zur Verfügung. Ursprünglich sind die Funktionen und Optionen dieses Werkzeuges dafür konzipiert, in transkribierten natürlichen Sprachdaten nach bestimmten Phänomenen zu recherchieren, ohne eine feste Vorstellung von den zu erwartenden sprachlichen Äußerungen zu haben. In Gesprächskorpora ist jede Äußerung einzigartig, und die Recherche zielt darauf ab, aus der unübersichtlichen Menge an Äußerungen, alle einer bestimmten

25 Man beachte die fehlerhafte automatische Normalisierung von plautdietsch *hat* als *hat* statt als *hatte* und plautdietsch *Morge* als *Morge* statt als *Morgen*. Lemmata und POS-Tags wurden in MEND noch nicht vergeben.

Form gleichenden Fälle zu finden, um sie in einem weiteren Arbeitsschritt z.B. anhand des Kontextes zu analysieren.

Mit experimentell strukturierten Daten, wie sie in MEND elizitiert wurden, ist die Anforderung an die Recherche genau umgekehrt. Hier wurden immer wiederkehrende Stimulussätze dargeboten und übersetzt. Eine Recherche zielt hier darauf ab, alle gegebenenfalls variierenden Formen einer konkreten Vorgabe zu finden. Die DGD, das Verbreitungs- und Analyseinstrument für die AGD-Daten, ist für derartige Daten noch nicht ausgerichtet. Inzwischen ist es, wie in Fußnote 17 erwähnt, aber immerhin möglich, sich die durchnummerierten Sätze in MEND systematisch als Satzlisten ausgeben zu lassen. Trotzdem blicken wir immer noch mit einem Werkzeug auf diese Goldmine, das eher für das Arbeiten im Kohlebergwerk konzipiert wurde.

Trotzdem lässt sich auch ohne Satzlisten so einiges ans Tageslicht befördern; denn gerade weil man nicht an Satzlisten gebunden ist, lassen sich mit den Funktionalitäten der DGD Analysen durchführen, wie sie beispielhaft in Abschnitt 3 erwähnt wurden. Die Arbeit mit einigen dieser Phänomene werden wir weiter unten demonstrieren. Diese betreffen unter anderem Kontextphänomene, wie z.B. die Suche nach bestimmten übersetzten Wörtern im Kontext von bestimmten Wörtern eines Stimulussatzes, sowie das Fokussieren auf die unterschiedlichen Ausgangssprachen, auf Sprechereigenschaften und auch auf akustische Phänomene, die erst durch das Text-Ton-Alignment analysierbar werden.

5.1 Browsing durch MEND

Durch das *Browsing* erhält man einen ersten Überblick über das Korpus. Der Reiter *Korpus* (vgl. Abbildung 7) enthält eine Kurzbeschreibung (vgl. den Anfang von Abschnitt 3) und Informationen wie zentrale Publikationen, Ereignisse, Sprecher, Transkripte, Audios und Zusatzmaterialien.

Abbildung 7: Korpusüberblick

Weitere Details zu jedem dieser Punkte liefern die anderen Reiter. Die einzelnen Datensätze sind untereinander verlinkt, sodass man zwischen Ereignissen, Sprechern, Transkripten und Audios navigieren kann. Der obere Teil der 321 Ereignisse umfassenden Tabelle ist in Abbildung 8 dargestellt. Die Archivierung der sonstigen Bezeichnungen dient zur nachhaltigen und transparenten Verknüpfung von Forschungsergebnissen, wie sie z.B. in Kaufmanns Publikationen zu diesen Daten bereits erschienen sind. Daneben werden hier der Erhebungsort (Ortsname) und das Erhebungsdatum genannt.

#	Ereignis-ID ▲ ▼	Sonstige Bezeichnungen ▲ ▼	Ortsname ▲ ▼	Erhebungsdatum ▲ ▼
1	MEND_E_00001 ►	Sm-1-1 ; USA-1	Seminole	1999
2	MEND_E_00002 ►	Sm-1-2 ; USA-2	Seminole	1999
3	MEND_E_00003 ►	Sm-1-3 ; USA-3	Seminole	1999
4	MEND_E_00004 ►	Sm-1-4 ; USA-4	Seminole	1999
5	MEND_E_00005 ►	Sm-1-5 ; USA-5	Seminole	1999
6	MEND_E_00006 ►	Sm-1-6 ; USA-6	Seminole	1999
7	MEND_E_00007 ►	Sm-1-7 ; USA-7	Seminole	1999
8	MEND_E_00008 ►	Sm-1-8 ; USA-8	Seminole	1999

Abbildung 8: Ereignisse mit DGD-Kennungen, Bezeichnungen des Erhebungs--projektes, Erhebungsort und Erhebungsjahr

Ebenso lang wie die Tabelle der Ereignisse (321 Einträge) ist die Tabelle aller Sprecher/innen (vgl. Abbildung 9). Hier erfährt man außerdem noch das Geburtsjahr und das Geschlecht der Gewährspersonen. Das Alter der Sprecher/innen zum Erhebungszeitpunkt ist jeweils in den Ereignissen dokumentiert.

