

Kann Nonstandard standardisiert werden? Ein Annotations- Standardisierungsversuch nicht nur von PoS- Tags am Beispiel des Spezialforschungsbereichs „Deutsch in Österreich“

Seltmann, Melanie E.-H.

melanie.seltmann@univie.ac.at
Universität Wien, Österreich

Einleitung

Unter einer Annotation von Sprachdaten wird eine Markierung, Kategorisierung und Interpretation derselben verstanden. Sie dienen – auch in anderen Wissenschaftsdisziplinen – in der Regel der wissenschaftlichen Auseinandersetzung und dem Forschungsprozess mit einem Datum und halten ebenso dessen Ergebnis fest (vgl. Breuer/Seltmann 2018: 145). Für die technische Realisierung können dabei sehr unterschiedliche Umsetzungen eingesetzt werden, abhängig sowohl von den Daten, die in einem Projekt untersucht werden, als auch den persönlichen Vorlieben der (entscheidenden) Mitarbeiter.

Der Spezialforschungsbereich „Deutsch in Österreich. Variation – Kontakt – Perzeption“ (SFB DiÖ)

Im Vortrag soll anhand des (Aufbaus des) Annotationssystems des Spezialforschungsbereichs „Deutsch in Österreich. Variation – Kontakt – Perzeption“ (FWF F60) (kurz: SFB DiÖ) und dessen Teilprojekts 11 „Kollaborative Online-Forschungsplattform“ die Annotation von Nonstandardvarietäten hinterfragt werden. Der SFB DiÖ beschäftigt sich mit der Vielfalt und dem Wandel der deutschen Sprache in Österreich. Er erforscht den Gebrauch und die subjektive Wahrnehmung von deutscher Sprache in Österreich und untersucht die Einflüsse von Kontaktsprachen auf sie. Der SFB ist an vier wissenschaftlichen Institutionen in Österreich angesiedelt: an den Universitäten Wien, Graz und Salzburg sowie an

der Österreichischen Akademie der Wissenschaften. Dabei sind unterschiedliche Forschungsbereiche und -institute beteiligt: von der Germanistik über die Slavistik bis hin zur Translationswissenschaft und der Schallforschung. Die drei thematischen Taskcluster befassen sich mit den inhaltlichen Säulen Variation, Kontakt und Perzeption und werden von zwei Taskclustern für die Organisation und Koordination sowie die Kollaborative Online-Forschungsplattform ergänzt und unterstützt (s. Abb. 1, vgl. Budin et al. 2018).

