

## Zur Syntax von Plauderchats

Burkhard Dietterle\*, Anke Lüdeling\*, Marc Reznicek†

\*Humboldt-Universität zu Berlin, † DAAD Madrid

### 0. Einleitung, Motivation

Dieser Beitrag beschäftigt sich mit zwei eng miteinander verbundenen Fragen:

Wie können die syntaktischen Strukturen in Chat-Texten beschrieben werden?

Welche syntaktischen Eigenschaften haben deutsche Chat-Texte?

Chats (und hier insbesondere sogenannte ‚Plauderchats‘) weichen in vielerlei Hinsicht von einer schriftlichen ‚Standardsprache‘ ab. Uns interessiert in diesem Beitrag vor allem die Syntax von Äußerungen<sup>1</sup> aus Plauderchats. In Abschnitt 1 werden wir zunächst kurz auf einige Grundannahmen von syntaktischen Beschreibungen eingehen und erläutern, warum diese für die Analyse von Chatdaten nicht immer geeignet sind. Dann, in Abschnitt 2, werden wir die Chatdaten aus dem NoSta-D-Korpus und ihre Vorverarbeitung vorstellen, bevor wir in Abschnitt 3 auf einige syntaktische Eigenschaften der Daten genauer eingehen.

### 1. Nichtstandardsprache: syntaktische Eigenschaften

Der Begriff Standardsprache ist problematisch<sup>2</sup>, aber was auch immer unter ‚Standardsprache‘ verstanden wird, die Sprache in Plauderchats wird darunter sicher nicht mitverstanden.

Fast alle Grammatiken (unabhängig davon, ob sie deskriptiv oder theoretisch ausgerichtet sind und welcher Theorie sie folgen) beschreiben konzeptuell geschriebene, geplante Register. Als wesentliche Einheit der syntaktischen Beschreibung wird immer ein Satz angenommen. Auch für den Begriff ‚Satz‘ gibt es unterschiedliche Definitionen, man kann aber verallgemeinernd sagen, dass als Satz ein finites Verb samt aller abhängigen Konstituenten (Argumente, Komplemente, Ergänzungen) sowie ggf. vorhandenen Modifikatoren (Adjunkte, Angaben) definiert wird. In vielen Registern und Varietäten des Deutschen kommen nun jedoch Äußerungen vor, die gemäß solch einer Definition keine Sätze sind und daher nicht wie Sätze analysiert werden dürfen. Dies betrifft beispielsweise Äußerungen mit Unflüssigkeiten (disfluencies, vgl. z.B. Eklund 2004 für einen Überblick) ebenso wie Äußerungen ohne Verb (Behr/Quintin 1996), Äußerungen ohne Subjekt (Schalowski 2009) oder andere nach Standardgrammatiken ‚unvollständige Äußerungen‘ (vgl. Reich 2011 zum Unterschied zwischen Ellipsen, die aus dem Satzkontext ergänzt werden können, und Auslassungen, in denen die Situation für das Verständnis bekannt sein muss). Wir möchten dies kurz an einem kurzen Abschnitt aus dem von uns behandelten Chat illustrieren (für die Referenz siehe Abschnitt 2). Hier unterhalten sich Chatteilnehmer TomcatMJ und Bochum über einen Umzug. Die Äußerung in Post 223 enthält kein Verb, trotzdem kann man sie problemlos verstehen. Eine Möglichkeit, dies zu erklären, ist, dass man implizit eine ‚kanonische‘ Syntax annimmt, die ein Verb enthält und beispielsweise aussehen könnte wie die Struktur in (1b).

(1a)

221	eigentlich kostet so  was die miete für nen lkw für nen tag und verköstigung für nen haufen helfender freunde @ zora
222	bochum-münster ohne küche 3500 euro ..

