RUHR UNIVERSITÄT BOCHUM

RUB

F-AG 7: Angewandte Sprachwissenschaft, Computerlinguistik

Kurationsprojekt 2

Linguistische Annotation von Nichtstandardvarietäten — Guidelines und „Best Practices"

Guidelines Vorverarbeitung

Version 1.2

Stand: 24.11.13

Marc Reznicek

Marc.Reznicek@hu-berlin.de

Guidelines für die Vorverarbeitung bauen teilweise auf den Regeln für konkurrierende Zielhypothesen. (Reznicek et al. 2012)

Inhalt

Vorverarbeitung

Die Vorverarbeitung der in NoSta-D enthaltenen Daten gliedert sich in zwei Teile:

1. die Aufbereitung der Inputformate für die Annotation mit WebAnno
2. die Erstellung einer (oder mehrerer) Normalisierungsebenen.

Die Normalisierungsebene wird dann zuerst annotiert. Aus den Annotationen der Normalisierungsebene werden die Annotationen der Originaldaten abgeleitet.

1. Datenaufbereitung

Ziel der hier beschriebenen Datenaufbereitung ist es, die verschiedenen Inputformate, aus denen die NoSta-D-Subkorpora kompiliert wurden, in ein einheitliches und dadurch vergleichbar zu bearbeitendes Format zu bringen. Für den Import nach WebAnno wurde ein tab-separiertes Format, TSV (ähnlich CONLL10), verwendet. In der aktuell von WebAnno unterstützten Version, ist keine Speicherung der referenzielle Ketten möglich. Als Speicherformat wird daher das TCF-Format (http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/The_TCF_Format) verwendet, das von WebLicht (http://weblicht.sfs.uni-tuebingen.de/weblichtwiki/index.php/Main_Page) unterstützt wird.

1.1. Linearisierung

In WebAnno lassen sich Daten unterschiedlicher Sprecher nur dann annotieren, wenn die einzelnen Beiträge in linearisierter Form vorliegen. Sich überlappende Redeabschnitte wie in Abbildung 1 müssen daher erst linearisiert werden.

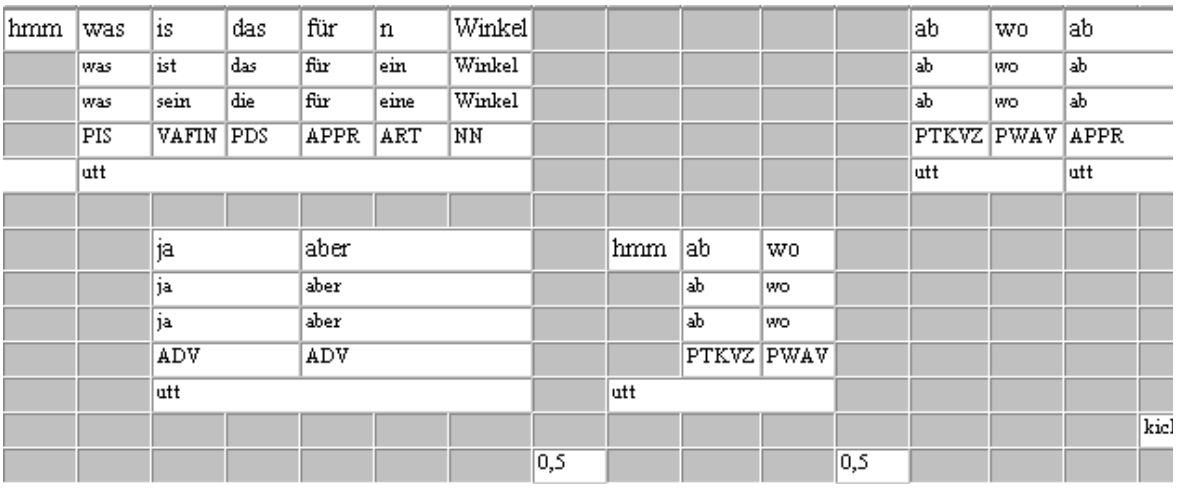

Abbildung 1:Sich überschneidende Sprechbeiträge in BeMaTaC 2011-12-14-A im EXMARaLDA Partitur-Editor