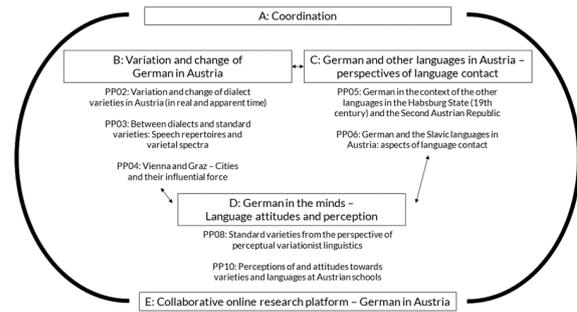


Abbildung 1. SFB-DiÖ Taskcluster und Teilprojekte

Durch die verschiedenen beteiligten Institute bedarf es schon innerhalb des SFB einer großen Flexibilität und Variabilität im Annotationssystem, da in den verschiedenen Teilprojekten nicht nur unterschiedliche Fragestellungen und linguistische Systemebenen untersucht werden, sondern dies auch aus unterschiedlichen disziplinären Perspektiven und in unterschiedlicher Granularität. Hinzu kommt, dass nicht nur sehr viele, sondern auch sehr heterogene empirische Daten (Methodenpluralismus) erhoben werden, welche einheitlich und methodenübergreifend annotiert werden sollen (vgl. Abb.2). Die Anzahl der dadurch entstehenden Annotationen sowie deren Vokabular ist sehr umfangreich. Da eines der Ziele der Forschungsplattform aber ist, die annotierten Daten einem möglichst breiten Publikum zur Verfügung zu stellen, muss das Annotationssystem auch „nach außen“ (intersubjektiv) nachvollziehbar sein und sich, wo möglich, an bereits bestehende Standards halten (z. B. bereits vorhandene Tagsets). Um diesem multidimensionalen Anspruch gerecht zu werden, wurde und wird ein eigenes Annotationssystem entwickelt, was auf ggf. vorhandene Standards (in seinem Vokabular) zurückgreift. Die Besonderheit liegt dabei in einer Zweiteilung von technischer Speicherung und Repräsentation.

UNTERSUCHUNGSOBJEKT (INTENDIERT)	SYSTEMEBENE IM FOKUS	METHODEN/SETTING	
Tendenz NON-STANDARD INTENDIERTER DIALEKT	SYNTAX (& MORPHOLOGIE, PHONOLOGIE)	Sprachproduktionsexperiment	
		Übersetzung (in Intendierten Dialekt)	
	PHONOLOGIE (& SYNTAX, MORPHOLOGIE)	(gelenktes) Freundesgespräch	
		(leitfadengesteuertes) Interview	
Tendenz STANDARD INTENDIERTE STANDARDSPRACHE	PHONOLOGIE	Übersetzung (in Intendierte Standardsprache)	
		Leseaufgaben (NWS, Einzelwörter, Bildbenennung)	
	SYNTAX (& MORPHOLOGIE, PHONOLOGIE)	Sprachproduktionsexperiment	

Abbildung 2. Erhebungssettings des SFB DiÖ

Das Annotationssystem des SFB DiÖ

Ziel der Annotation ist es, ein einheitliches Gesamtbild über alle Systemebenen der gesprochenen Sprache hinweg zu verfolgen. Die Annotation wird dabei phänomenbezogen auf einer Annotationsebene je betrachtetem Phänomen vorgenommen. Neben den Systemebenen wie Phonetik/Phonologie, Morphologie/Lexik und Syntax kommen auch qualitativ-inhaltliche sowie ggf. gesprächsanalytische Annotationen hinzu. Um diesem multidimensionalen Anspruch gerecht zu werden, wurde und wird ein eigenes Annotationssystem entwickelt, was auf ggf. vorhandene Standards (in seinem Vokabular) zurückgreift. Die Besonderheit liegt dabei in einer Zweiteilung von technischer Speicherung und Repräsentation. Die Speicherung funktioniert linear innerhalb einer relationalen Datenbank auf mehreren Tagebenen, die zeitgleich im Mehrbenutzerzugriff bearbeitet werden können. Die Repräsentation wird jedoch hierarchisch geliefert. Dies hat den Vorteil, das Annotierende durch die komplexe Annotation geleitet werden und weniger Obacht auf die richtige Reihenfolge der Tags legen müssen, da sich diese aus der Hierarchie automatisch ergibt.

Das vorgestellte Annotationssystem ist insofern hierarchisch, dass es eine child-parent-Struktur für ein spezifisches phänomenbezogenes Tagset fordert. Dies bedeutet, dass einerseits immer Kategorien zu den zu taggenden Features angegeben werden müssen, andererseits ihr Aufbau sinnvoll beschrieben werden muss. Dadurch wird die Kohärenz und Nachvollziehbarkeit des Tagsets erhöht. Auch in der praktischen Anwendung wird hierdurch die Annotation vereinfacht, indem durch die Vorgabe der Struktur selbige ebenso beim Annotationsprozess abgerufen wird: Durch diesen Vorgang wird es erst möglich, dass z.B. eine Tagging-Eingabemaske

aus den unzähligen vorhanden Tags jene herausfiltern kann, die für Annotierende in diesem Moment relevant sind.