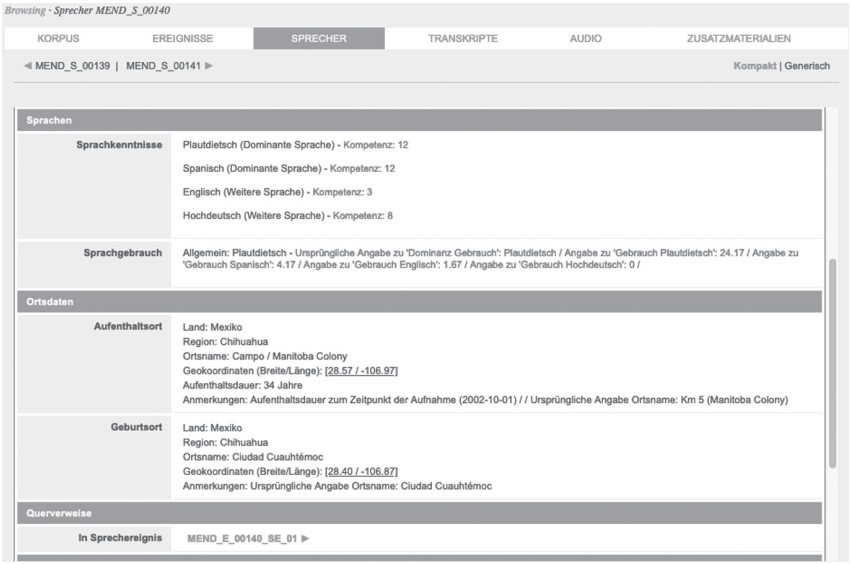
Browsing · Sprecherliste MEND

KORPUS | EREIGNISSE | SPRECHER | TRANSKRIPTE | AUDIO | ZUSATZMATERIALIEN

#	Sprecher-ID ▲ ▼	Sonstige Bezeichnungen ▲ ▼	Geburtsjahr ▲ ▼	Geschlecht ▲ ▼
1	MEND_S_00001 ▶	Sm-1-1 ; USA-1	1970	Weiblich
2	MEND_S_00002 ▶	Sm-1-2 ; USA-2	1984	Männlich
3	MEND_S_00003 ▶	Sm-1-3 ; USA-3	1985	Weiblich
4	MEND_S_00004 ▶	Sm-1-4 ; USA-4	1985	Männlich
5	MEND_S_00005 ▶	Sm-1-5 ; USA-5	1983	Männlich
6	MEND_S_00006 ▶	Sm-1-6 ; USA-6	1979	Männlich
7	MEND_S_00007 ▶	Sm-1-7 ; USA-7	1983	Weiblich
8	MEND_S_00008 ▶	Sm-1-8 ; USA-8	1985	Weiblich

Abbildung 9: Sprecherübersicht mit ursprünglichen Bezeichnungen des Erhebungsprojekts, dem Geburtsjahr und dem Geschlecht der Gewährspersonen

Zu den Sprechern wurden auch einige weitere Sozialdaten und Informationen zur Sprachkompetenz erhoben, wie in Abbildung 10 beispielhaft an einem Sprecher dargestellt. Die Sprachkompetenzen und der Sprachgebrauch sind hier aufgelistet (vgl. Fußnote 6). MEND_S_00140/Mex-68 ist ambilingual Plautdietsch-Spanisch (jeweils 12 Punkte), gebraucht in intra-ethnischen Domänen aber überwiegend das Plautdietsche (24,17 aus 30 Punkten).



Browsing · Sprecher MEND_S_00140

KORPUS | EREIGNISSE | SPRECHER | TRANSKRIPTE | AUDIO | ZUSATZMATERIALIEN

◀ MEND_S_00139 | MEND_S_00141 ▶ Kompakt | Generisch

Sprachen

Sprachkenntnisse	Plautdietsch (Dominante Sprache) - Kompetenz: 12 Spanisch (Dominante Sprache) - Kompetenz: 12 Englisch (Weitere Sprache) - Kompetenz: 3 Hochdeutsch (Weitere Sprache) - Kompetenz: 8
Sprachgebrauch	Allgemein: Plautdietsch - Ursprüngliche Angabe zu 'Dominanz Gebrauch': Plautdietsch / Angabe zu 'Gebrauch Plautdietsch': 24.17 / Angabe zu 'Gebrauch Spanisch': 4.17 / Angabe zu 'Gebrauch Englisch': 1.67 / Angabe zu 'Gebrauch Hochdeutsch': 0 /

Ortsdaten

Aufenthaltsort	Land: Mexiko Region: Chihuahua Ortsname: Campo / Manitoba Colony Geokoordinaten (Breite/Länge): [28.57 / -106.97] Aufenthaltsdauer: 34 Jahre Anmerkungen: Aufenthaltsdauer zum Zeitpunkt der Aufnahme (2002-10-01) / / Ursprüngliche Angabe Ortsname: Km 5 (Manitoba Colony)
Geburtsort	Land: Mexiko Region: Chihuahua Ortsname: Ciudad Cuauhtémoc Geokoordinaten (Breite/Länge): [28.40 / -106.87] Anmerkungen: Ursprüngliche Angabe Ortsname: Ciudad Cuauhtémoc

Querverweise

In Sprechereignis	MEND_E_00140_SE_01 ▶

Abbildung 10: Weitere Sozialdaten wie detaillierte Sprachkenntnisse und Ortsdaten

In jedem Ereignis gibt es genau ein Sprechereignis, d.h. die Aufnahme der Übersetzung der 46 Stimulussätze. Jedes dieser Sprechereignisse hat ein Transkript (vgl. Abbildung 11). Hier erfährt man etwas über die genaue Anzahl an Types (von 256 bis 599), Tokens (von 578 bis 1669) und über die Aufnahmedauer (von 4:45 bis 13:58 Minuten).

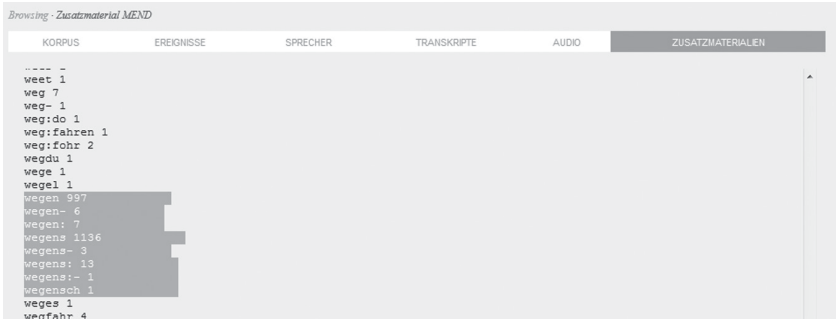
Browsing · Transkriptliste MEND

KORPUS | EREIGNISSE | SPRECHER | TRANSKRIPTE | AUDIO | ZUSATZMATERIALIEN

#	Transkript-ID ▲ ▼	Types ▲ ▼	Tokens ▲ ▼	Dauer ▲ ▼	Format ▲ ▼	Alignment (j/n) ▲ ▼
1	MEND_E_00001_SE_01_T_01 ▶	416	954	00:05:55	FLN	ja
2	MEND_E_00002_SE_01_T_01 ▶	393	954	00:06:56	FLN	ja
3	MEND_E_00003_SE_01_T_01 ▶	410	977	00:07:13	FLN	ja
4	MEND_E_00004_SE_01_T_01 ▶	388	953	00:08:42	FLN	ja
5	MEND_E_00005_SE_01_T_01 ▶	398	959	00:07:50	FLN	ja
6	MEND_E_00006_SE_01_T_01 ▶	397	965	00:05:58	FLN	ja
7	MEND_E_00007_SE_01_T_01 ▶	393	977	00:09:47	FLN	ja
8	MEND_E_00008_SE_01_T_01 ▶	386	960	00:07:20	FLN	ja

Abbildung 11: Übersicht der Transkripte

In der Rubrik *Audio* (ohne Abbildung) sind alle Aufnahmen aufgelistet, verlinkt und abspielbar. Schließlich lassen sich in der Rubrik *Zusatzmaterial* (ebenfalls ohne Abbildung) alle Stimulussätze einsehen (vgl. Fußnote 12) und die Fragebögen der Erhebung nachvollziehen. Die Wortlisten (in alphabetischer Reihenfolge und nach Frequenz) bieten einen guten Einstieg in den Inhalt des Korpus. So wird z.B. ersichtlich, dass es viele Vorkommen für die Kausaleinleiter *wegen(s)* gibt (vgl. Abbildung 12 und (14a)).