(1b)

222	[Bochum-Münster ohne Küche] <sub>SUBJ</sub> kostet [3500 Euro] <sub>MOD</sub> .
-----	---

<sup>1</sup> Wir verwenden den Begriff ‚Äußerung‘ hier prätheoretisch. Technisch sind Chats in Posts eingeteilt, ein Post kann mehrere Äußerungen enthalten. Für eine Modellierung der Struktur von Chat-Daten vgl. Beißwenger et al. (2012).

<sup>2</sup> Eine Aufarbeitung der Thematik ‚Standardsprache‘ sprengt den Rahmen dieses Beitrags. Nur wenige Grammatiken thematisieren ihre Beschreibungsgrundlage überhaupt (Eisenberg 2007 definiert Standardsprache als die Sprache überregionaler Tageszeitungen, die Dudengrammatik 2005 (Duden 4) hat einen Abschnitt zu gesprochener Sprache, geht aber in allen anderen Abschnitten auch implizit von einer geschriebenen, geplanten Varietät mit vollständigen Sätzen aus). Diskutiert wird die Beschreibungsgrundlage von Grammatiken und der Begriff ‚Standardsprache‘ ausführlicher in Beschreibungen von bestimmten ‚Nichtstandard‘-Varietäten, vgl. z. B. Hennig (2006) oder Maas (2010) zu gesprochener Sprache oder Wiese/Freywald/Mayr (2009) zu Kiezdeutsch.

Ohne Verb kann man zwar den einzelnen Bestandteilen eines Satzes bestimmte Informationen zuordnen (z.B. Flexionsinformationen wie Kasus), aber eigentlich keine grammatischen Funktionen wie Subjekt oder Modifikator, da die grammatischen Funktionen inhärent relational sind (ein Subjekt ist immer ein Subjekt von einem Verb).<sup>3</sup> Wenn einem nur unverbundene ‚Chunks‘ oder Phrasen als Analysemittel zur Verfügung stehen, bleibt viel Information unzugänglich. Die Ergänzung in (1b) wird normalerweise implizit vorgenommen und es gibt oft viele Möglichkeiten, eine Äußerung zu ergänzen. Wir werden in Abschnitt 2 eine Methode vorstellen, in der man solche Ergänzungen explizit macht.

Konzeptuell kann man bei der Beschreibung von solchen ‚Nichtstandardvarietäten‘ zwei Wege gehen, die beide in der Literatur vorgeschlagen und diskutiert werden:

- 1) Man formuliert eine spezifische Grammatik je Varietät.
- 2) Man formuliert eine übergreifende Grammatik und versteht die Varietäten als spezifische Instanzierungen dieser einer Grammatik (spezifisch in Bezug auf die Gewichtung der von der Grammatik zur Verfügung gestellten Strukturen).

Wenn man für jede Varietät eine spezifische Grammatik annimmt, kann man alle Besonderheiten bis ins kleinste Detail analysieren und entgeht der sog. comparative fallacy (Bley-Vroman 1983), d.h. der Versuchung, eine Varietät durch unpassende und unzutreffende Kategorien einer anderen Varietät zu beschreiben. Man kann dann allerdings Varietäten nicht direkt und vor allem nicht quantitativ (zum Beispiel für Registerstudien) miteinander vergleichen. Daher wählen wir hier einen Ansatz, der zunächst aussieht wie Option 2: Wir formulieren eine kanonische Ebene (Normalisierungsebene oder auch Zielhypothese genannt)<sup>4</sup>, die es uns ermöglicht, alle Varietäten – also einschließlich (statt ausschließlich!) Chat – auf die gleiche Weise zu beschreiben (vgl. Hirschmann/Doolittle/Lüdeling 2007). Dabei ist es uns wichtig, dass dies eine rein methodische Entscheidung ist und wir damit nichts über eine kognitive ‚Wahrheit‘ aussagen wollen. Auch soll unser Ansatz nicht sprachpflegerisch verstanden werden, unsere kanonischen Ebenen sind nicht als ‚Verbesserungen‘ zu verstehen.

