[WLG]
WIENER LINGUISTISCHE GAZETTE

Annotation von Sprachdaten eines variationslinguistischen Großprojekts am Beispiel des Spezialforschungsbereichs ›Deutsch in Österreich‹

Markus Pluschkovits/Katharina Kranawetter

Special print from: *Wiener Linguistische Gazette* (WLG) 89 (2021): 167–189

University of Vienna · Department of Linguistics · 2021

Owner, editor and publisher:
University of Vienna, Department of Linguistics
Sensengasse 3a
1090 Vienna
Austria

Editorial board: Markus Pöchtrager (General Linguistics),
Mi-Cha Flubacher & Florian Grosser (Applied Linguistics),
Stefan Schumacher (Historical Linguistics)
Contact: wlg@univie.ac.at
Homepage: *http://wlg.univie.ac.at*

ISSN: 2224-1876
NBN: BI,078,1063

The WLG journal is published in loose sequence and in open access format.
All issues as of 72 (2005) are available online.

Annotation von Sprachdaten eines variationslinguistischen Großprojekts am Beispiel des Spezialforschungsbereichs ›Deutsch in Österreich‹

Markus Pluschkovits/Katharina Kranawetter*

Wiener Linguistische Gazette (WLG)
Department of Linguistics
University of Vienna
Issue 89 (2021): 167–189

Abstract

The present paper is concerned with the annotation of linguistic phenomena from a variationist linguistic perspective. The Special Research Programme ›German in Austria‹ serves as example for annotation in a large-scale variationist research project. After contextualizing both the corpus and the annotation logic used in the project, we situate the annotation system in a linguistic philosophy of science, focusing on the epistemological status of the classification of spoken language. During this discussion, the multidimensional annotation system used in the project is highlighted for its advantages. We conclude with a discussion of standardized

* Markus Pluschkovits, Zentrum für Translationswissenschaft, Gymnasiumstraße 50, 1190 Wien, markus.pluschkovits@univie.ac.at (Corresponding author).

annotation schemes in a variationist context, and stress the importance of publishing not just results, but annotation schemes as well in the spirit of open science.

Keywords: annotation, tagging, variationist linguistics, corpus linguistics, data processing

1 Vorwort

In der vorliegenden Arbeit[1] wollen wir den Fokus auf ein Element linguistischer Forschung legen, dem oftmals zu wenig Beachtung geschenkt wird, obwohl es teilweise das Gros der geleisteten Arbeit ausmacht und für die Weiternutzung der Primärforschungsdaten von essenzieller Bedeutung ist – die Annotation von Sprachdaten.

Bei der linguistischen Annotation werden Sprachdaten mit beschreibenden bzw. analytischen Vermerken versehen (vgl. Ide 2017: 2). Diese Annotationen werden dazu genützt, dem Korpus Informationen hinzuzufügen und sind daher essenziell für eine weitere Verarbeitung und Analyse von Sprachdaten (vgl. Newman & Cox 2021: 25). Besonders in Großprojekten kommen verschiedene Anforderungen an die Funktionalität eines Annotationssystems hinzu. Der Spezialforschungsbereich ›Deutsch in Österreich: Variation – Kontakt – Perzeption‹ (F 60, im Folgenden SFB oder SFB:DiÖ) beschäftigt sich mit verschiedenen Ebenen variationslinguistischer Forschung und hat somit ganz eigene Anforderungen an ein Annotationssystem, denen Standardsysteme oft nicht gerecht werden können.

Daher soll sich diese Arbeit vor allem der Annotation in variationistischen elektronischen Korpora widmen. Im Folgenden wird anhand der Sprachdaten, auf deren Annotationssystem kurz eingegangen wird, der epistemologische Status phänomenbezogener Annotation innerhalb der Variationslinguistik näher beleuchtet. Es wird auch der Status standardisierter Annotationsparadigmen betrachtet, insbesondere im Kontext

1 Wir möchten uns ganz herzlich bei den beiden Gutachtern, Christian Huber und Ludwig Maximilian Breuer, für ihre Anregungen und Vorschläge bedanken, die uns sehr geholfen haben.

des heterogenen Forschungsgegenstands. Ein kurzes Plädoyer für die Veröffentlichung von Annotationsschemata im Sinne der Open-Science-Bewegung bildet den Abschluss.

Der vorliegende Artikel zielt im Besonderen darauf ab, sich den Fragen zu stellen, wie in einem großen Forschungsprojekt wie dem SFB Annotation betrieben wird, wie ein dafür geeignetes Annotationssystem aussieht und funktioniert und wie es den Ansprüchen eines solchen Großprojekts gerecht werden kann. Des Weiteren wird darauf eingegangen, wie sich die Vorgehensweise des SFB in die Annotationsstandards der Variationslinguistik einordnen lässt und welche Erkenntnisse durch die Annotation gewonnen werden können.

2 Annotation im SFB:DiÖ – die Ordnung der Dinge

Bevor wir uns der Annotation der Daten zuwenden, soll zunächst gesagt sein, um welche Daten es sich dabei handelt. Der SFB:DiÖ untersucht – an sechs Instituten verteilt auf die Universitäten Graz, Salzburg und Wien sowie der Österreichischen Akademie der Wissenschaften – vor allem die Variation und Perzeption verschiedener Varietäten des Deutschen in Österreich, sowie deren Kontakt mit anderen Sprachen (für einen Überblick über den SFB siehe auch Budin et al. 2018). Die konkreten Forschungsanliegen sind dabei vielfältig und erstrecken sich von historischem und gegenwärtigem Sprachkontakt über phonologische, syntaktische und morphologische Phänomene in verschiedenen Varietäten sowie lexikalische Variation und Spracheinstellungsforschung. Notwendigerweise werden dafür auch verschiedenste Arten von Daten gesammelt, die unter anderem Sprachproduktionstests, Aufnahmen von Interviews, Freundesgesprächen, Lesetexten und Übersetzungsaufgaben, aber auch Fragebögen und Ähnliches umfassen.

Ein Großteil dieser Daten wird dabei in einer zentralen Datenbank erfasst. Durch die Heterogenität der Daten werden diese dabei unterschiedlich behandelt: Sprachaufnahmen wie Interviews von Gewährspersonen, die Freundesgespräche, die in Abwesenheit der Explorator:innen aufgenommen wurden, oder kurze Lesetexte (›Nordwind und Sonne‹) liegen dabei als zeitalignierte Transkripte vor. Sprachaufnahmen aus

kontrollierteren Settings wie den Sprachproduktionstests oder Übersetzungsaufgaben wurden zwar ebenfalls zu weiten Teilen transkribiert (für weiteres zu den Sprachproduktionstests im SFB:DiÖ siehe exemplarisch Fingerhuth & Breuer 2020), liegen aber als sogenannte ›Aufgaben‹ vor, was bedeutet, dass die meist relativ kurzen *responses* zu der entsprechenden Aufgabe gespeichert wurden, und sich diese über ein eigenes Interface abrufen lassen. Obwohl auch diese *responses* größtenteils transkribiert wurden, unterscheiden sie sich von den Transkriptdaten darin, was deren kleinste annotierbare Einheit ist. Während in den Transkriptdaten die kleinste annotierbare Einheit das Einzelwort ist – sprich, hier auf Wortebene tokenisiert wurde – ist bei den *responses* zu Sprachproduktionsexperimenten oder Übersetzungsaufgaben eine Antwort – das entspricht hier einem Satz oder Halbsatz – die kleinste annotierbare Einheit. Die *responses* wurden daher nicht auf Wortebene tokenisiert, sondern werden als sogenannte Antwort erfasst.