Browsing · Zusatzmaterial MEND

KORPUS | EREIGNISSE | SPRECHER | TRANSKRIPTE | AUDIO | ZUSATZMATERIALIEN

```
---- -
weet 1
weg 7
weg- 1
weg:do 1
weg:fahren 1
weg:fohr 2
wegdu 1
wege 1
wegel 1
wegen 997
wegen- 6
wegen: 7
wegens 1136
wegens- 3
wegens: 13
wegens:- 1
wegensch 1
weges 1
wegfahr 4
```

Abbildung 12: Wortliste (alphabetisch mit Häufigkeit) aus dem Zusatzmaterial mit plautdietschen Formen für die Kausaleinleiter *wegen(s)*

Eine weitere Suche nach der Zeichenkette *wegen* führt zu weiteren Wortformen, wie *Wegens* (1), *dawegen* (10), *dawegens* (3), *deswegen* (3), *dewegen* (8), *dewegens* (2) und *Dowegen* (2). Wenn wir auf diese Weise eine Liste an möglichen Kandidaten an Wortformen für *wegen* gefunden haben, können wir diese in die Funktion *Tokensuche* übernehmen (siehe Details zu dieser Art Recherche weiter unten), indem wir entweder eine Liste der Vorkommen verwenden oder einen regulären Ausdruck ersinnen, der auf alle diese Formen passt. In der Tokensuche ist unter dem Fragezeichen-Icon eine kurze Hilfestellung zu regulären Ausdrücken vorhanden. In diesem Fall würde der reguläre Ausdruck *(W|da|des|de|Do)?(w)?egen(s)?* in der normalisierten Tokensuche zu 2161 Treffern und diesen Formen führen:

Transkribierte Formen
wegens (1136) ; wegen (997) ; wegens: (13) ; dawegen (10) ; dewegen (8) ; wegen: (7) ; dawegens (3) ; deswegen (3) ; dewegens (2) ; Dowegen (1) ; Wegens (1)

Normalisierte Formen
wegen (2153) ; deswegen (14) ; dawegen (10) ; dewegens (2) ; Dowegen (1) ; Wegens (1)

Die transkribierten Formen beruhen auf der initial erstellten Transliteration (vgl. Fußnote 14). Die meisten Korpora in der DGD wurden zusätzlich mit einer normalisierten Annotationsspur versehen. Wie in Abschnitt 4 erklärt, wurde die Normalisierung für MEND teilweise automatisch erstellt, basierend auf einem Modell, das auf etwa 10 % händisch annotierter MEND-Daten trainiert wurde. Deshalb verwundert es nicht, dass das eine Vorkommen von *Wegens* nicht wie alle anderen 1136 Vorkommen von *wegens* zu *wegen* normalisiert wurde. Vermutlich liegt es daran, dass *Wegens* am Satzanfang, also in Großschreibung, nur einmal vorkommt und dadurch nicht in der Portion von 10 % Trainingsdaten war und deshalb auch das Modell auf diesen Fall nicht trainiert wurde.

Eine weitere lohnende Suche bzw. eine Möglichkeit, die Liste der Kausaleinleiter zu ergänzen, wäre die Suche nach *weils*, *weil*, *wiels* und *wiel* (normalisiert *weil*; vgl. (14b)). Bei dieser Fülle an Daten stößt eine manuelle Analyse allerdings schnell an Grenzen. Der entsprechende reguläre Ausdruck wäre *((W|da|des|de|Do)?(w)?egen(s)?|weil)* und führte zu 3288 Treffern, wie in randomisierter Form in Abbildung 13 dargestellt:

Transkribierte Formen
wegens (1136) ; wegen (997) ; weils (638) ; wiels (263) ; weil (150) ; wiel (52) ; wegens: (13) ; dawegen (10) ; dewegen (8) ; wegen: (7) ; dawegens (3) ; deswegen (3) ; dewegens (2) ; weilst (2) ; wielst (2) ; Dowegen (1) ; Wegens (1)

Normalisierte Formen
wegen (2153) ; weil (1107) ; deswegen (14) ; dawegen (10) ; dewegens (2) ; Dowegen (1) ; Wegens (1)

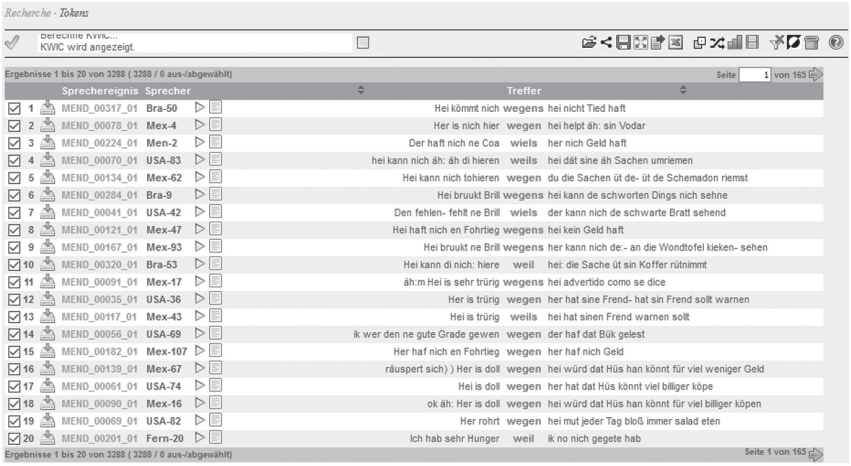


Abbildung 13: KWIC-Trefferliste (*Keywords in Context*) zum regulären Ausdruck *((W|da|des|de|Do)?(w)?egen(s)?|weil)* nach dem randomisierten Durchmischen der Liste

5.2 Strukturierte Metadatensuche und virtuelle Korpora

Eine relativ große Menge an weiteren Daten wurde im Erhebungsprojekt von den Gewährspersonen abgefragt und ein Teil davon ist in die Metadaten des Korpus eingeflossen. Startet man z.B. eine Suche nach dem Land des Aufnahmeortes und wählt den Wert *Brasilien* (vgl. Abbildung 14), werden alle Ereignisse mit diesen Eigenschaften zusammengestellt.