Viele der vorliegenden theoretischen oder deskriptiven grammatischen Modelle können nicht alle Phänomene, die in Chat vorkommen, adäquat erfassen. Durch den Abgleich mit einer kanonischen Ebene können wir jedoch diejenigen Stellen finden, die grammatisch interessant sind. Um die Eigenschaften von Chat gut zu verstehen, müssen wir auch wissen, wie sich Chat von anderen Varietäten – wie zum Beispiel von gesprochenen Varietäten oder Zeitungssprache – unterscheidet.

Unser Korpus mit Originaltexten und zugehörigen Annotationen wird in einer Mehrebenenarchitektur bereitgestellt, die es möglich macht, alle Originaltext-Informationen erhalten und beliebige weitere Annotationsebenen hinzufügen zu können. Die Hinzufügung einer kanonischen Ebene bedeutet also nicht, dass der Originaltext verändert wird, sondern dass der Originaltext um eine explizite Interpretation ergänzt wird. In Abschnitt 3 zeigen wir einige Eigenschaften von Chat-Daten, die sich so finden und beschreiben lassen. Vorher möchten wir die Aufbereitung der Daten genauer erläutern.

## 2. Annotation der Chat-Daten in NoSta-D

Die Chat-Daten, die im Folgenden beschrieben werden, stammen aus dem Dortmunder Chat-Korpus (DCK, Beißwenger 2013) und wurden als ein Teil von NoSta-D<sup>5</sup> (Dipper/Lüdeling/Reznicek 2013) aufbereitet. Zweck des NoSta-D-Korpus ist der quantitative und qualitative Vergleich der darin enthaltenen fünf Nichtstandardvarietäten und einer Standardvarietät<sup>6</sup>. NoSta-D-unicum, das Chat-Subkorpus von NoSta-D enthält 11.312 Tokens in 787 Segmenten aus einer Sequenz (unicum\_21-02-2003\_1) eines unmoderierten Chat-Forums, eines sogenannten Plauderchats. In Abschnitt 3 werden wir diese Daten verglichen mit Daten aus NoSta-D-bematac<sup>7</sup>, einem Korpus mit gesprochenen Map-Task-Dialogen, und NoSta-D-tuebadz, einem Ausschnitt aus dem TüBa-D/Z-Korpus mit Zeitungssprache. Alle Subkorpora

<sup>3</sup> Dies ist ein wenig vereinfacht dargestellt: Ein Subjekt ist immer ein Subjekt von einem Prädikat. Prädikate sind in den meisten Sprachen und auch im Deutschen häufig, aber nicht notwendig Verben.

<sup>4</sup> Die Grundidee für diese Normalisierungsvariante stammt aus der Annotation von Lernerdaten. Vgl. Lüdeling (2011) und Reznicek/Lüdeling/Hirschmann (2013) für eine ausführliche Motivation und Beschreibung.

<sup>5</sup> NoSta-D Projekt-Webseite, siehe Abschnitt „Quellen“

<sup>6</sup> Das Korpus entstand im Rahmen des Clarin-D Kurationsprojektes „Linguistic Annotation of Non-standard Varieties — Guidelines and Best Practices“. Siehe Abschnitt „Quellen“

<sup>7</sup> Zu BeMaTac siehe Sauer/Lüdeling (erscheint).

wurden gemäß eigens für NoSta-D zusammengestellten Richtlinien für Vorverarbeitung und Annotation<sup>8</sup> erstellt.

Die Kernidee dieser Richtlinien ist es, Interpretation und Annotation des Originaltextes nicht in einem Schritt zu vereinen, sondern in zwei Schritte zu trennen:

- 1) Eine Interpretation des Originaltextes wird in der bereits erwähnten kanonischen Ebene fixiert.
- 2) Sämtliche Annotationen am Originaltext werden nicht durch den Originaltext selbst motiviert, sondern durch die kanonische Ebene.