Streng genommen ist dabei die Antwort innerhalb der Datenbank die einzige annotierbare Einheit. Die Speicherlogik der Datenbank des SFB erlaubt es nicht, Annotationen direkt an ein Worttoken zu vergeben, sondern erstellt stattdessen eine sogenannte Antwort, die genau dieses Token enthält. Die Antwort, die mit diesem Token verknüpft ist, erhält dann die Annotation. Jegliche Annotation zu diesem Token ist daher nur über die Antwort mit dem Token assoziiert. Dies hat verschiedene Gründe, unter anderem erlaubt es, das gleiche technische System zur Annotation sowohl für Tokens aus den Transkriptdaten als auch für die *responses* zu kontrollierten Aufgaben zu verwenden. Gleichzeitig bietet dieser Zwischenschritt über Antworten auch den Vorteil, dass innerhalb einer Antwort mehrere Wort-Tokens zusammengefasst werden können, sprich, dass etwa bei der Annotation einer Phrase nicht ein einzelnes Worttoken die Annotation erhält, sondern stattdessen alle Tokens der Phrase in einer Antwort zusammengefasst werden können, und alle diese Tokens über die mit ihnen assoziierte Antwort die Annotation erhalten, diese aber bei einer Suche trotzdem nur als ein Ergebnis (da nur eine Antwort) aufscheinen. Diese Datenbankarchitektur geht auf das Dissertationsprojekt Breuer (2021, siehe insbesondere Kap. 2.2.5) zurück.

Es wurde bereits erwähnt, dass die Sprachdaten aus den freieren Settings und den Lesetexten größtenteils zeitalligniert transkribiert wurden, es soll allerdings darauf hingewiesen werden, dass innerhalb der verschiedenen Forschungskontexte des SFB für das Gesamtkorpus kein einheitliches Transkriptionssystem verwendet wurde. Prinzipiell ist den Transkriptdaten gleich, dass sie drei verschiedene Tiers zur Transkription anbieten (wobei eines davon als *default-tier* gilt, das unbedingt gefüllt werden muss, damit eine Transkription vorliegt, während die anderen beiden nur im Bedarfsfall verwendet werden), mit welchem Transkriptionssystem jedoch im *default-tier* transkribiert wurde, unterscheidet sich je nach Teilprojekt des SFB. Ein großer Teil wurde dabei orthographisch normalisiert transkribiert, sprich in das Standarddeutsche übertragen, wobei besonders salienter Dialektgebrauch oder andere Auffälligkeiten zusätzlich in einem eigenen Tier mittels literarischer Umschrift transkribiert wurden. Ein drittes Tier steht zusätzlich für phonetische Transkription mittels IPA zur Verfügung und wird im Bedarfsfall verwendet. Andere Teilprojekte nutzen für die Transkription im *default-tier* stark an GAT2 angelegte Transkriptionskonventionen. Auf die Herausforderungen, die sich durch die Verwendung verschiedener Transkriptionssysteme innerhalb eines Großprojekts ergeben, kann im Folgenden leider nicht detailliert eingegangen werden, es lohnt sich allerdings ein Spezifikum herauszustellen: den Umgang mit Interpunktionen.

Innerhalb von Transkription ist der Umgang mit Interpunktion teilweise sehr heikel, insbesondere wenn die Transkription nicht zeitalligniert ist bzw. das zugehörige Audiomaterial nicht weitergegeben werden darf, und so Interpunktion benutzt wird, um beispielsweise Intonation zu markieren. Die Interpretation solcher Interpunktion kann allerdings sehr verschieden ausfallen (vgl. Nagy & Sharma 2013: 242). Auch im SFB wird Interpunktion in den beiden Hauptformen der Transkription (sowohl standardorthographisch als auch an GAT2 angelehnt) verwendet, allerdings wird diese vom verwendeten Transkriptionstool[2]

2 Dieses Transkriptionstool, das momentan den Arbeitstitel ›Transcribe‹ trägt, wurde innerhalb des SFB auf Open-Source-Basis entwickelt und wird demnächst auch

automatisch mittels eines Type-Token-Parsers interpretiert, der für entsprechende Interpunktion die Eigenschaften des betreffenden Tokens ändert. Das Token ⟨.⟩, das beispielsweise fallende Intonation signalisieren kann, wird von der Transkriptionssoftware als ›delimiter‹, also Begrenzer einer Intonationsphrase erkannt, und dementsprechend anders dargestellt. Besonders relevant dabei ist, dass auch zu Anonymisierendes eindeutig markiert – und dadurch in der weiteren Verarbeitung der Daten ausgeblendet – werden kann (durch die Verwendung eckiger Klammern). Auch Klitika lassen sich so leicht auflösen – ein Klitikon wie *gemma* (›gehen wir‹) würde dabei als ⟨ge_ _ma⟩ transkribiert werden, wobei die Unterstriche markieren, dass diese beiden Einzeltoken eigentlich Fragmente einer Einheit sind – und dementsprechend graphisch als zueinander gehörig markiert werden. Den beiden Tokens, die aus dem Klitikon entstehen, wird als Eigenschaft zugewiesen, dass sie Fragmente zueinander sind. Dieser Type-Token-Parser ist dabei flexibel, und kann mithilfe von Regular Expressions an verschiedene Transkriptionskonventionen angepasst werden.

Obwohl eine zeitalignierte Transkription eigentlich bereits eine Form der Annotation darstellt (vgl. Bird & Libermann 2001: 24), ist im Folgenden Annotation im engeren Sinne gemeint – also Annotation im Sinne von »information/elements added to provide specifically linguistic/grammatical/structural information such as part of speech, semantics, pragmatics, prosody, interaction, and many others« (Gries & Berez 2017: 382). Wir grenzen dabei noch weiter ein, und verstehen unter Annotation im engeren Sinne damit nur manuell vergebene Tags auf Antworten, und schließen damit Transkription, aber auch Informationen, die als Metadaten direkt zu den Tokens gespeichert werden, aus. Diese im Rahmen des SFB vergebenen Annotationen dienen dabei vor allem zur Klassifikation von sprachlichen Variationsphänomenen.