Abbildung 14: Strukturierte Metadatensuche mit Auswahl an Deskriptoren

Das Ergebnis kann dann als virtuelles Korpus im eigenen Benutzeraccount (vgl. *Meine DGD*) gespeichert werden und so als Grundlage weiterer Recherchen dienen (vgl. Abbildung 15). Dadurch muss man nicht bei jeder Tokensuche oder Recherche nach diesem Metadatum (oder anderen Metadaten) filtern.

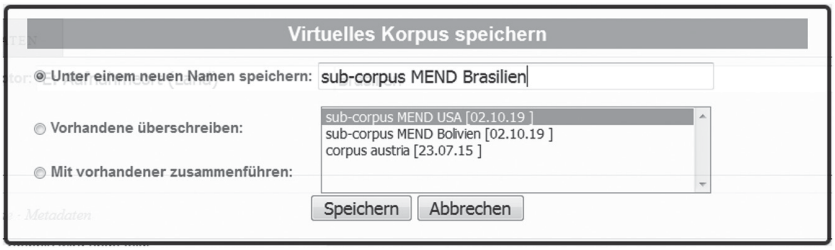


Abbildung 15: Speichern von virtuellen Korpora (dies ist mit allen Rechercheergebnissen möglich, also auch denen, die bei der Tokensuche entstehen)

Ergebnisse wie dieses virtuelle Subkorpus können auch per ID mit anderen DGD-Nutzer/inne/n geteilt werden (vgl. Abbildung 16).

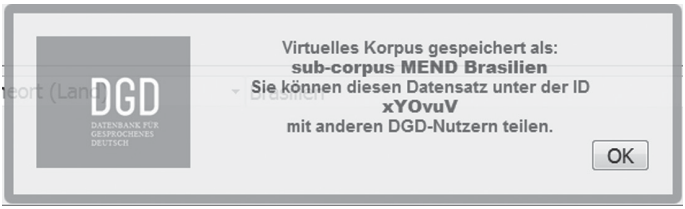


Abbildung 16: Teilen von virtuellen Korpora mit Kolleg/inn/en mit Zugang zur DGD

Solch ein Vorgehen kann dann kollaboratives Arbeiten unterstützen.

5.3 Struktursensitive Tokensuche

Ein häufig gewähltes Vorgehen bei der Recherche in der DGD ist die Tokensuche in Kombination mit dem Filtern des Kontextes (zusätzlich können auch Metadaten mit einbezogen werden). Als Beispiel suchen wir nach dem Wort *dogs* ('Hunde'), welches vom Interviewer in Stimulussatz <10> *He didn't know that he should have fed the dogs this morning* geäußert wird. Insgesamt werden 119 Treffer gefunden. Dass die Zahl höher ist als die Anzahl der in den USA interviewten Gewährspersonen, liegt unter anderem daran, dass in Mexiko und Paraguay viele Informant/inn/en das Englische als Stimulussprache dem Spanischen vorzogen. In der Übersetzung von Satz <10> würde man dann ein Wort ähnlich der standarddeutschen Form *Hunde* erwarten. In der Wortliste (aller Wörter in allen Transkripten, vgl. Zusatzmaterial) lassen sich folgende Kandidaten an Formen für *Hund* (Stimulus <46>) und *Hunde* (Stimulus <10>) finden (Probanden

variieren in der Übersetzung manchmal vom Plural *dogs* zum Singular *Hund*, also interessieren uns zunächst beide Formen):

> H:und (3), H:und- (1), H:ung (1), Hund (239), Hund- (13), Hund: (4), Hunde (2), Hundje (102), Hundje- (2), Hundjes (4), Hundke (1), Hunds (1), Huntjes (1), Hung (269), Hung- (4), Hung: (3), Hunge (1), Hungen (2), Hungs (1), Hunje (1), Hunn (5), Hüng (1)

Auch ohne einen komplizierten regulären Ausdruck zu erstellen, lässt sich nach all diesen Formen der Kontext von *dogs* filtern (vgl. Abbildung 17). Dazu wählt man unter *Kontext* in den beiden Feldern *Kontext* z.B. „50 Tokens" und „beidseitig", im Feld *Skopus Transkript* aus und fügt im Suchfeld *Transkribiert* folgenden regulären Ausdruck ein:

> *(H:und|H:und-|H:ung|Hund|Hund-|Hund:|Hunde|Hundje|Hundje-|Hundjes|Hundke|Hunds|Huntjes|Hung|Hung-|Hung:|Hunge|Hungen|Hungs|Hunje|Hunn|Hüng)*

Zusätzlich setzt man den Haken bei *Reguläre Ausdrücke* und drückt den Knopf *Kontext filtern*.

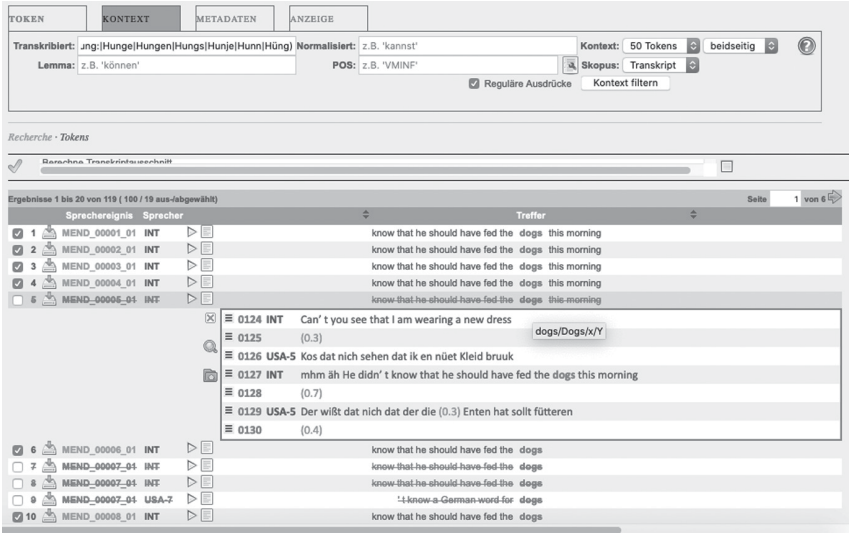


Abbildung 17: Filtern der KWIC-Liste *(Keywords in Context)* nach der Tokensuche und beispielhafte Anzeige eines Transkriptabschnitts