In Schritt 1) werden die Originaltexte aus den verschiedenen Varietäten auf Texte einer Normalisierungsvarietät abgebildet, wobei die Abbildung nach ausgearbeiteten Richtlinien für alle Originaltexte gleichermaßen erfolgt. Bei den normalisierten Texten handelt es sich in keiner Weise um „korrekte“ Versionen der Originale. Sie dienen, wie oben beschrieben, lediglich als *Tertium Comparationis*.

## 2.1 Vorverarbeitung

Die Annotation syntaktischer Strukturen hängt in hohem Maße vom Ergebnis der Vorverarbeitung der Daten ab. Für NoSta-D-unicum sind hier folgende Schritte entscheidend: Linearisierung, Satzsegmentierung, Tokenisierung, Normalisierung und Wortarten-Tagging.

### 2.1.1 Linearisierung

Die Originaldaten des DCK liegen in einem HTML-Tabellenformat vor. Schreiberalias und Schreiberbeitrag (Post) stehen in getrennten Spalten, wie in (2) zu sehen. In unserer Version ist der Schreiberalias gelöscht, bei der Linearisierung entfallen also die Sprecherwechsel.

(2)

1	<b>System</b>	JustChat 4.0r0.204 (55.204) developed by Medium.net.
2	<b>System</b>	Du betrittst den Raum.
3	<b>Quaki</b>	was echt zori?
4	<b>System</b>	little15 betritt den Raum.
5	<b>Quaki</b>	das küssen??
6	<b>Pharao</b>	na gut marc. kein servicepaket nr.1 für dich :)
7	<b>Zora</b>	was?
8	<b>System</b>	TomcatMJ kommt aus dem Raum Go-Rin-No-Sho herein.
9	<b>TomcatMJ</b>	Hi
10	<b>System</b>	TomcatMJ ist wieder da.

### 2.1.2 Satzsegmentierung

Wie in Abschnitt 1 ausgeführt, ist für die meisten grammatischen Beschreibungen die wesentliche Einheit ein Satz. Auch in NoSta-D werden Texte in Sätze segmentiert. Als Satzsegment gilt dabei eine kontinuierliche Tokenkette, über deren sämtliche Token ein (und genau ein) Dependenzgraph aufgespannt wird. Einem NoSta-D-Satzsegment entsprechend also kanonische Sätze im Sinne von Matrixsätzen inklusive aller ihrer abhängigen Sätze. Viele Posts in NoSta-D-unicum sind kanonische Sätze und können ohne weiteres als Satzsegmente in NoSta-D übernommen werden. Natürlich sind auch miteinander koordinierte Sätze gemäß Standardgrammatiken kanonisch und natürlich können koordinierte Sätze auch in einem NoSta-D-Satzsegment abgebildet werden. Auch in diesem Fall wird ein (und nur genau ein) Dependenzgraph über die gesamte Tokenkette aufgespannt.

Miteinander koordinierte Sätze werden in den geschriebenen Varietäten (also auch in den Chatdaten) dann über eine Koordinierungssatzkante (CS) miteinander verbunden, wenn sie durch Satzzeichen als asyndetische Koordination markiert sind. Ansonsten werden sie getrennt. Da die gesprochene Varietät diese Unterscheidung nicht zulässt, wurden hier alle Kandidaten für asyndetische Koordination getrennt. Auf diese Weise geht keine Information verloren und die Vergleichbarkeit kann weiterhin hergestellt werden.

<sup>8</sup> Siehe NoSta-D Projekt-Webseite.

Problematisch wird es, wenn eine Äußerung kein finites Verb enthält und daher nicht als Satz, sondern lediglich als Sequenz von (infiniten, nominalen, adverbialen usw.) Fragmenten analysiert werden kann. In solchen Fällen wird die Satzsegmentierung anhand der kanonischen Ebene motiviert, in die Verben ergänzt werden (siehe Abschnitt 2.2.3). So werden Fragmente in ein Satzsegment aufgenommen, die sich ein gemeinsames Verb auf der kanonischen Ebene teilen (siehe (3) aus NoSta-D-bematac). Im Übrigen kann auch die Segmentierung kanonischer Sätze als eine anhand ihrer Normalisierung motivierte verstanden werden, mit der Besonderheit, dass das Verb in der kanonischen Ebene nicht ergänzt, sondern aus dem Originaltext übernommen wird.