Wenn diese Hierarchie jedoch auch in die Speicherung Eingang finden würde, wären Probleme vorprogrammiert. Die Annotationen werden daher linear gespeichert, um eine größtmögliche datenstrukturelle Flexibilität zu ermöglichen. Dies ist insbesondere dann nötig, wenn das Annotationssystem iterativ zum Forschungsprozess erweiterbar sein soll, wie es im SFB DiÖ der Fall ist.

Vergleich von Annotationssystemen

Im Vortrag wird jedoch nicht nur das Annotationssystem des SFB Deutsch in Österreich vorgestellt, sondern auch mit anderen Annotationssystemen verglichen. Hierfür wird insbesondere der Bereich des Part-of-Speech-Taggings herausgegriffen. Unter Part-of-Speech-Tagging oder PoS-Tagging wird die automatische Zuweisung von Wortarten verstanden. Es wird jedem Token automatisch durch einen Tagger eine Wortart zugewiesen (vgl. Lemnitzer/Zinsmeister 2010: 72).

Vergleichssysteme sind etwa das Stuttgart-Tübingen-Tagset (STTS, Schiller et al. 1999) oder das European Dialect Syntax (EDiSyn)-Tagset (Barbiers/Wyngaerd o.J.). Trotz der häufigen Verwendung des STTS ist es in verschiedenen Punkten relativ problematisch. Ziel ist einmal nur die Wortartenbestimmung selbst, zudem im Falle des häufig verwendeten großen Tagsets die Repräsentation von Morphologie und Derivation (vgl. Telljohann et al. 2013: 2). Das Tagset ist in Anbetracht verschiedener Klassifizierungskategorien aufgebaut, die Kategorisierung erfolgte nach morphologischen, syntaktischen sowie semantischen Kriterien (vgl. Schiller et al. 1999: 4). Es besteht aus elf Hauptwortarten, welche unterschiedlich tief klassifiziert sind (vgl. Schiller et al. 1999: 5). Diese für einige Forschungsfragen sehr hilfreichen Subklassifizierungen stellen jedoch auch ein Problem dar, da keine Einheitlichkeit im Aufbau des Tagvokabulars gegeben ist und die Auswahl nicht ohne Vorkenntnisse zu diesem Aufbau ersichtlich ist. Durch den uneinheitlichen Aufbau in Bezug auf Analysetiefe sowie der Struktur der Teilelemente im Tag ist das Set einerseits nur für eingeschränkte Zwecke verwendbar und – wenn dieses Problem ausgeglichen werden soll – nur schwierig nachvollziehbar um die dazu nötigen, neuen Teilelemente eines Tags ergänzbar.

Dies wird insbesondere beim PoS-Tagging von Nonstandardvarietäten ersichtlich, das eine Herausforderung darstellt, da viele varietätenspezifische Ausprägungen von Wortarten bzw. Features von Wortarten und Besonderheiten dieser nicht abgedeckt werden. Eine klarere Strukturierung und weitaus mehr Möglichkeiten bietet hier das European Dialect Syntax (EdiSyn)-Tagset, welches zumindest in Ansätzen die Wortarten als Kategorien und deren Ausprägungen (grammatische Features) als Features der Kategorien modelliert, d.h.

eine Trennung von Tags auf verschiedenen Ebenen durchführt. Die Features selbst werden wiederum nicht benannten Kategorien (grammatischen Kategorien wie Kasus, Genus etc.) zumindest zur Gruppierung zugeordnet. Der große Vorteil daran ist, dass Wortarten und Features unabhängig voneinander erweitert werden können und dass Features potentiell auch mehreren Wortarten zugeschrieben werden können – was nicht nur für den sprachtypologischen Vergleich verschiedener Sprachen sinnvoll erscheint, sondern auch, da innerhalb ein und derselben Sprache dieselben Features für unterschiedliche Wortarten auftreten. Diesen Ansatz erweitert das im SFB genutzten Annotationssystem, setzt jedoch für Kategorien und Features eigene Generationen an und lässt damit eine m:n-Verbindung zu.