In NoSta-D werden daher alle Sprechbeiträge, die als Äußerungseinheiten verstanden werden können, als Blöcke nacheinander realisiert (siehe Abbildung 2).

hmm	was	is	das	für	n	Winkel						ab	wo	ab
							ja	aber	hmm	ab	wo			

Abbildung 2: Linearisierung sich überlappender Sprechbeiträge für NoSta-D-Annotation

Die Sequenz eines Sprechers wird nur dann durch den anderen Sprecher unterbrochen, wenn keine Überlappung vorhanden ist, oder der jeweils andere Sprecher auf einen parallelen Abschnitt reagiert.

1.2. Satzsegmentierung

Sind Satzgrenzen im Text markiert, so werden diese auch in der Vorverarbeitung respektiert. Dies kann in Einzelfällen zu sehr langen und komplizierten Strukturen führen (Kafka), anders als das Arborator-Tool ist die Annotation solch langer Abschnitte in WebAnno allerdings kein Problem.

Die Subkorpora ohne Satzsegmentierung (BeMatAC, unicum, AnselmBerlin) werden so segmentiert, dass Matrixsätze uns alle abhängigen Sätze soweit möglich ein einem Segment auftauchen. Miteinander (asyndetisch) koordinierte Sätze werden in einzelne Segmente getrennt, wobei die Konjunktion am Beginn jedes Segmentes steht.

B1_1v,16	biſchof tete ſente anſhel(=)
B1_1v,17	m⁹ bat marien manch
B1_1v,18	iar myt heyſen trenen·
B1_2r,01	das ſy ym offenbarte wy
B1_2r,02	vnſer here ihus criſtus
B1_2r,03	ſyne marter irleden hatte
B1_2r,04	do ſprach vnſe vrouwe Anſ=
B1_2r,05	helme ich ſage dir das
B1_2r,06	myn here iheſus criſtus·
B1_2r,07	alzo groſe martir irleden
B1_2r,08	hot· das ſy nyrkeyn men=
B1_2r,09	ſche us gelegen mak ⁊ Doch
B1_2r,10	ſaltu wiſſen· daz ich an ſo ta=
B1_2r,11	ne wirdekeit komen byn·
B1_2r,12	das ich nvmmermer be=
B1_2r,13	trubet mak werden ⁊ dar(=)

sente anshelmus bat marien manch iar myt heysen trenen das sy ym offenbarte wy vnser here ih-us cristus syne marter irleden hatte
do sprach vnse vrouwe Anshelme ich sage dir das myn here ihesus cristus alzo grose martir irleden hot . das sy nyrkeyn mensche usgelegen mak
Doch salt u wissen daz ich an sotane wirdekeit komen byn das ich nvmmermer betrubet mak werden

Abbildung 3: Segmentierung von AnselmBerlin Text ohne Satzgrenzenmarkierung in einzelne Matrixsätze für die Annotation in WebAnno.

Im Chat (NoSta-D_unicum) müssen nicht alle Wortformen eines Postings in ein Segment fallen. Emoticons, Interjektionen und Antwortpartikel werden als isolierte Segmente behandelt.

Eingebettete Parenthesen werden nicht aus dem Matrixsatz getrennt.

Obwohl dieser Abschnitt der Normalisierung vorangestellt ist, ist die erfolgreiche Segmentierung in einigen Fällen erst durch die Normalisierung möglich. So werden Fragmente gemeinsam in ein Segment aufgenommen, die ein gemeinsames Verb auf der Normalisierungsebene teilen, auch dann, wenn dieses Verb im Originaltext nicht vorhanden ist.

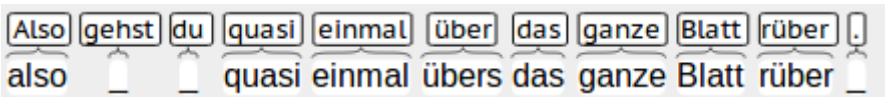

Abbildung 4: Segmentierung auf der Basis der Normalisierung

1.3. Tokenisierung

Die Tokenisierung wird aus den Subkorpora übernommen. Die Annotation verlangt allerdings in einigen Fällen eine Retokenisierung von Verschmelzungen und Konkatenierungen in den Daten.