Als Ergebnis werden von den ursprünglich 119 Treffern 100 aus- und 19 abgewählt. Dies bedeutet, dass im Kontext von *dogs* 19-mal keine Übersetzung in einer der Formen von *Hunde* vorkommt. Diese Treffer lassen sich nun genauer betrachten. Es stellt sich heraus, dass einige Male *dogs* mit *Enten* übersetzt wurde, woraus man entweder schließen könnte, dass einige Probanden statt *dogs ducks* verstanden oder, was ebenso wahrscheinlich ist, dass der Interviewer Göz Kaufmann die Auslautverhärtung des Deutschen aufs Englische übertrug. Zu entscheiden, ob es sich hierbei um ein Produktions- oder ein Perzeptionsproblem handelt, wäre ein durchaus interessantes Unterfangen. Je nachdem, welches Phänomen eine/n Forschende/n interessiert, kann die Suche weiter eingeschränkt werden. Interessiert z.B. die Syntax, wäre es eigentlich egal, ob Hunde oder Enten gefüttert werden, und man könnte die Formen für *Enten* (Ent|Ente|Enten) in den Filter mit einbauen. Übrig bleibt dann genau ein Treffer, bei dem es ein Problem beim lexikalischen Zugriff gab: „*I don't know a German word for dogs*." Weitere Funktionen, welche die Trefferliste nach dem Filtern übersichtlicher machen, sind das Invertieren des Filters ( ) und das Löschen der abgewählten Treffer ( ). Zur weiteren Verarbeitung kann das Suchergebnis auf verschiedene Weisen verwendet werden:

- Analog zum Teilen von virtuellen Korpora kann auch das Suchergebnis gespeichert ( ) und mit anderen DGD-Nutzer/innen/n über eine ID geteilt werden ( )
- Als Tabelle kann das Ergebnis im Excel-Format ( ) oder als tabulatorseparierte Textdatei ( ) heruntergeladen werden
- Das Ergebnis kann auch quantifiziert werden ( ), wie wir es beim Erstellen von Tabelle 1 getan haben

Wie bei anderen Lexemen gibt es auch bezüglich *Hund* einen weiteren Stimulussatz, nämlich Satz <46> *I should have shown the little dog to the kids*, in dem *dog* im Singular vorkommt. Auch hier wurde *dog* manchmal mit *Ente*, also *duck* (im Plautdietschen *Ent*) verwechselt. Interessant ist auch die Verwechslung mit *doc*, denn einige Male wurde der Satz mit *Doktor* übersetzt.[26]

26 Als Übung kann der/die interessierte Nutzer/in ein ähnliches Phänomen in Stimulussatz <36> *The doctor who wants to see my foot is very worried* durchspielen. Manche Gewährspersonen haben hier beim Wort *foot* (‚Fuß') *food* (‚Essen') verstanden. Entweder ist dies wieder ein Perzeptionsproblem, oder der Interviewer war wieder in seiner Aussprache unklar, indem er eine Hyperkorrektur zur Vermeidung der Auslautverhärtung durchführte.

Wenn man nun das gesamte Prozedere umkehrt und nach allen Tokens sucht, die etwas mit der Übersetzungsaufgabe bezüglich *Hund* oder *Hunde* zu tun haben, so kann man zunächst nach dem Ausdruck

(H:und|H:und-|H:ung|Hund|Hund-|Hund:|Hunde|Hundje|Hundje-|Hundjes|Hundke|Hunds|Huntjes|Hung|Hung-|Hung:|Hunge|Hungen|Hungs|Hunje|Hunn|Hüng|Ent|Ente|Enten|Doktor)

suchen und im Anschluss die KWIC-Liste (*Keywords in Context*) jeweils anhand der unterschiedlichen stimulussprachlichen Kontexte *dog*, *dogs*, *perrito*, *perros*, *cachorrinho(s)*, *cachorros* filtern. Wenn man das Ergebnis quantifiziert, gelangt man zu Tabelle 1:

Wort \ Kontext	dog	dog normalisiert	dogs	dogs normalisiert	perrito	perrito normalisiert	perros	perros normalisiert	cachorinho(s)	cachorinho(s) normalisiert	cachorros	cachorros normalisiert	total (ohne normalisiert)	total (normalisiert)
H:und			1	1	1	1			1	1			3	3
H:und-											1	1	1	1
H:ung			1										1	0
Hund	**80**	**81**	**20**	**81**	**100**	**101**	**12**	**161**	**25**	**26**		**55**	**237**	**505**
Hund-	3	3			6	6	2	2	2	2			13	13
Hund:	1	1			1	1			1	1	1	1	4	4
Hunde							1	1			1	2	2	3
Hundje					**69**				**32**				**101**	**0**
Hundje-					1	1			1	1			2	2
Hundjes	1				3								4	0
Hundke					1	1							1	1
Hunds											1	1	1	1
Hung	**1**		**60**		**1**		**149**		**1**		**55**		**267**	**0**
Hung-			2	2			2	2					4	4
Hung:							2	2			1	1	3	3
Hunge											1		1	0
Hungen			2	2									2	2
Hungs							1	1					1	1
Hunje					1	1							1	1
Hunn	1	1			1	1	2	2			1	1	5	5
Hüng			1	1									1	1
Hündchen		1				72				32			0	105
Ent	11	11											11	11
Ente			2	2									2	2
Enten			12	12									12	12
Doktor	6	6											6	6
total	104	104	101	101	185	185	171	171	63	63	62	62	686	686

Tabelle 1: Vorkommnisse von *Hund* bzw. *Hunde* im Kontext der jeweiligen ausgangssprachlichen Stimuli[27]

27 Hervorgehoben sind die zentralen Formen von *Hund* (Singular, Plural und Diminutiv). Aus Gründen der Transparenz haben wir alle transkribierten Formen aufgelistet, wie sie die DGD liefert. Ein/e Forscher/in kann auf dieser Basis aktiv entscheiden, die anderen Formen in die eine oder die andere Kategorie einzuteilen. Normalisiert bedeutet, dass das Wort in der linken Spalte so oft in normalisierter Form vorkommt. Zum Beispiel wurde im Kontext *dogs* das Wort 20-mal mit *Hund* übersetzt, 60-mal mit *Hung* (Plural) usw. In normalisierter Form kommt *dogs* allerdings 81-mal als *Hund* (Singular) und nie als *Hung* (Plural) vor. Tabelle 1 zeigt also einerseits die Komplexität der Normalisierungsaufgabe und andererseits auch die Fehler und Lücken, die es noch zu beheben gilt (vgl. Abschnitt 6).