(3)

<b>Normalisierung:</b>	<b>Also</b>	gehst	du	quasi	einmal	über	das	ganze	Blatt	rüber	
<b>Original:</b>	<b>also</b>	_	_	quasi	einmal	über	das	ganze	Blatt	rüber	

### 2.1.3 Tokenisierung

Eine auffällige Eigenschaft von Chat-Sprache besteht in der Häufung von nichtkanonischen Zusammenschreibungen.<sup>9</sup> Diese können in einem Korpus auf drei unterschiedliche Weisen tokenisiert werden.

Als ein Token, das dann als komplexe syntaktische Einheit, die im Satz mehrere Funktionen erfüllen kann, zu beschreiben wäre. Beispielsweise ist das Element *sparste* in (4) als ein Token mit zwei syntaktischen Kategorien repräsentiert.

(4)

<b>Token</b>	<i>sparste</i>	<i>den</i>	<i>umzug</i>
<b>syntaktische Funktionen<sup>10</sup></b>	<b>VERB + SUBJ</b>	DET	OBJA

Als zwei Token, die dann als zwei einfache syntaktische Einheiten zu repräsentieren wären. So wird *sparste* in (5) in die beiden Token *sparst* und *e* getrennt, denen jeweils eine Kategorie zugeordnet ist.<sup>11</sup>

(5)

<b>Token</b>	<i>Sparst</i>	<i>e</i>	<i>den</i>	<i>umzug</i>
<b>syntaktische Funktionen</b>	<b>VERB</b>	<b>SUBJ</b>	DET	OBJA

Während in der ersten Tokenisierung beliebig viele (auch unvorhergesehene) kombinierte Tags entstehen könnten (zum Beispiel, weil es auch Kombinationen aus mehr als zwei Elementen geben kann oder die Kombinationen nicht vorhersehbar sind, siehe Abschnitt 3.1) und in der zweiten Tokenisierung die Information über die ursprüngliche Schreibung verloren geht, haben wir uns in NoSta-D für eine dritte Lösung entschieden. Wir trennen *sparste* in syntaktische Einheiten auf, markieren aber die ursprüngliche Zusammenschreibung durch einen senkrechten Strich. Jede Einheit bekommt eine eigene syntaktische Kategorie zugewiesen (siehe (6)).

(6)

<b>Token</b>	<i>sparst </i>	<i>e</i>	<i>den</i>	<i>umzug</i>
<b>syntaktische Funktionen</b>	<b>VERB</b>	<b>SUBJ</b>	DET	OBJA

<sup>9</sup> Vergleichbar mit einigen Zusammenschreibungen sind Verschmelzungen oder Klitisierungen in der gesprochenen Sprache, siehe Abschnitt 3.1.

<sup>10</sup> Die hier angedeutete Ebene „syntaktische Funktionen“ ist im Korpus nicht als Spanne wie im Beispiel, sondern als Label für die Kanten eines Abhängigkeitsgraphen gelöst. Die Abbildung dient nur zur Illustration der Tokenisierungsproblematik.

<sup>11</sup> Der Plosiv kann sowohl zum Verb als auch zum Pronomen gehören. Um die (unbeantwortbare) Frage, wo genau segmentiert wird, geht es hier nicht. Uns ist wichtig, dass es zwei Elemente mit grammatischen Kategorien gibt.