Die retokenisierten Tokens erhalten den „|"-Charakter an der rechten Kante, wenn nach ihnen ein Subtoken abgetrennt wurde (siehe Abbildung 5).

1) l*ach| wech mich @ bochum*
2) 1 *an| ne stirn bapp*
3) * *na| gut| 50| cm| lauflaufleine* *

Entsprechend werden Zusammenschreibungen in der Normalisierung ebenfalls durch | angezeigt.

4) *Hier werden Beiträge von kleinen Leuten veraast, die von ehrenamtlichen Kassierern fünf| mark| weise gesammelt werden.*(NoSta-D_tuebadz-r8.0:23)

Die folgenden Sonderzeichen werden zusätzlich als Tokengrenzen annotiert

@ siehe Beispiel 1)

* siehe Beispiel 3)

2. Normalisierung

Ziel der Normalisierung ist es einerseits die Vergleichbarkeit zwischen den sehr unterschiedlichen Varietäten herzustellen, als auch die die Annotationen von nicht-kanonischen Strukturen zu motivieren und somit nachvollziehbar zu machen. Die Normalisierung ist so angelegt, dass sie dem Standard entspricht, und somit durch die gängigen Annotationsschemata ebenso gut abgedeckt ist, wie diese. In Teilen

(Explizitmachung von Ellipsen & Auslassungen, 2.5) geht sie allerdings darüber hinaus, sodass auch das Standardsubkorpus aus TüBa-D/Z eine Normalisierungsebene enthält.

Die Darstellung der Ebenen ist durch die Beschränkungen der verwendeten Annotationstools nicht optimal gelöst. Während Formate wie PAULA (Chiarcos et al. 2008) oder SALT (Zipser 2009) prinzipiell unbestimmt viele konkurrierende Tokenebenen und darauf aufsetzende Annotationen jeglicher Art (tokenbasiert, Spannen, Bäume, Pointer) abbilden können, ist dies für WebAnno bzw. Arborator nicht möglich. Die Umsetzung ist daher in NoSta-D bisher über verschiedene Dateien gelöst, in denen die jeweils andere Ebene als Annotation dargestellt wird.

Die in NoSta-D enthaltenen Annotationen verlangen teilweise mehrere Normalisierungsebenen (vgl. Reznicek et al. (erscheint) für ähnliche Anforderungen bzgl. L2-Lernersprache). Dies konnte im Rahmen des Kurationsprojektes allerdings nicht geleistet werden. Der Umgang mit den zusätzlichen Ebenen verhält sich aber prinzipiell genauso wie die hier beschriebene. Unterschiede werden an den relevanten Stellen in dieser Dokumentation erwähnt.

2.1. Löschungen & Einfügungen

Um dem Ansatz treu zu bleiben, dass jedes Token auf er Originalebene im syntaktischen Baum repräsentiert sein soll, muss auch für jedes Token eine Wortart annotiert sein. Da die Normalisierungsebene aber eine Standard-Entsprechung der Originaldaten darstellen soll, müssen in dieser teilweise Tokens hinzugefügt werden (Beispiel 5), die im Originaltext nicht enthalten sind. Problematisch wird es in Fällen, in denen eine grammatische Normalisierung keine Entsprechung vorsieht (Beispiel 6 & 7).