Aus Tabelle 1 lässt sich schließen, dass die Singularform *Hund* im Plautdietschen auch *Hund* ist, die Pluralform *Hunde Hung* und der Diminutiv *Hündchen Hundje*. Da der englische Stimulussatz *little dog* ('kleiner Hund') verwendet, kommen plautdietsche Diminutivformen in Übersetzungen fast ausschließlich dann vor, wenn sie auf spanischen oder portugiesischen Stimulussätzen fußen. Diese verwenden nämlich *perrito* und *cachorrinho* ('Hündchen'). Hier existiert also ein klarer Einfluss der Stimulussprache (vgl. Fußnote 8). Auch gibt es weitere Realisierungsvarianten, was in den unterschiedlich transkribierten Formen und den zusätzlichen Symbolen der literarischen Umschrift zu erkennen ist (':' bezeichnet eine Segmentdehnung, '-' eine(n) Reparatur/Abbruch; vgl. die Fußnoten 14 und 15).

Ein Vorteil der literarischen Umschrift ist, dass auffällige phonetische Variationen zumindest teilweise bereits transkribiert sind. Eine auditive Kontrolle der Aussprache ist aber trotzdem nötig, da sie zum Zeitpunkt der Verschriftung der Daten nicht im Zentrum stand.[28] Dennoch lässt sich z.B. nach der normalisierten Form *ich* suchen und man erhält eine quantifizierte Übersicht:

Transkribierte Formen
ik (4214) ; Ik (2158) ; ich (148) ; Ich (140) ; ik: (20) ; Ich::- (1) ; I:k: (1)[29]

Normalisierte Formen
ich (6682)

Abgesehen von den Spezialfällen *ik:*, *Ich::-* und *I:k:* lässt sich zwischen *ik* bzw. *Ik* und der standarddeutschen Form *ich* bzw. *Ich* unterscheiden. Eine stichprobenartige Analyse zeigt, dass *ich* oft in standarddeutschen Kommentaren zu Übersetzungen vorkommt, also nicht in den Übersetzungen selbst. Zumeist kommt es also vor, wenn die Probanden denken, dass sie ins Standarddeutsche übersetzen sollen (vgl. Abbildung 4). Manchmal erscheint *ich* aber auch in plautdietschen Übersetzungen, was einen Extremfall der bei den Übersetzungen (2a-c) angesprochenen Palatalisierung von /*k*/ in *ik* darstellen könnte.

---

28 Interessierte Nutzer/innen können ausgewählte Beispiele aus der KWIC (*Keywords in Context*) herunterladen (). Neben den Metadaten sind auch Audio und Transkript in einer bereitgestellten Zip-Datei enthalten. Wenn man z.B. das Format zu Praat TextGrid einstellen will, kann man das tun, indem man den Treffer der KWIC-Liste aufklappt, () und den Ausschnitt in einem Browser-Tab öffnet (). Dort kann man dann die Download-Optionen einstellen und den Vorgang des Downloads abschließen ().

29 Die transkribierte Form *isch* kommt in dieser Aufzählung nicht vor, da sie nicht als *ich* normalisiert wurde und daher nicht gefunden werden konnte. Von *isch* gibt es im Korpus genau einen Treffer in der standardnahen Übersetzung von Abbildung 4.

6. Zusammenfassung und Ausblick

MEND als Korpus mit elizitierten Daten einer Minderheitensprache hat für die DGD einen besonderen, eher experimentellen Status. Bislang fokussiert die DGD ihre Recherchemöglichkeiten auf den prototypischen Fall eines Gesprächskorpus, also auf Sammlungen natürlicher Interaktionsdaten, die gerade nicht die Systematik und Vergleichbarkeit der MEND-Daten aufweisen. Andererseits kann der Aufbereitungszustand des MEND-Korpus noch nicht als komplett angesehen werden – es fehlen z.B. noch die Lemmatisierung und das POS-Tagging der Transkript-Daten[30] – und daher kann die Breite an Analysefunktionen der DGD in MEND noch nicht voll ausgeschöpft werden. Eine bessere sprach- und texttechnologische Unterstützung für Minderheitensprachen wie das Plautdietsche bleibt ein Desiderat, exemplarische Unternehmungen wie die hier beschriebene können hierfür ein erster Schritt sein.

Für die Frage der Nachhaltigkeit und Nachnutzbarkeit aufwändig erhobener Daten ist das MEND-Korpus beispielhaft. Über das AGD werden die aufbereiteten Daten nun dauerhaft bewahrt und können über den ursprünglichen Forschungskontext hinaus von anderen Forscher/inne/n genutzt werden. Der nicht unbeträchtliche Aufbereitungsaufwand könnte künftig deutlich reduziert werden, wenn Daten von vornherein digital erhoben (was heute eine Selbstverständlichkeit sein dürfte, zum Zeitpunkt der Erhebung 1999–2002 aber noch kaum praktikabel war) und mit Werkzeugen wie EXMARaLDA (oder auch FOLKER, Praat, ELAN) erschlossen werden. Dies trifft auf das oben erwähnte und vergleichbare Forschungsprojekt zum Pomerano Brasiliens zu (vgl. Fußnote 13), bei dem alle Übersetzungen bzw. alle freien Gespräche digital aufgenommen wurden und alle Transkriptionen in EXMARaLDA durchgeführt werden. Auch dieses Korpus wird vom AGD übernommen und in die DGD integriert werden. Abschließend sei daher noch einmal betont, dass das AGD sehr gerne bereit ist, bei einer möglichen Archivierung bestehender oder einer *Best-Practice*-konformen Erhebung neuer Daten zu beraten.