Als Konsequenz werden im Folgenden auch die Part-of-Speech-Tags nicht näher behandelt, da diese einerseits durch ihre teilweise relativ weit

anwender:innenfreundlich veröffentlicht. Für vor allem die technischen Aspekte von Transcribe siehe Graf et al. (in Vorbereitung), für den Code das entsprechende GitHub-Repositorium (DiÖ 2021: transcribe).

vorangeschrittene Standardisierung und Unabhängigkeit von einzelnen linguistischen Phänomenen eine qualitativ andere Betrachtung erfordern (die den gegebenen Rahmen sprengen würde), andererseits aber diese innerhalb der Datenbank des SFB als Information zu einzelnen Tokens direkt gespeichert werden und nicht als eigentliches Tag. Das bedeutet, dass in der Speicherlogik die Part-of-Speech-Klassifikation kein genuines Tag, das vergeben wurde, darstellt, sondern lediglich eine Zusatzinformation, die in der gleichen Tabelle wie die Tokens gespeichert wird – im Gegensatz zu manuell vergebenen Annotationen, die sich auf Antworten beziehen. Im Folgenden wird die Logik dieser Annotationen im engeren Sinne, wie sie im SFB verwendet werden, expliziert, und es wird gezeigt, wieso das Etablieren eines Annotationssystems bereits elementare Forschungsarbeit ist. Den Abschluss bilden Überlegungen zu Annotationsstandards in der Variationslinguistik und ein Plädoyer für die Offenlegung von Annotationssystemen aus der Perspektive der Open-Science-Bewegung.

2.1 SFB-Annotationssystem

In ihrer simpelsten Form erfüllen Annotationen – sowohl innerhalb des SFB als auch in der Korpuslinguistik generell – die einfache Aufgabe, Linguist:innen zu befähigen, alle Vorkommnisse eines bestimmten (linguistischen) Phänomens innerhalb eines Korpus (wieder)auffindbar zu machen (vgl. Gries & Berez 2017: 403). Diese basale Erkenntnis beeinflusst aber selbstverständlich das Design der Annotationsarchitektur, insbesondere wenn die untersuchten Phänomene innerhalb eines gemeinsamen Korpus auf verschiedenen linguistischen Systemebenen aufzufinden sind. Innerhalb des SFB wird daher auf verschiedenen Annotationsebenen annotiert, die nicht mit linguistischen Systemebenen (z. B. Phonetik, Phonologie, Syntax, Morphologie) korrespondieren, sondern stattdessen phänomenbezogen sind (vgl. Breuer & Seltmann 2018: 146–147). Für jedes linguistische Phänomen, das in den Fokus der Untersuchungen tritt, wird dementsprechend eine eigene Annotationsebene erstellt. Die Tags, die auf diesen Annotationsebenen vergeben werden können, sind Entitäten in der Datenbank und werden zentral verwaltet und mit den

einzelnen Forscher:innen abgesprochen. Ihr Status als Entität in der Datenbank bedeutet, dass Tags nicht einfach als Abfolge von orthographischen Symbolen auf eine Antwort gespeichert werden, sondern jedes vergebene Tag einen Verweis auf eine Entität in der Datenbank darstellt – jedes Mal etwa wenn das Tag `IRR` (›irrelevant‹) vergeben wird, müssen die Forschenden nicht `IRR` in einem Annotationsfeld ausfüllen, sondern klicken auf das vorhandenen Tag, das einen Verweis auf das im Vorhinein definiertes Tag `IRR` darstellt. Dies minimiert einerseits Tippfehler und ermöglicht es andererseits, Tags zentral zu definieren.

Neben der Zweckmäßigkeit, beim tatsächlichen Annotieren die mögliche Auswahl aus dem gesamten Tagset auf die für das Phänomen relevanten Tags zu reduzieren, sobald die relevante Annotationsebene ausgewählt wurde, – sprich: beim Anlegen eines Tags wird festgelegt, auf welcher Annotationsebene dieser überhaupt erscheinen kann – basiert auch die Ausgabe des getaggten Materials auf diesen Annotationsebenen. Einfacher gesagt ermöglichen die verschiedenen Annotationsebenen das schnelle Wiederauffinden der getaggten Belege eines Phänomens. Dies geht außerdem einher mit Lüdelings (2017: 133) Forderung, dass »[i]n einer idealen variationistischen Korpusstudie […] eine Variable eine Annotationsebene darstellen [sollte, K. K. + M. P.], und darauf sollten alle Varianten annotiert sein.« Ungeachtet der Tatsache, dass ›linguistisches Phänomen‹ und ›Variable‹ sich nicht immer gleichsetzen lassen, ist dies ein methodisch sauberes und forschungspraktisch zweckmäßiges Vorgehen der Annotation (zur Diskussion des Begriffs der ›Variable‹ in der Variationslinguistik siehe exemplarisch Kallenborn 2019: 50–56).

Wie bereits erwähnt schränken die individuellen Tagebenen ein, welche einzelnen Tags überhaupt vergeben werden können. Dies ist nicht nur praktisch, sondern auch notwendig, da das momentane Taginventar im Annotationssystem des SFB knapp 900 individuelle Einzeltags umfasst. Im Folgenden wird dessen Aufbau kurz erklärt.

Prinzipiell ist die Tagging-Struktur des SFB:DiÖ an zwei Charakteristiken festzumachen: es handelt sich einerseits um ein hierarchisches, andererseits aber auch lineares System. Dieser scheinbare Widerspruch ergibt sich folgendermaßen: Sowohl an der Oberfläche, die den User:innen präsentiert wird, als auch in der Logik, in der Annotationen

entworfen werden, sind die einzelnen Tags hierarchisch angeordnet – sie stehen zueinander in *parent/child*-Beziehungen, die sich auf Generationen aufteilen. Einfach gesagt bewirkt das, dass auf gewisse Tags (die *parents*) nur gewisse andere Tags (die *children*) folgen können. Die einzelnen Generationen sind dabei nicht arbiträr festgelegt, sondern folgen einer spezifischen Logik: die sog. nullte Generation bezeichnet das linguistische Phänomen, welches untersucht wird, die erste Generation spezifiziert Kategorien dieses Phänomens, und die zweite Generation entspricht spezifischen Features dieser Kategorien, die sich gemeinhin innerhalb der Kategorie gegenseitig ausschließen (siehe Abbildung 1).[3] Ein einfaches Beispiel dazu wäre die Klassifikation von finiten Verben im Deutschen. Würde man finite Verben als unser Phänomen betrachten, so wären die Kategorien, die die Verben als finit definieren ›Person‹, ›Numerus‹, ›Genus verbi‹, ›Tempus‹ und ›Modus‹, und damit Tags der ersten Generation. Die Features dieser Kategorien wären die entsprechenden *children* der Tags der ersten Generation, und damit auf Generation zwei angesiedelt: für das Tag ›Person‹ wären die möglichen *children* dementsprechend ›1. Person‹, ›2. Person‹, ›3. Person‹, für das Tag ›Numerus‹ wären die *children* ›singular‹ und ›plural‹, usw. Innerhalb einer Kategorie schließen sich die Features dabei gemeinhin aus – ein konkretes finites Verb kann nicht singular und plural sein. Die einzelnen Kategorien ergänzen sich dabei in der Beschreibung des Phänomens, und finden Anwendung bei jedem Vorkommnis des Phänomens – jedes finite Verb im Deutschen lässt sich anhand dieser fünf Kategorien bestimmen, ansonsten ist es kein finites Verb. Diese Logik beherrscht das Design der verschiedenen Annotationskategorien, und unterstützt beim tatsächlichen Annotationsprozess darin, dass nur wohlgeformte Annotationen entstehen. Gleichzeitig werden die einzelnen Annotationen allerdings als lineare Abfolgen von Tags gespeichert, ohne ihre Hierarchie beim konkreten Speichervorgang abzubilden – dies erhöht die Flexibilität des