5) *Ob sie **sich** ethisch und moralisch nicht auszahlt, oder ob sie materiell keinen Gewinn zufügt?* (NoSta-D_fk002_2006_08)
6) *Aber wer wird in unserer fortgeschrittenen Gesellschaft den Migranten, die Bankräuber, die Mafia und jeden Verbrecher ~~zu~~ bestrafen?* (NoSta-D_fk002_2006_08)
7) *Die Rechte der Frauen, ~~um~~ zu arbeiten, ~~um~~ ein soziales Leben wie Männer zu haben, ~~um~~ eine gewählte (nicht gelittene) Sexualität dank der Kontrazeptionsmittel zum Beispiel zu haben, sind heute für uns natürlich, was vor fünfzig Jahren nicht der Fall war.* (NoSta-D_fk012_2006_07)

Um jedem Token der Originalebene eine Wortart zuweisen zu können, werden in er Normalisierungsebene Token auch dann nicht gelöscht, wenn ein Satz dadurch nicht mehr dem Standard entspricht. Folgeprojekte sollten in diesen Fällen eine weitere Normalisierung erarbeiten, in der diese Tokens gelöscht werden.

2.2. Orthographie

Auf der Normalisierungsebene wird die Orthographie auf die Standardschreibung der aktuellen Dudenversion (26. Auflage) angepasst. Dies gilt auch für das Kontrollkorpus (NoSta-D_tuebadz-r8.0).

8) *Wie| viel da monatlich fällig wird, weiß sie aber nicht.* (NoSta-D_tuebadz-r8.0:27)

9) *Die 88-jährige wurde vom rechten Außenspiegel erfa~~ß~~sst* (NoSta-D_tuebadz-r8.0:190)

Interjektionen werden phonetisch auf eine minimale Version reduziert. So werden beispielsweise alle Varianten in Abbildung Tabelle 1 links auf eine einheitliche Version rechts normalisiert.

oh	→	Oh
ohhhh	→	Oh
ohh	→	Oh
ohhhhh	→	Oh

Tabelle 1: Normalisierung von Interjektionen

2.3. Interpunktion

Die Interpunktion wird für alle Subkorpora im Sinne der neuen deutschen Rechtschreibung angepasst, dabei werden keine vorhandenen Zeichen gelöscht (siehe 2.1).

10) Wie ist es mit illegalen Ab- und Verhören, Überwachung und Lügen, im Namen der Gesellschaft und ihrer Sicherheit? (NoSta-D_fk002_2006_08)

Eine Ausnahme ist die Markierung von direkter Rede. Im Subkorpus NoSta-D_AnselmBerlin werden mehrere direkte Reden ineinander verschachtelt, sodass das jeweilige Ende nicht sauber getrennt werden kann. In diesen Fällen wurde die Einleitung der direkten Rede lediglich mit einem Doppelpunkt markiert. In Beispiel 11 sind die Ebenen direkter Rede eingerückt dargestellt.

11) *Da sprach unsere Frau:*

Sie führten ihn in eines Juden Haus, der hieß Annas. Dabei stand der Tempel. Da fragte Annas mein liebes Kind, was er die Leute lehrte.

Da sprach mein liebes Kind:

Ich habe offenbarlich gelehrt und nicht heimlich.

(NoSta-D_AnselmBerlin)

2.4. Morphologie & Syntax

In der Normalisierung wird die Originalwortstellung erhalten. Das bedeutet, dass die Normalisierung in NoSta-D im Gegensatz zu den Zielhypothesen, die in Reznicek et al. (2012) beschrieben werden, in vielen Fällen keine grammatische Entsprechung des Originaltextes

darstellt. Für die anschließende manuelle Dependenzanalyse kann die Wortstellung allerding vernachlässigt werden. Im Beispiel 11 wurden

12) *Doch sollst du wissen, dass ich an sotane Würdigkeit gekommen bin, dass ich nimmermehr* ***betrübt mag werden****.* (NoSta-D_AnselmBerlin)

Flexion und Derivation werden an die Anforderungen ihrer Mütter (in der anschließenden Dependenzanalyse angepasst. Dies bedeutet, dass sich die Normalisierung in Einzelfällen anhand der Annotation ändern kann. In der Abbildung 5 wird für den das Präpositionalobjekt für das Verb „telefonieren“ „an einen Staatsanwalt“ ersetzt durch „mit einem Staatsanwalt“.