30 Diese würden wiederum – wie die Pilotierung der orthographischen Normalisierung – eine eingehende Auseinandersetzung mit den Daten erfordern, sowohl auf konzeptueller Ebene (Welche Lemmata und POS sind für das Plautdietsche anzusetzen?) als auch in Form von manueller Annotationsarbeit (Trainingsdaten für den Tagger, Lemmatisierungslexikon). So sinnvoll und lohnend (auch mit Blick auf andere Varietäten) diese Aufgabe wäre, das AGD kann sie aus eigenen Mitteln nicht bestreiten. Wir sind für diesbezügliche Kooperationen aber offen.

Bibliografie

Bangalore, Srinivas / Behrens, Bergljot / Carl, Michael / Ghankot, Maheshwar / Heilmann, Arndt / Nitzke, Jean / Schaeffer, Moritz / Sturm, Annegret (2016): „Syntactic variance and priming effects in translation". In: Carl, Michael / Bangalore, Srinivas / Schaeffer, Moritz (Hrsg.): *New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB*. Cham: Springer, 211–238.

Barbiers, Sjef / Bennis, Hans / de Vogelaer, Gunther / Devos, Magda / van der Ham, Margreet (2005): *SAND – Syntactische Atlas van de Nederlandse Dialecten* (Band II). Amsterdam: Amsterdam University Press.

Blevins, Margo (2019): „Orthographic Normalization of Language Contact Data". In: *German(ic) in language contact: Grammatical and sociolinguistic dynamics: Book of Abstracts*. Berlin: Freie Universität, 19–21. [https://www.geisteswissenschaften.fu-berlin.de/v/namdeutsch/Workshop/Programme/book_of_abstracts_GiLC_190701.pdf; zuletzt aufgerufen am 14.06.2021].

Bucheli, Claudia / Glaser, Elvira (2002): „The syntactic atlas of Swiss German dialects: Empirical and methodological problems". In: Barbiers, Sjef / Cornips, Leonie / van der Kleij, Susanne (Hrsg.): *Syntactic Microvariation*. Amsterdam: Meertens Institute (http://www.meertens.knaw.nl/books/synmic; zuletzt aufgerufen am 14.06.2021), 41–74.

Clyne, Michael (1981): *Deutsch als Muttersprache in Australien. Zur Ökologie einer Einwanderersprache*. Wiesbaden: Franz Steiner.

Collins, Chris / Postal, Paul (2014): *Classical NEG Raising: An Essay on the Syntax of Negation*. Cambridge, MA: MIT Press.

Deppermann, Arnulf (2008): *Gespräche analysieren: Eine Einführung*. Wiesbaden: Verlag für Sozialwissenschaften.

Dorian, Nancy C. (1981): *Language Death: The Life Cycle of a Scottish Gaelic Dialect*. Philadelphia: University of Pennsylvania Press.

Engel, Ulrich / Vogel, Irmgard (Hrsg.) (1975): *Gesprochene Sprache. Bericht der Forschungsstelle Freiburg*. Tübingen: Narr.

Fandrych, Christian / Meißner, Cordula / Slavcheva, Adriana (2012): „The GeWiss corpus: Comparing spoken academic German, English and Polish". In: Schmidt, Thomas / Wörner, Kai (Hrsg.): *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam/Philadelphia: John Benjamins, 319–338.

Fleischer, Jürg / Lenz, Alexandra N. / Weiß, Helmut (2015): „Syntax hessischer Dialekte (SyHD)". In: Kehrein, Roland / Lameli, Alfred / Rabanus, Stefan (Hrsg.): *Regionale Variation des Deutschen*. Berlin/New York: de Gruyter, 261–287.

Frick, Elena / Schmidt, Thomas (2020): „Using full text indices for querying spoken language data". In: Bański, Piotr / Barbaresi, Adrien / Clematide, Simon / Kupietz, Marc / Lüngen, Harald / Pisetta, Ines (Hrsg.): *Proceedings of the LREC 2020 Workshop, Language Resources and Evaluation Conference, 11–16 May 2020, 8th Workshop on Challenges in the Management of Large Corpora (CMLC-8)*. Paris: European Language Resources Association, 40–46.

Gorisch, Jan / Schmidt, Thomas / Stift, Ulf-Michael (2022): „Data of German speech minorities in the Archive for Spoken German: An overview". In: Boas, Hans C. (Hrsg.): *Comparative Language Island Research: Data, Methods, and Goals*. Leiden: Brill.

Götze, Angelika / Lindenfelser, Siegwalt / Lipfert, Salome / Neumeier, Katharina / König, Werner / Maitz, Peter (2018): „Documenting Unserdeutsch (Rabaul Creole German): A workshop report". In: Maitz, Péter / Volker, Craig A. (Hrsg.): *Language and Linguistics in Melanesia* (Themenheft des *Journal of the Linguistic Society of Papua New Guinea*), 65–90.

Hill, Jane / Hill, Kenneth (1977): „Language Death and Relexification in Tlaxcalan Nahuatl". *Linguistics* 191, 55–68.

Huffines, Marion Lois (1991): „Acquisition Strategies in Language Death". *Studies in Second Language Acquisition* 13, 43–55.

Kaufmann, Göz (1997): *Varietätendynamik in Sprachkontaktsituationen: Attitüden und Sprachverhalten rußlanddeutscher Mennoniten in Mexiko und den USA*. Frankfurt a.M.: Peter Lang.

Kaufmann, Göz (2003): „The verb cluster in Mennonite Low German". In: Mattheier, Klaus J. / Keel, William (Hrsg.): *German Language Varieties Worldwide: Internal and External Perspectives*. Frankfurt a.M.: Peter Lang, 177–198.

Kaufmann, Göz (2004): „Eine Gruppe – Zwei Geschichten – Drei Sprachen: Rußlanddeutsche Mennoniten in Brasilien und Paraguay". *Zeitschrift für Dialektologie und Linguistik* 71(3), 257–306.

Kaufmann, Göz (2005): „Der eigensinnige Informant: Ärgernis bei der Datenerhebung oder Chance zum analytischen Mehrwert?" In: Lenz, Friedrich / Schierholz, Stefan (Hrsg.): *Corpuslinguistik in Lexik und Grammatik*. Tübingen: Stauffenberg, 61–95.

Kaufmann, Göz (2007): „The Verb Cluster in Mennonite Low German: A new approach to an old topic". *Linguistische Berichte* 210, 147–207.