3 Selbstverständlich unterstützt die Annotationsarchitektur mehr als drei Generationen. Die dritte Generation, die auch teilweise Verwendung findet, ist zwar nicht formal definiert, spezifiziert aber zumeist Features genauer. Alle weiteren, tieferen Generationen finden in Ausnahmefällen ebenfalls Verwendung, folgen dabei aber keiner strikten globalen Logik.

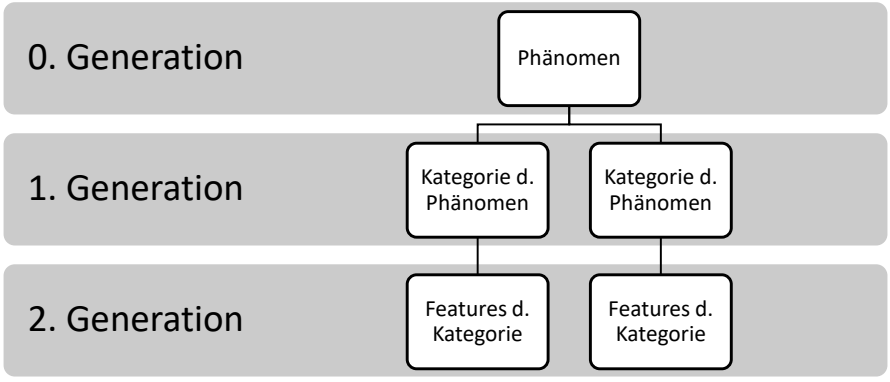

Abbildung 1: Schematische Abbildung der Logik des Tagging-Systems im SFB:DiÖ

Systems, und erlaubt es, einzelne Kategorien oder Features im Nachhinein leicht hinzuzufügen, da die konkret gespeicherten Annotationen als lineare Abfolge von Tags gespeichert werden, ohne dass ihre Hierarchisierung im Speicherakt (etwa durch Klammerung) ausgedrückt wird (vgl. Breuer & Seltmann 2018: 147).

Diese relativ rigide Logik des Tagging-Systems impliziert, dass Kategorien und mögliche Features der Kategorien bereits früh durchdacht werden müssen. Selbstverständlich ist ein iteratives Vorgehen möglich, und Tagging-Komplexe werden regelmäßig um Kategorien oder Features erweitert, aber zumeist werden bereits basale Kategorien- und Feature-Tags festgelegt, bevor tatsächlich im Korpus annotiert wird. Dies hat zwei Vorteile: Einerseits wird so implizite Annotation, die oft in opaker Weise nur Varianten zuweist, die auf undurchsichtigen Kriterien basieren und dementsprechend methodisch problematisch sind (vgl. Xiao 2009: 995), vermieden. Andererseits erzwingt eine solche Annotationslogik, bereits frühzeitig zumindest grundlegende Kategorien zur Charakterisierung der Varianten einer Variablen zu erstellen. Gerade hierbei ist es wichtig, die Charakteristika dieser Varianten möglichst transparent zu halten – nicht nur aus methodischen (vgl. Lüdeling 2017: 137), sondern auch

aus forschungspraktischen Gründen, da mehrere Personen an teilweise verschiedenen Projektstandorten die gleichen Phänomene annotieren. Daher wird jedem einzelnen Tag beim Erstellen eine Erläuterung zu Dokumentationszwecken beigefügt, was nicht nur dafür sorgt, dass das Tag intern kohärent bleibt, sondern die Nachnutzung der Daten auch abseits von Standards (siehe Abschnitt 2.3) erlaubt.

Die interne Logik des Annotationssystems trägt selbstverständlich auch gewisse strukturalistische Züge – die Idee, Sprache anhand von Kategorien und Features zu beschreiben, ist Linguist:innen natürlich bekannt, denn nichts anderes wird de facto bei einer (ohren)-phonetischen Transkription nach dem internationalen phonetischen Alphabet getan. Ein [b]-Laut etwa ist durch die Features bzw. Kategorien ›Stimmhaftigkeit‹, ›Artikulationsort‹ und ›Artikulationsweise‹ bestimmt und dem Laut wird hierbei beispielsweise ›+/−stimmhaft‹, ›bilabial‹ oder ›plosiv‹ zugewiesen. Das Graphem ⟨b⟩ repräsentiert dabei strenggenommen nicht ›einen‹ tatsächlichen Laut, sondern diese Ansammlung an Features gewisser Kategorien, die konventionalisiert sind. Tatsächliche Sprechereignisse, also Laute, die mit diesem Graphem kodiert werden, können so intersubjektiv nachvollziehbar gemacht werden, auch wenn sie sich individuell fein unterscheiden. Die Kategorie-und-Feature-Logik ist dementsprechend eigentlich bereits in der Linguistik erprobt. Sie eignet sich jedoch durchaus auch für Phänomene außerhalb der Phonetik und Phonologie, wie ein kurzes Beispiel aus dem pragmatischen Bereich zeigen soll.

Abbildung 2 zeigt die prinzipielle Tag-Struktur für eine Tag-Question.[4] Die Kategorien, die diese charakterisieren, und für die Untersuchung des Phänomens im Rahmen des SFB relevant sind, sind dabei (v. l. n. r.): die oberflächliche Form (d. h. welcher Wortart das Question-Tag zugeordnet wird), die Intonation, die Position in Satz und Turn, die Antwort, die auf diese folgt, und ihre (angenommene) Funktion. Die zugehörigen Features im Beispiel sind die der Konjunktion – das Question-Tag wird also als Konjunktion realisiert, z. B. *oder?* (`QTForm`→

4 Dieses Annotationsschema wurde gemeinsam mit PP04 und PP03 (insbesondere Stefanie Edler und Katharina Korecky-Kröll) des SFB:DiÖ erarbeitet.