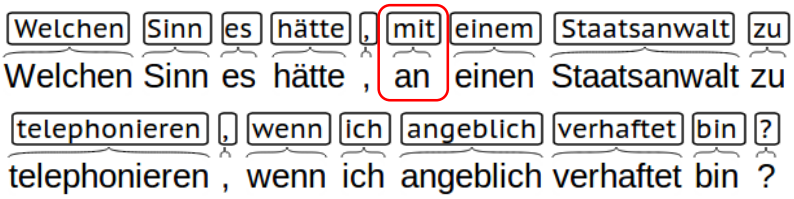

Abbildung 5: Anpassung der Argumente an die Argumentstruktur in Standard. Untere Ebene = Originaltext; Obere Ebene = Normalisierung

13) *Das muss abgeklärt werden, bevor man dazu Stellung nehmen kann, ob Interessen geschadet wurde oder diese genutzt worden sind.* (NoSta-D_cbs009_2006_09)

Die in Überschriften oft übliche Auslassung der Artikel wird korrigiert (Beispiel 12).

14) *Aber Bremerhavens AfB fordert jetzt einen Untersuchungsausschuß.* (NoSta-D_tuebadz-r8.0:37)

2.5. Ellipsen & Auslassungen

Um Fragmente in Dependenzstrukturen zu motivieren und um nicht an der Oberfläche realisierte Referenten annotieren zu können, werden in NoSta-D alle Ellipsen explizit in der Normalisierung eingefügt. Dies gilt für Argumente in koordinierten Sätzen (Abbildung 6), für Verben in Satzfragmenten (Abbildung 7) sowie für die Ergänzung von Argumenten für Inflektive (Abbildung 8).

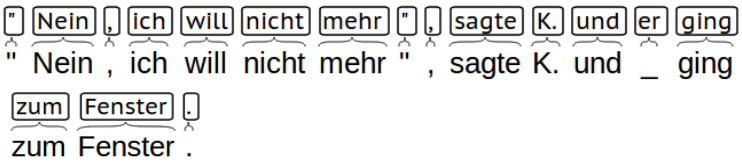

Abbildung 6: Ellipsen in Normalisierung explizit eingefügt: "er" im koordinierten Satz in Kafka

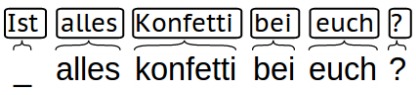

Abbildung 7: Explizitmachung elliptischer Verben in NoSta-D_unicum_21-02-2003_1

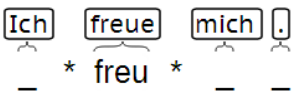

Abbildung 8: Ergänzung der Argumentstruktur von Inflektiven in NoSta-D_unicum_21-02-2003_1

Die eingefügten Tokens sollten so wenig lexikalische Information wie möglich mitbringen und so wenige nicht im Original vorhandene obligatorische Argumente mitbringen, die dann ebenfalls in der Normalisierung realisiert werden müssten.

Dabei sollte schrittweise in der folgenden Art vorgegangen werden:

1) Kann das eingesetzte Token durch einen Parallelismus aus dem Kontext identifiziert werden?
2) Wenn nicht, gibt es ein prototypisches Verb für die vorhandene Argumentstruktur und Bedeutung? Häufig: *„sein“*, *„haben“*, *„existieren“*, *„machen“*, *„geben“*, *„gehen“*, *„sagen“*
3) Wenn nicht, setze Dummy-Verben und Argumente ein! Häufig *„VERBen“*, *„VERBst“*, etc. sowie *„jemand“*, *„etwas“*, *„irgendwie“*, *„irgendwo“* etc.

Freie NPen in Grüßen und Wünschen werden in die Argumentstruktur prototypischer Verben eingebettet.

15) *Ich wünsche ein schönes Wochenende.* (NoSta-D_cbs009_2006_09)

In koordinierten Verbalphrasen werden die Auxiliare und Nomen nicht ergänzt.