Standardisierungsversuch von Annotationen

Darauf basierend wird herausgearbeitet, bis zu welchem Grad eine Standardisierung von Annotationssystemen, insbesondere auch sprachen- und varietätenübergreifend arbeitenden Systemen, möglich, handhabbar und sinnvoll ist. Schließlich soll untersucht werden, inwiefern mit Abweichungen von einer Normierung umgegangen werden kann, inwiefern Mehrsprachigkeit (auch im Sinne der „inneren Mehrsprachigkeit“, vgl. Wandruszka (1979)) die Standardisierungsmöglichkeiten beeinträchtigt, bzw. inwiefern Möglichkeiten bestehen, ein Annotationssystem zu modifizieren und zu erweitern, um trotz der großen und nicht trivialen Anforderungen einer Standardisierung standhalten zu können. Hierbei ist vor allem zu hinterfragen, ob eine Standardisierung des Vokabulars möglich und nötig ist oder ob vielmehr die Struktur und Beschreibung der zu standardisierende Aspekt der Annotation ist.

Forschungsfragen

Ziel des Vortrags ist es, sich den folgenden Forschungsfragen zu widmen:

- Welche Anforderungen entstehen an ein Annotationssystem für variationslinguistische Daten und insbesondere Nonstandardvarietäten und wie können diese erfüllt werden? Welche Anforderungen entstehen durch Methoden- und Theorienpluralismus?
- Inwiefern ist ein solches Framework standardisierbar bzw. welche Aspekte davon?
- Welche Vor- und Nachteile birgt ein standardisiertes Annotationssystem?

Bibliographie

Barbiers, Sjef / Wyngaerd, Guido Vanden (o.J.): *Tagging protocol*. <http://www.meertens.knaw.nl/pdf/variatielinguistiek/dialectsyntax/Tagging-protocol.pdf> [zuletzt abgerufen 13. Oktober 2018].

Breuer, Ludwig M. / Seltmann, Melanie E.-H. (2018): *Sprachdaten(banken) – Aufbereitung und Visualisierung am Beispiel von SyHD und DiÖ* in: **Börner, Ingo / Straub, Wolfgang / Zolles, Christian (eds):** *Germanistik digital*. Digital Humanities in der Sprach- und Literaturwissenschaft. Wien: Facultas, 135-152.

Budin, Gerhard / Elspaß, Stephan / Lenz, Alexandra N. / Newerkla, Stefan M. / Ziegler, Arne (2018): *Der Spezialforschungsbereich ‚Deutsch in Österreich (DiÖ). Variation – Kontakt – Perzeption‘* in: Zeitschrift für germanistische Linguistik 46(2), 300-308. DOI: 10.1515/zgl-2018-0017.

Lemmitzer, Lothar / Zinsmeister, Heike (2010): *Korpuslinguistik*. Tübingen: Gunter Narr Verlag.

Schiller, Anne / Teufel, Simone / Stöckert, Christine / Thielen, Christine (1999): *Guidelines für das Tagging deutscher Textcorpora mit STTS* (Kleines und großes Tagset) <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf> [zuletzt abgerufen 13. Oktober 2018].

Telljohann, Heike / Versley, Yannick / Beck, Kathrin / Hinrichs, Erhard / Zastrow, Thomas (2013): *STTS als Part-of-Speech-Tagset in Tübinger Baumbanken* in: **Zinsmeister, Heike / Heid, Ulrich / Beck, Kathrin (eds.):** *Das Stuttgart-Tübingen Wortarten-Tagset – Stand und Perspektiven*. Journal for Language Technology and Computational Linguistics 28, 1/2013. 1-15.

Wandruszka, Mario (1979): *Die Mehrsprachigkeit des Menschen*. München: Piper.