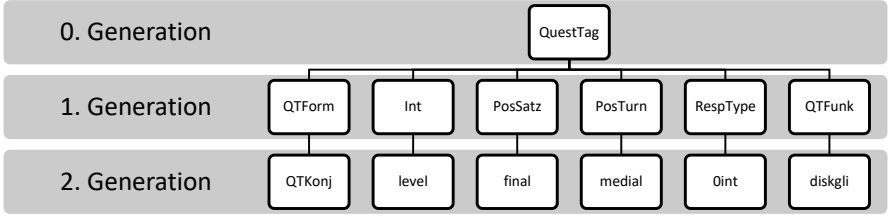

Abbildung 2: Tag-Struktur für Tag-Questions

`QTKonj`). Weiters ist die Intonation level (`Int`→ `level`), die Position im Satz ist final (`PosSatz`→ `final`), aber medial im Turn (`PosTurn`→ `medial`), die Tag-Question wurde nicht beantwortet, stattdessen hat der:die Sprecher:in den Turn gehalten (es war also nicht intendiert, eine Antwort zu erhalten; `RespType`→ `0int`), und die angenommene Funktion ist die der Diskursgliederung (`QTFunk`→ `diskgli`). Wie diese Tag-Kombination im Annotationstool aussieht wird in Abbildung 3 dargestellt. Das Sprachereignis, das damit getagged wurde, könnte also so etwas sein wie:

0504: *ja das geht jetzt schnell oder, zwei Wochen Urlaub und dann ist sie noch zwei Wochen da*[5]

Daran lassen sich zwei Dinge beobachten: Einerseits sind die zur Klassifikation verwendeten Kategorien heterogen und beziehen sich sowohl auf das betreffende Token selbst (in linguistischer Form und der Intonation), als auch auf den linguistischen Kotext (durch Klassifikation nach Position in Satz und Turn) sowie auf pragmatische Faktoren (in der intendierten Antwort und der Zuweisung einer Funktion). Das Zuweisen der Features zu den einzelnen Kategorien ist dementsprechend entweder einfach – die oberflächliche Form, die ein Question-Tag annimmt, ist transparent – oder kompliziert und mit wesentlichem Interpretationsaufwand verbunden (was insbesondere die Funktion betrifft). Der Vorteil

5 Dieser Beleg stammt aus dem SFB-Korpus.

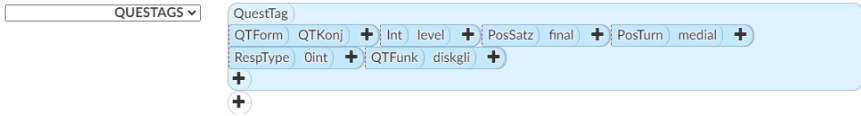

Abbildung 3: Tag-Struktur für Tag-Questions in der Datenbankansicht

gegenüber einer kürzeren, nicht mehr-dimensionalen Annotation ist allerdings, dass die Nachvollziehbarkeit der Zuordnung bestehen bleibt. Denn was andererseits bei dieser Form der Annotation auffällt, ist, dass die Zuweisung zu einer expliziten Variante einer Variablen ausbleibt. Wie Lüdeling (2017: 137) festhält, ist die Definition von Variablen und Varianten insbesondere abseits der Phonetik schwierig, und das Bestimmen oberflächlicher und transparenter Kategorien der gangbarere Weg. Selbstverständlich können anhand der Features später Varianten zu einer – wie auch immer definierten – pragmatischen Variable zusammengezogen werden (etwa verschieden intonierte Varianten einer Konjunktion als Variable, oder deren Funktion als Variable, mit beispielsweise unterschiedlichen linguistischen Formen als Varianten), diese Variablen- und Variantenbildung bleibt allerdings transparent und für die weitere wissenschaftliche Gemeinschaft nachvollziehbar.

Wie bereits ersichtlich wurde, ist Annotation in unserem Verständnis – sowohl als tatsächlicher Prozess des Annotierens als auch bereits im Design eines Annotationsschemas – keineswegs ein unproblematischer Prozess. Das Folgende geht daher auf den epistemologisch schwierigen Status der Annotation genauer ein und zeigt, wie Annotation nicht nur als (linguistische) Forschung (vgl. Lüdeling 2017: 141) verstanden werden kann, sondern essentieller Grundteil jeglicher (linguistischen) Erkenntnis ist – de Marneffe & Potts (2017: 426) beispielsweise bezeichnen die Annotationen dabei als genauso wichtig wie die daraus resultierenden Daten.

2.2 Annotation im SFB DiÖ als Forschungsarbeit/Erkenntnisgewinn

Die Annotation ist dabei keineswegs ein Element nur moderner Forschung – Lordick et al. (2016: 187) charakterisieren das Annotieren insgesamt als »eine der ältesten und allgegenwärtigsten wissenschaftlichen Praktiken […], die es gibt.« Nichtsdestotrotz bleibt eines der elementaren Probleme des Annotierens erhalten – es gibt de facto kein perfektes Annotationssystem, da eine perfekte Klassifikation eines wie auch immer gearteten Objekts allumfassendes Wissen über das Objekt voraussetzt (vgl. Sperberg-McQueen 2015: 378), was ein solches Objekt natürlich als Forschungsgegenstand disqualifizieren würde. Insbesondere in der Variationslinguistik werden solche Fragen der Klassifikation relevant, da die Heterogenität des Forschungsobjekts – sprich die sprachliche Variation, und der Versuch, diese Variation zu erklären und klassifizieren – in der Natur der Sache liegt. Daher wird im Folgenden kurz über den epistemologischen Status der Annotation (und damit verbundenen Klassifikation) von sprachlichen Variationsphänomenen im sprachdynamisch-variationistischen Paradigma reflektiert.

Die prinzipiellen Kernklassifikationen sprachdynamisch-variationistischer Linguistik lassen sich als Horizontalität und Vertikalität beschreiben (vgl. Herrgen & Schmidt 1989: 304; Schmidt & Herrgen 2011: 73). Die Variation, die also betrachtet wird, wird in elementarster Weise zunächst zugehörig zu geographischem Raum und Formalität/Situativität betrachtet. Abstrakt-ideal ist ein Kernforschungsbereich dementsprechend das Abbilden wie auch immer gearteter Varianten einer Variablen auf einem Koordinatensystem von Horizontalität (also geographische Verteilung einer Sprachgemeinschaft) und Vertikalität (also Verortung auf einem Spektrum intraindividueller Variation, oftmals anhand von Förmlichkeit oder Distanz zu Standardvarietäten). Dies soll dabei keinesfalls das variationistische Projekt abwerten – in gewisser Weise ist eine solch quasi-tabellarische Taxonomie Voraussetzung für alle Formen empirischer Erkenntnis in einem rationalistischen Paradigma (vgl. Foucault 1994 [1966]: 54). Es geht im Folgenden allerdings nicht um die Verortung von sprachlichen Phänomenen auf den Achsen der Horizontalität

und Vertikalität – diese sind sozusagen raison d'être variationistischer Linguistik, diese Verortung auf den Achsen ist gleichzeitig Voraussetzung für, und Produkt dieser: sprich, ohne linguistische Variation, die durch geographische Verteilung oder Formalität der Gesprächssituation bestimmt wird, wäre variationistische Linguistik nicht denkbar. Gleichzeitig gelten diese Pole als Steuerungsfaktoren der Variation, und die Verortung der Varianten an diesen Polen ist auch ein dezidiertes Forschungsinteresse.