#### 2.1.4 Normalisierung

Wie bereits erwähnt, dient die NoSta-D-Normalisierungsebene als Grundlage für eine vergleichende Analyse über unterschiedliche Varietäten hinweg. Gleichzeitig motiviert sie die Annotationsentscheidungen für die Originaldaten, deren Strukturen ansonsten zu wenige Hinweise für eine Entscheidung zwischen konkurrierende Annotationen geben würden.

Die wichtigsten Schritte bei der Erstellung der Normalisierung für das Chat-Subkorpus sind:

- Angleichung der Orthographie und Interpunktion an den Duden (Duden 1, 2013) und Abbildung von reduzierten Formen auf Vollformen, wie in (7)
- Standardisierung von Interjektionen und Namen der Chatteilnehmer (siehe (8))
- Auffüllung von Ellipsen und Auslassungen. Um die Annotation von Fragmenten zu motivieren, werden in NoSta-D Ellipsen bzw. Auslassungen explizit in der Normalisierung aufgefüllt. Das gilt für nichtrealisierte Verben in Satzfragmenten wie in (9) sowie für nichtrealisierte Argumenten in Inflektiven wie in (10). Auf Ellipsen und Auslassungen gehen wir in Abschnitt 3.3 näher ein. Die lexikalische und syntaktische Information der eingefügten Elemente wird entweder durch Parallelismus aus dem Kontext identifiziert, oder durch allgemeine, semantisch relativ schwache Verben (z.B. *sein*, *haben*) bzw. Dummy-Verben („VERBen“, „VERBst“) gefüllt.

(7)

<b>Normalisierung:</b>	Lantonie	redet	wie	<b>eine</b>	Bewährungshelferin	.
<b>Original:</b>	Lantonie	redet	wie	<b>ne</b>	Bewährungshelferin	_

(8)

Original		Normalisierung
ohhhh	→	Oh
oh	→	Oh
ohhhhhhh	→	Oh
lantonieeeee	→	Lantonie
LANTOOO	→	Lantonie
Lantööö	→	Lantonie

(9)

<b>Normalisierung:</b>	<b>Ist</b>	alles	Konfetti	bei	euch	?
<b>Original:</b>	_	alles	konfetti	bei	euch	?

(10)

<b>Normalisierung:</b>	<b>Ich</b>		freue		<b>mich</b>	.
<b>Original:</b>	_	*	freu	*	_	_

## 2.2 Syntaktische Annotation

### 2.2.1 Wortartenannotation

Die NoSta-D-Korpora werden mit dem STTS-Tagset (Schiller et al. 1999) annotiert. Wie in den übrigen Annotationsschritten wird die auch Wortartannotation des Originaltexts durch eine entsprechende Annotation der Normalisierung motiviert. So wird beispielsweise das Wortartentag PPER für *dudu* in (11) aus der Normalisierung (*du*) übernommen. Eine Ausnahme wird bei Inflektiven gemacht: diese erhalten in der Normalisierung das POS-Tag VVFIN, aber im Originaltext VVIN, wie in (12) gezeigt.<sup>12</sup>

<sup>12</sup> Das STTS wurde für konzeptionell geschriebene Sprache entwickelt und deckt viele Formen aus gesprochener Sprache, internetbasierter Kommunikation und anderen Varietäten nicht adäquat ab. Daher gibt es inzwischen Initiativen, die das STTS entsprechend erweitern wollen. Für Chat-Daten liegt dazu ein Vorschlag von Bartz/Beißwenger/Storrer (2014) vor. Da das NoSta-D-Korpus aber unterschiedliche Varietäten vergleichen will, wird hier nur das ursprüngliche STTS angenommen.