Kaufmann, Göz (2008): „Where Syntax meets Morphology: Varianten des bestimmten Artikels und die Variation satzfinaler Verbcluster im Plattdeutschen texanischer Mennoniten". In: Patocka, Franz / Seiler, Guido (Hrsg.): *Dialektale Morphologie, dialektale Syntax: Beiträge zum 2. Kongress der IGDD*. Wien: Praesens, 87–119.

Kaufmann, Göz (2011): „Looking for Order in Chaos: Standard convergence and divergence in Mennonite Low German". In: Putnam, Mike (Hrsg.): *Sudies on German-Language Islands*. Amsterdam/Philadelphia: John Benjamins, 187–230.

Kaufmann, Göz (2015): „Rare phenomema revealing basic syntactic mechanisms: The case of unexpected verb-object-sequences in Mennonite Low German". In: Adli, Aria / García García, Marco / Kaufmann, Göz (Hrsg.): *Variation in Language: System- and Usage-Based Approaches*. Berlin/Boston: de Gruyter, 113–146.

Kaufmann, Göz (2016): *The World beyond Verb Clusters: Aspects of the Syntax of Mennonite Low German*. (Habilitationsschrift) (http://paul.igl.uni-freiburg.de/kaufmann/userfiles/downloads/Formatiert%20(Englisch).pdf; zuletzt aufgerufen am 18.02.2022).

Kaufmann, Göz (2017): „‚Sorvete und Tema is nich Dütsch': Zur Integration portugiesischer Lehnwörter in drei deutschen Varietäten Südbrasiliens". In: Eller-Wildfeuer, Nicole / Maitz, Péter / Wildfeuer, Alfred (Hrsg.): *Sprachkontaktforschung – explanativ* (ZDL-Themenheft). Stuttgart: Steiner, 260–307.

Kaufmann, Göz (2018a): „Mennonitenplautdietsch in Nord- und Südamerika (MEND)". In: Institut für Deutsche Sprache (Hrsg.): *Archiv für Gesprochenes Deutsch* (http://agd.ids-mannheim.de/MEND_extern.shtml; zuletzt aufgerufen am 14.06.2021).

Kaufmann, Göz (2018b): „Relative markers in Mennonite Low German: Their forms and functions". In: Speyer, Augustin / Rauth, Philipp (Hrsg.): *Syntax aus Saarbrücker Sicht 2: Beiträge der SaRDiS-Tagung zur Dialektsyntax* (ZDL-Beihefte). Stuttgart: Steiner, 109–148.

Kaufmann, Göz (2022): „In the thick of it: scope rivalry in past counterfactuals of Pomerano". *Journal of Comparative Germanic Linguistics* 25(3), 333–384. https://doi.org/10.1007/s10828-022-09137-9.

Kleiner, Stefan (2015): „‚Deutsch heute' und der Atlas zur Aussprache des deutschen Gebrauchsstandards". In: Kehrein, Roland / Lameli, Alfred / Rabanus, Stefan (Hrsg): *Regionale Variation des Deutschen: Projekte und Perspektiven*. Berlin/Boston: de Gruyter, 489–518.

Labov, William / Ash, Sharon / Boberg, Charles (2005): *The Atlas of North American English: Phonetics, Phonology and Sound Change*. Berlin/Boston: de Gruyter.

Maitz, Péter / König, Werner / Volker, Craig A. (2016): „Unserdeutsch (Rabaul Creole German): Dokumentation einer stark gefährdeten Kreolsprache in Papua-Neuguinea". *Zeitschrift für germanistische Linguistik* 44(1), 93–96.

Mathussek, Andrea (2016): „On the problem of field worker isoglosses". In: Côté, Marie-Hélène / Knooihuizen, Remco / Nerbonne, John (Hrsg.): *The Future of Dialects: Selected Papers from Methods in Dialectology* XV. Berlin: Language Science Press, 99–115.

Milroy, Lesley / Gordon, Matthew (2003): *Sociolinguistics: Method and Interpretation*. Malden/MA: Blackwell.

Schmidt, Thomas (2018): „Gesprächskorpora". In: Kupietz, Marc / Schmidt, Thomas (Hrsg.): *Korpuslinguistik* (Band 5). Berlin/Boston: de Gruyter, 209–230.

Schütte, Wilfried / Schmidt, Thomas (2017): „Niederdeutsche Aufnahmen aus dem Korpus ‚Deutsche Mundarten': Sprachliche und kulturhistorische Aspekte". *Sprachreport* 2017(1), 8–19.

Siemens, Heinrich (2012): *Plautdietsch: Grammatik, Geschichte, Perspektiven*. Bonn: Tweeback.

Stift, Ulf-Michael / Schmidt, Thomas (2014): „Mündliche Korpora am IDS: Vom Deutschen Spracharchiv zur Datenbank für Gesprochenes Deutsch". In: Institut für Deutsche Sprache (Hrsg.): *Ansichten und Einsichten. 50 Jahre Institut für Deutsche Sprache*. Mannheim: Institut für Deutsche Sprache, 360–375.

Winterscheid, Jenny / Deppermann, Arnulf / Schmidt, Thomas / Schütte, Wilfried / Schedl, Evi / Kaiser, Julia (2019): *Normalisieren mit OrthoNormal. Konventionen und Bedienungshinweise für die orthografische Normalisierung von FOLKER-Transkripten*. Mannheim: Institut für Deutsche Sprache. [https://doi.org/10.14618/ids-pub-9326; zuletzt aufgerufen am 14.06.2021].

Wurmbrand, Susi (2017): „Verb clusters, verb raising, and restructuring". In: Everaert, Martin / Riemsdijk, Henk van (Hrsg.): *The Blackwell Companion to Syntax*. Oxford: Wiley-Blackwell, 1–109.

Zimmer, Christian / Wiese, Heike / Simon, Horst J. / Zappen-Thomson, Marianne / Bracke, Yannic / Stuhl, Britta / Schmidt, Thomas (2020): „Das Korpus Deutsch in Namibia (DNam): Eine Ressource für die Kontakt-, Variations- und Soziolinguistik". *Deutsche Sprache* 2020(3), 210–232.

Zwirner, Eberhard / Bethge, Wolfgang (1958): *Erläuterungen zu den Texten. Spracharchiv, Deutsches: Lautbibliothek der deutschen Mundarten*. Göttingen: Vandenhoeck & Ruprecht.