Analog dazu werden diese Informationen im Annotationssystem des SFB nicht ausgezeichnet, sondern sind als Meta-Informationen über eine Sprecher:innentabelle, die den Ort vermerkt, für den diese Person als Informant:in gilt, und das Experimentalsetting, das im weitesten Sinne die Zugehörigkeit zu einem vertikalen Spektrum der Formalitität ausweist, zu den einzelnen Tokens zugewiesen. Das meint, dass grundsätzlich weder die intendierte Förmlichkeit eines Settings noch die geographische Zugehörigkeit des:der Sprecher:in tatsächlich annotiert wird, sondern stattdessen zu den Tokens eines:einer Sprecher:in gespeichert wird, woher der:die Sprecher:in kommt, und in welchem Setting die Sprachdaten produziert wurden. Diese Zuweisung wird benötigt, um in diesem Paradigma valide Erkenntnisse zu gewinnen – anhand des Vorkommens gewisser, wie auch immer definierter Varianten an verschiedenen Punkten des Koordinatensystems aus horizontaler und vertikaler Variation werden diese Varianten festgeschrieben. Dieses Festschreiben – also die Zuordnung von sprachlicher Variation (als ›Variante‹ einer ›Variable‹) zu einem Register oder Dialekt – ist oft auch das Produkt variationistischer Forschung.

Innerhalb dieses (ontologischen) Rahmens findet also unsere Form variationslinguistischer Annotation statt. Die implizite oder explizite Annotation als Voraussetzung für eine Auswertung ist dementsprechend in den meisten Fällen empirischer Linguistik zwingend notwendig, um Sprachdaten als linguistische Varianten einer Variablen zu klassifizieren, und anschließend diese Varianten innerhalb eines horizontal/vertikalen Koordinatensystems festschreiben zu können. Das ›Festschreiben‹ ist hier nahezu wörtlich zu verstehen, denn normalerweise werden nur diskrete Punkte angenommen – denn insbesondere auf der vertikalen

Ebene fällt eine Klassifizierung schwer. Was auf der horizontalen Ebene teilweise über politische Grenzen (wie Ortsgebiete oder Landesgrenzen), teilweise über Georeferenz, erfolgen kann – gemeint sind vermeintlich objektive Orientierungspunkte – fehlt auf der vertikalen. Hier werden entweder bewusst Vereinfachungen im Sinne einer Modellbildung angenommen, in der Varianten zu (teilweise arbiträr definierten) Sprachlagen oder Varietäten (wie ›Standardsprache‹ oder ›Basisdialekt‹) gebündelt werden, oder es wird versucht, graduelle Unterschiede zu treffen, z. B. mithilfe von Dialektalitätsmessungen wie von Herrgen & Schmidt (1989) – die verständlicherweise den Nachteil haben, sich hauptsächlich auf phonetisch-phonologische Varianten zu beziehen, und Fragen der Salienz und Pertinenz der Merkmale nur bedingt akkommodieren zu können. Der tatsächliche Annotationsprozess folgt einem ähnlichen Schema – während Sprachereignisse, die annotiert werden, große Heterogenität aufweisen können, kann auch ein auf Kategorien und Features basiertes Annotationssystem diese nur diskreten Punkten zuweisen. Diese Kategorisierung findet allerdings nicht in einem zwei-, sondern in einem *n*-dimensionalen Raum statt, wobei *n* der Anzahl der Kategorien – und damit der Achsen – entspricht, und jedes individuelle Feature einer Kategorie einen möglichen Punkt auf der jeweiligen Achse repräsentiert (vgl. hierzu Sperberg-McQueen 2015: 380). Der Nachteil der diskreten Kategorisierung, die sich aus der Annotationslogik ergibt, kann also durch die Multidimensionalität des Systems zumindest ein Stück weit ausgeglichen werden.

Denn auch wenn das Festschreiben bestimmter Features einer Kategorie – wie etwa die der Funktion eines Question-Tags im obigen Beispiel – drastisch verkürzend wirkt (was es, in gewissem Sinne, auch ist), so muss bewusst bleiben, dass dies de facto bei allen Klassifikationen der Fall ist. Auch innerhalb von gut definierten Kategorien gibt es notwendigerweise Heterogenität innerhalb der einzelnen Features, wie wohl jede:r, die:der phonetische Transkriptionen derselben Aufnahme von unterschiedlichen Transkriptor:innen verglichen hat, bezeugen kann. Annotation ist also in diesem Sinne auch eine Zuschreibung, die notwendigerweise verkürzend und teilweise subjektiv sein muss. Annotation im hier gemeinten Sinne – als explizite Kategorisierung von sprachlichen

Phänomenen – ist in dieser Weise auch immer Interpretation der Daten, basierend auf linguistischer Theorie. Problematisch kann dabei sein, dass diese linguistische Theorie teilweise an den mit ihr annotierten Daten überprüft wird, was einen Zirkularitätseffekt zur Folge haben kann (vgl. Consten & Loll 2012: 711–712.). Wie bei jeder anderen Interpretation der Daten ist es daher methodologisch notwendig, mehr als nur das Ergebnis zu präsentieren – der Akt der Interpretation selbst ist bereits eine epistemologische Leistung. Die beste Möglichkeit, die Linguist:innen dementsprechend haben, um ihre Ergebnisse intersubjektiv nachvollziehbar zu machen, ist daher das Veröffentlichen und Teilen der zugrunde liegenden Annahmen, Analysen und Annotationsprinzipien und im besten Falle auch der annotierten Forschungsdaten.

2.3 Annotationsstandards in der Variationslinguistik

Während die Verwendung standardisierter Annotationsschemata oftmals als Goldstandard für die Interoperabilität und Wiederverwendbarkeit von Korpora gilt (siehe Ide, Calzolari et al. 2017 für einen Überblick über standardisierte Annotationsschemata), besteht eine andere Möglichkeit in der Offenlegung des verwendeten Annotationsvokabulars, um dieses verständlich zu machen (vgl. Lehmberg & Wörner 2008: 484). Dies geschieht im Rahmen des SFB:DiÖ durch eine verbindliche Erläuterung zu jedem einzelnen Tag bei dessen Erstellung (vgl. Breuer & Seltmann 2018: 147). Die Digitalisierung spielt bei Annotationssystemen eine große Rolle (siehe Wong et al. 2011), worauf auch im SFB eingegangen wird. Im weiteren Verlauf des SFB werden diese Erläuterungen in einer online-Plattform zugänglich gemacht. Andere Möglichkeiten für kleinere Forschungsvorhaben bestehen etwa in der Ablage der Annotationsrichtlinen auf einem etablierten Repositorium wie Github, wie etwa exemplarisch Clausen & Scheffler (2020).