(11)

<b>Normalisierung:</b>	Ja	,	das	bist	<b>du</b>	.
<b>Norm-POS:</b>	PTKANT	\$,	PDS	VAFIN	<b>PPER</b>	\$.
<b>Original:</b>	jepp	–	–	bist	<b>dudu</b>	–
<b>ORIG-POS:</b>	PTKANT			VAFIN	<b>PPER</b>	

(12)

<b>Normalisierung:</b>	Ich	mal	<b>gucke</b>	,	wo	Quaki	sich	nun
<b>Norm-POS:</b>	PPER	ADV	<b>VVFIN</b>	\$,	PWAV	NE	PRF	ADV
<b>Original:</b>	–	mal	<b>guck</b>	–	wo	quaki	sich	nu
<b>ORIG-POS:</b>		ADV	<b>VVINF</b>		PWAV	NE	PRF	ADV

### 2.2.2.2 Dependenzannotation

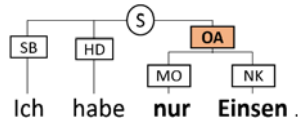
Für die Annotation syntaktischer Strukturen sind sehr unterschiedliche Modelle entwickelt worden, die unterschiedliche Stärken und Schwächen mit sich bringen (vgl. Frank 2013). Für die Analyse von Nichtstandardsprache haben sich Dependenzmodelle allerdings als vorteilhaft herausgestellt (Kübler/Prokic 2006; Nivre et al. 2007). Daher verwenden wir auch für das NoSta-D-Korpus ein Dependenzmodell. Für deutsche Zeitungssprache gibt es bereits sehr weit entwickelte automatische Dependenzparser mit einer hohen In-domain accuracy, also einer hohen Zuverlässigkeit für trainingsähnliche Daten (u.a. der MaltParser von Nivre et al. 2007, der Stanford Parser von Rafferty/Manning 2008, der MATE Parser von Bohnet 2010). Die Modelle ziehen ihr „Wissen“ vor allem aus statistischen Maßen, die zuvor in einer Trainingsphase aus annotierten (Zeitungskorpora) extrahiert wurden. Für Out-of-domain-Daten müssen neue Trainingskorpora erstellt und die Modelle neu trainiert werden. Für unsere Nichtstandardvarietäten gibt es bisher keine adäquaten Trainingskorpora, so dass wir unsere Daten manuell annotieren müssen.

Richtlinien für die manuelle Annotation von Dependenzstrukturen wurden zwar in Foth (2006) vorgeschlagen, diese Regeln decken aber selbst einen erheblichen Teil der in TüBa-D/Z vorhandenen sprachlichen Strukturen nicht befriedigend ab. Zwei andere für das Deutsche sehr viel umfassendere Richtlinien für die syntaktische Analyse sind Telljohann et al. (2005) für die TüBa-Baumbank und Albert et al. (2003) für das TIGER-Korpus. In beiden Fällen handelt es sich allerdings um hybride Konstituentenstrukturbäume, für deren automatische Übersetzung in eine Dependenzstruktur zwar Werkzeuge entwickelt wurden (vgl. Forst et al. 2004, Seeker/Kuhn 2012), die aber menschlichen Annotatoren keine Entscheidungshilfe bei der Annotationen von nicht-kanonischen Daten bieten.

Die Dependenz-Annotation des Originaltexts in NoSta-D wird, wie in den anderen Annotationsschritten, aus einer Dependenz-Annotation der Normalisierung abgeleitet. Diese wiederum wird aus einer TIGER-Annotation der Normalisierung übersetzt<sup>13</sup>. Eine unmittelbare Dependenz- oder TIGER-Annotation des Originaltextes wäre aufgrund der Nichtkanonizität nicht konsequent durchhaltbar. So muss die Äußerung *nur einsen* (NoSta-D-unicum, Post 65) abhängig von der Normalisierung beispielsweise als Akkusativobjekt (13a) oder als Subjekt (13b) annotiert werden (gemäß den TIGER-Konventionen).

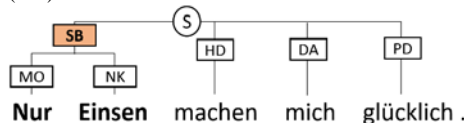
<sup>13</sup> Ausnahmen gibt es für Fälle, in denen TIGER vom STTS abweichende Wortartenannotationen zugrundelegt (siehe das NoSta-D-Annotationsschema).