16) *Er hatte* [die Beine übereinandergeschlagen] *und* [einen Arm auf die Rückenlehne des Stuhls gelegt]. (NoSta-D_Kafka: 141)

2.6. Abbrüche & Selbstkorrekturen

Die in NoSta-D realisierte Normalisierung erlaubt keine Löschung von Tokens, sondern lediglich Einfügungen. Dies liegt darin begründet, dass jedem Token im Original eine Dependenzkante zugewiesen werden können soll, was nur vermittels eines Normalisierungstokens geht. In einer NoSta-D-Erweiterung sollten aber auch weitere Normalisierungsebenen eingefügt werden, auf denen Tokens gelöscht werden können.

Wichtig wäre dies für die Annotation von Selbstkorrekturen. Diese werden bisher über die Dependenzkanten markiert. Es wäre aber eine bessere Lösung, diese als Spannen über die Differenz einer Normalisierung (auf der die korrigierten Token nicht repräsentiert sind) und dem Original abzuleiten. Weder abgebrochene noch selbstkorrigierte Strukturen werden durch fehlende Töchter ergänzt.

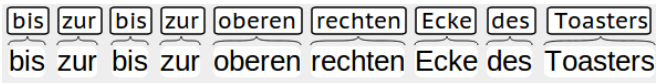

Abbildung 9: Keine Ergänzung von Abbrüchen und selbstkorrigierten Abschnitten in NoSta-D Bematac 2012-10-31-C

2.7. Lexik

Lexikalische Korrekturen werden auf der Normalisierungsebene möglichst vermieden, wenn eine grammatisch korrekte Lesart für den Originaltext gefunden werden kann.

17) *Trotzdem ist diese Theorie unvermeidbar, auch wenn eher praxisorientierte Überlegungen stattfinden.* (NoSta-D_cbs013_2006_09)

Für die historischen Texte wurden die Normalisierungen der Lexik aus den verfügbaren Quellen übernommen und lediglich orthographisch angepasst.

18)	*do*	*wurden*	*dy*	*iungere*	*entslofen*	.
	Da	wurden	die	Jünger	entschlafen	.

(NoSta-D_AnselmBerlin)

Dies gilt für:

- NoSta-D_AnselmBerlin (http://www.linguistics.ruhr-uni-bochum.de/anselm/)
- NoSta-D_DDB (http://korpling.german.hu-berlin.de/ddb-doku/index.htm)

2.8. Referenz

Als Vorverarbeitungsschritt für die Annotation von Koreferenz wird in der vorliegenden Normalisierung die Referenz kontextgeleitet angepasst.

19) *Wenn **man** sich mit dieser Frage im Rahmen der Ethik beschäftigt, wird ~~er~~man fast auf jeden Fall sagen, dass Kriminalität sich nicht auszahlt.* (NoSta-D_fk002_2006_08)

3. Literatur

Chiarcos, Christian; Dipper, Stefanie; Götze Michael; Ritz, Julia; Stede, Manfred (2008): A Flexible Framework for Integrating Annotations from Different Tools and Tagsets. In: Proceeding of the Conference on Global Interoperability for Language Resources, Hong Kong, January 2008.

Reznicek, Marc; Lüdeling, Anke; Schwantuschke, Franziska; Walter, Maik; Schmidt, Karin; Hirschmann, Hagen et al. (2012): Das Falko-Handbuch. Korpusaufbau und Annotationen. Version 2.01. Institut für deutsche Sprache und Linguistik, Humboldt-Universität zu Berlin. Berlin.https://www.linguistik.hu-

berlin.de/institut/professuren/korpuslinguistik/forschung/falko/Falko-Handbuch_Korpusaufbau%20und%20Annotationen_v2.01.

Scholze-Stubenrecht, Werner (2013): Duden, die deutsche Rechtschreibung. 26., völlig neu bearb. und erw. Aufl. Mannheim [u.a.]: Dudenverl. (Der Duden in zwölf Bänden 1).

Zipser, Florian (2009): Entwicklung eines Konverterframeworks für linguistisch annotierte Daten auf Basis eines gemeinsamen (Meta-)modells. Diplomarbeit. Humboldt-Universität zu Berlin, Berlin. Institut für Informatik. http://hal.archives-ouvertes.fr/docs/00/60/61/02/PDF/Diplomarbeit_FZ_final.pdf

Alle Quellen wurden geprüft am 27.09.2013