Ein solches Vorgehen entspricht nicht nur dem Credo von Lordick et al. (2016: 195), solche elementaren Forschungsdaten »aus der Schublade« zu holen und damit breiter nutzbar zu machen (schlussendlich wird dadurch die Nachvollziehbarkeit der Forschungsergebnisse vereinfacht), sondern verhilft auch dazu, einen heterogenen Forschungsgegenstand

wie sprachliche Variation besser zu konzeptualisieren. Wie oben erwähnt handelt es sich bei der Annotation um die Zuweisung von Sprachdaten zu bestimmten Kategorien, die idealtypischerweise in Varianten einer Variable zusammengefasst werden. Um diesen Vorgang transparent und nachvollziehbar zu halten, müssen die Annahmen und Prinzipien offengelegt werden, auf denen die Annotation beruht (vgl. Consten & Loll 2012: 705).

Hier stellt sich auch die Frage, inwieweit standardisierte Annotationsschemata für variationistische Forschung geeignet sind. Selbstverständlich spricht nichts gegen die Verwendung dieser, problematisch kann hingegen sein, dass diese oft nicht gegenstandsadäquat sind – sie eignen sich oft für konventionalisierte Kategorien, wie etwa Part-of-Speech-Tagging, oder syntaktische Annotation der Satzstruktur. Variationistische Phänomene können aber oftmals nicht genau genug beschrieben werden – was aber natürlich auch nicht der Zweck standardisierter Annotationsschemata ist und sich auch in einer verminderten Genauigkeit der verschiedenen Systeme zeigt (vgl. Zupan et al. 2019: 651). Standardisierte Annotationsschemata müssen Versatilität und Generalisierbarkeit aufweisen (Ide & Romary 2004: 212), was sie oft zu breit für spezifisch variationistische Fragestellungen machen kann. Newman & Cox (2021: 36–37) weisen auf die Herausforderungen und Schwierigkeiten hin, die konventionalisierte Annotationen vor allem für Non-Standard-Varietäten bergen. Bei variationslinguistischen Korpora muss auf die Besonderheiten Rücksicht genommen werden, die durch automatisierte bzw. standardisierte Annotationssysteme nicht abgedeckt werden können. So müssen bei der bereits anfangs beschriebenen Auswahl der Methoden, der geeigneten Annotationen, der Selektion der Konventionen sowie der Tools die besonderen Gegebenheiten eines Variationskorpus bedacht werden und es muss auf die verschiedenen sich daraus ergebenden Problemstellungen adäquat eingegangen werden. Newman & Cox (2021: 37) beschreiben dies so:

> […] many lesser-studied languages and varieties […] may require the development of annotation conventions that »fit« the linguistic features of the source materials, as well as the implementation of

> these conventions in existing annotation tools, adding additional complexity to the overall corpus annotation workflow.

Jegliche Kategorisierungsversuche sind in gewissem Sinne auch ontologische Aussagen über den Gegenstand, den sie kategorisieren (vgl. Sperberg-McQueen 2015: 378) – wenn dieser Gegenstand sprachliche Variation ist, so wird er mit standardisierten Kategorisierungen, die oft nur auf einem einzelnen Label beruhen, nicht adäquat abgebildet. Die Standardisierung für variationistische Sprachmaterialien ist daher äußerst schwierig umzusetzen und für viele Phänomene impraktikabel bis nicht anwendbar (vgl. Ide & Romary 2004: 213). Auch ist händische Annotation von Sprachdaten sehr aufwändig, besonders für variationistische Korpora (vgl. Rehbein et al. 2012: 2).

Aus diesen Gründen versucht sich die Annotation im SFB:DiÖ zwar – wann immer möglich – an standardisierten Annotationsschemata zu orientieren, allerdings weniger mit dem Ziel, diese auch tatsächlich zu implementieren, und mehr in dem Sinne, dass das interne Annotationssystem nicht zu hermetisch gestaltet wird. Gemeinsam mit der umfassenden Dokumentation und der subsequenten Publikation aller Annotationen am Projektende wird damit die Wiederverwertbarkeit der Daten im Sinne der Open-Science-Idee gewährleistet, und der komplette Erkenntnisprozess möglichst nachvollziehbar gemacht.

3 Zusammenfassung

Im Vorangegangenen wurde beschrieben, wie in einem Großprojekt linguistische Annotation von Sprachdaten gestaltet werden kann.

Nachdem das Annotationssystem des SFB:DiÖ erläutert und dessen Modell aus wissenschaftstheoretischer Sicht beleuchtet wurde, kam der Einsatz standardisierter Annotationsschemata zur Diskussion. Hier wurde auf die verschiedenen Herausforderungen und Besonderheiten, die bei der Erstellung eines Annotationssystems in einem variationslinguistischen Projekt zu beachten sind, hingewiesen. Dass gerade die Annotation eines variationslinguistischen Korpus durchaus eine zukunftsweisende

Vorgehensweise ist, wurde auch von Newman & Cox (2021: 44) beschrieben, die festhalten, dass gerade die Annotation variationistischer und vergleichbarer Korpora eine der Herausforderungen in der Computer- und Korpuslinguistik im 21. Jahrhundert darstellen wird.

Auch wenn die Verwendung rein standardisierter Annotationsschemata für den SFB nicht angebracht ist, wäre es durchaus wert, die Verbindung mit möglichen Referenzmodellen, wie etwa OLiA (Ontologies of Linguistic Annotation, vgl. Chiarcos & Sukhareva 2015), auf ihre Praktikabilität zu prüfen. Auf jeden Fall ist die Weiternutzung und Interoperabilität der primären Forschungsdaten und deren Annotationen ein wichtiges Anliegen des Projekts.

Dieser Fokus auf die oftmals (zu) wenig beachteten Aspekte empirischer Linguistik hat hoffentlich einen Beitrag zur Schärfung dieser Konzepte und ihrer Applikation in der Forschung geleistet.

Abkürzungsverzeichnis

`0int`	keine Beantwortung intendiert
`diskgli`	Diskursgliederung
`Int`	Intonation
`IRR`	Irregulär
`PosSatz`	Position im Satz
`PosTurn`	Position im Turn
`QTForm`	Form des Question-Tag
`QTFunk`	Funktion des Question-Tag
`QTKonj`	Konjunktiv
`QuestTag`	Question-Tag
`RespType`	Art der Beantwortung
SFB:DiÖ	Spezialforschungsbereich ›Deutsch in Österreich‹

Literatur

Bird, Steven & Mark Libermann. 2001. A formal framework for linguistic annotation. *Speech Communication* 33(1–2). 24–60. *https://doi.org/10.1016/S0167-6393(00)00068-6* (Abruf 1. Juli 2021).

Breuer, Ludwig Maximilian. 2021. *„Wienerisch" vertikal. Theorie und Methoden zur stadtsprachlichen syntaktischen Variation am Beispiel einer empirischen Untersuchung in Wien*. Wien: Universität Wien. Dissertation.

Breuer, Ludwig Maximilian & Melanie E. H. Seltmann. 2018. Sprachdaten(banken) – Aufbereitung und Visualisierung am Beispiel von SyHD und DiÖ. In Ingo Börner, Wolfgang Straub & Christian Zolles (Hgg.), *Germanistik digital: digital humanities in der Sprach- und Literaturwissenschaft*, 135–152. Wien: facultas.

Budin, Gerhard, Stephan Elspaß, Alexandra N. Lenz, Stefan Michael Newerkla & Arne Ziegler. 2018. Der Spezialforschungsbereich „Deutsch in Österreich (DiÖ). Variation – Kontakt – Perzeption". *Zeitschrift für germanistische Linguistik* 46(2). *https://doi-org.uaccess.univie.ac.at/10.1515/zgl-2018-0017* (Abruf 12. September 2021).

Chiarcos, Christian & Maria Sukhareva. 2015. OLiA: Ontologies of linguistic annotation. *Semantic Web* 6(4). *https://doi.org/10.3233/SW-140167* (Abruf 18. Juli 2021).

Clausen, Julia & Tatjana Scheffler. 2020. *Annotation manual for German question tags*. O. O. *https://github.com/TScheffler/TagQuestions* (Abruf 18. Juli 2021).

Consten, Manfred & Annegret Loll. 2012. Circularity effects in corpus studies – why annotations sometimes go round in circles. *Language Sciences* 34. 702–714.

DiÖ. 2021. *Github Repositorien des Spezialforschungsbereich Deutsch in Österreich*. O. O. *https://github.com/german-in-austria* (Abruf 12. September 2021).

Fingerhuth, Matthias & Ludwig Maximilian Breuer. 2020. Language production experiments as tools for corpus construction: A contrastive study of complementizer agreement. *Corpus Linguistics and Linguistic Theory [ahead of print]. https://doi-org.uaccess.univie.ac.at/10.1515/cllt-2019-0075* (Abruf 12. September 2021).

Foucault, Michel. 1994 [1966]. *The order of things: An archeology of the human sciences*. New York: Vintage.

Graf, Arnold, Ludwig Maximilian Breuer, Tahel Singer & Markus Pluschkovits. in Vorbereitung. *Transcribe: a web-based linguistic transcription tool*. O. O.

Gries, Stefan Thomas & Andrea L. Berez. 2017. Linguistic annotation in/for corpus linguistics. In Nancy Ide & James Pustejovsky (Hgg.), *Handbook of linguistic annotation*, 379–409. Dordrecht: Springer.

Herrgen, Joachim & Jürgen Erich Schmidt. 1989. Dialektalitätsareale und Dialektabbau. In Wolfgang Putschke, Werner Veith & Peter Wiesinger (Hgg.),

Dialektgeographie und Dialektologie. Günter Bellmann zum 60. Geburtstag von seinen Schülern und Freunden (Deutsche Dialektgeographie), 304–346. Marburg: Elwert.

Ide, Nancy. 2017. Introduction. In Nancy Ide & James Pustejovsky (Hgg.), *Handbook of linguistic annotation*, 1–18. Dordrecht: Springer.

Ide, Nancy, Nicoletta Calzolari, Judith Eckle-Kohler, Dafydd Gibbon, Sebastian Hellmann, Kiyong Lee, Joakim Nivre & Laurent Romary. 2017. Community standards for linguistically annotated resources. In Nancy Ide & James Pustejovsky (Hgg.), *Handbook of linguistic annotation*, 113–165. Dordrecht: Springer.

Ide, Nancy & Laurent Romary. 2004. International standard for a linguistic annotation framework. *Natural Language Engineering* 10(3–4). 211–225.

Kallenborn, Tim. 2019. *Regionalsprachliche Syntax: Horizontal-vertikale Variation im Moselfränkischen* (Zeitschrift für Dialektologie und Linguistik. Beihefte 176). Stuttgart: Steiner.

Lehmberg, Timm & Kai Wörner. 2008. Annotation standards. In Anke Lüdeling & Merja Kytö (Hgg.), *Corpus linguistics: An international handbook, volume 1* (Handbücher zur Sprach- und Kommunikationswissenschaft 29.1), 484–501. Berlin & New York: De Gruyter.

Lordick, Harald, Rainer Becker, Michael Bender, Luise Borek, Canan Hastik, Thomas Kollatz, Beata Mache, Andrea Rapp, Ruth Reiche & Niels-Oliver Walkowski. 2016. Digitale Annotationen in der geisteswissenschaftlichen Praxis. *Bibliothek – Forschung und Praxis* 40(2). 186–199.

Lüdeling, Anke. 2017. Variationistische Korpusstudien. In Marek Konopka & Angelika Wöllstein (Hgg.), *Grammatische Variation: Empirische Zugänge und theoretische Modellierung* (Jahrbuch des Instituts für deutsche Sprache 2016), 129–144. Berlin & Boston: De Gruyter. *https://doi.org/10.1515/9783110518214-009* (Abruf 10. Juli 2021).

Marneffe, Marie-Catherine de & Christopher Potts. 2017. Developing linguistic theories using annotated corpora. In Nancy Ide & James Pustejovsky (Hgg.), *Handbook of linguistic annotation*, 411–438. Dordrecht: Springer.

Nagy, Naomi & Devyani Sharma. 2013. In Robert J. Podesva & Devyani Sharma (Hgg.), *Research Methods in Linguistics*, 235–256. Cambridge: Cambridge University Press.

Newman, John & Christopher Cox. 2021. Corpus annotation. In Magali Paquot & Stefan Thomas Gries (Hgg.), *A practical handbook of corpus linguistics*, 25–48. Cham: Springer International Publishing.

Rehbein, Ines, Josef Ruppenhofer & Caroline Sporleder. 2012. Is it worth the effort? Assessing the benefits of partial automatic pre-labeling for frame-semantic annotation. *Language Resources and Evaluation* 46(1). 1–23.

Schmidt, Jürgen Erich & Joachim Herrgen. 2011. *Sprachdynamik: Eine Einführung in die moderne Regionalsprachenforschung* (Grundlagen der Germanistik 49). Berlin: Erich Schmidt Verlag.

Sperberg-McQueen, Christopher Michael. 2015. Classification and its structures. In Susan Schreibman, Ray Siemens & John Unsworth (Hgg.), *A new companion to digital humanities*, 377–393. New York: John Wiley & Sons, Ltd. *https://doi.org/10.1002/9781118680605.ch26* (Abruf 10. Juli 2021).

Wong, Deanna, Steve Cassidy & Pam Peters. 2011. Updating the ICE annotation system: tagging, parsing and validation. *Corpora* 6(2). 115–144.

Xiao, Richard. 2009. Theory-driven corpus research: using corpora to inform aspect theory. In Anke Lüdeling & Merja Kytö (Hgg.), *Corpus linguistics: An international handbook, volume 2* (Handbücher zur Sprach- und Kommunikationswissenschaft 29.2), 987–1008. Berlin & New York: De Gruyter.

Zupan, Katja, Nikola Ljubešić & Tomaž Erjavec. 2019. How to tag non-standard language: Normalisation versus domain adaptation for Slovene historical and user-generated texts. *Natural Language Engineering* 25(5). 651–674.