(13a)



Normalisierungsvariante für *nur einsen* als Akkusativobjekt.

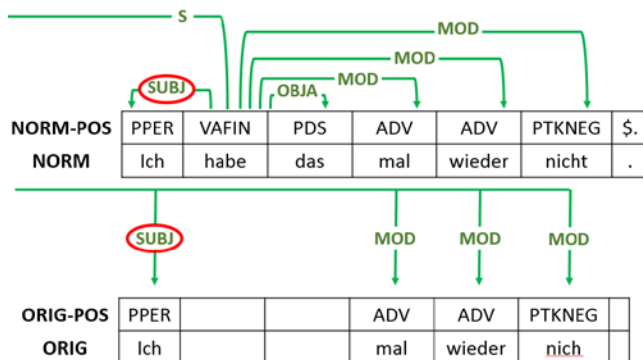
(13b)



Normalisierungsvariante für *nur einsen* als Subjekt.

Im Folgenden erklären wir die Übertragung der Annotation von der Normalisierung auf die Originaldaten. In vielen Fällen kann die Annotation der Normalisierung 1:1 auf die Originaldaten übertragen werden, da die Originalsätze syntaktisch nicht auffällig sind und vielleicht orthographisch und morphologisch, aber nicht syntaktisch normalisiert wurden. In Abschnitt 3.3 kommen wir auf diese Fälle unter der Bezeichnung parallele Annotationen zurück. In einigen Fällen kann die Annotation der Normalisierung *nicht* 1:1 auf die Originaldaten übertragen werden, weil nicht nur orthographisch und morphologisch, sondern auch syntaktisch normalisiert wurde. Ein Beispiel für eine solche abweichende Annotation findet sich in (14). Wie in Abschnitt 1 angesprochen, sind syntaktische Rollen wie Subjekt, Objekt oder (Verb-)Modifikator ohne Verb nicht zuweisbar. In (14) sieht man, dass im Originaltext für *Ich* das Dependenzlabel SUBJ (Subjekt) aus der Normalisierung übernommen wird, die Dependenzkante aber nicht an ein Verb (es gibt ja keins), sondern an die Segmentwurzel<sup>14</sup> gebunden wird.

(14)



Die in diesem Abschnitt dargestellte Vorverarbeitung und syntaktische Annotation erlaubt es, syntaktische Phänomene zwischen so unterschiedlichen Varietäten wie Zeitungstexten, gesprochener Sprache und Chat quantitativ zu vergleichen. In Abschnitt 3 werden wir Chat-Syntax daher hauptsächlich nicht anhand von Strukturen beschreiben, die nur in Chat zu finden sind (Emoticons etc.), sondern mithilfe weiterer NoSta-D-Subkorpora die relativen Häufigkeiten syntaktischer Muster gegenüberstellen.

### 3. Syntaktische Strukturen im NoSta-D Chat

Die meisten der zahlreichen Artikel, die sich aus linguistischer Sicht mit Chat befassen, beschäftigen sich mit lexikalischen oder orthographischen Aspekten (so zum Beispiel Myslin/Gries 2010), mit graphisch auffälligen Eigenschaften wie Emoticons, Inflektiven und Asterisk-Ausdrücken (Teuber 1998, Schlobinski 2001) oder mit pragmatischen Phänomenen (siehe dazu die Artikel in Herring/Stein/Virtanen 2013). Zur Syntax von Chats gibt es weniger Arbeiten. Allgemein wird oft angenommen, dass (Plauder-)Chats

<sup>14</sup> Im Gegensatz zu allen anderen uns bekannten Dependenzschemata ist bei uns das finite Verb der höchste Regent, sondern eine phonetisch leere „Segmentwurzel“, an die realisierte finite Verben oder eben Dependente von nicht realisierten Verben gebunden werden.

konzeptionell mündlich seien (vgl. Storrer 2001, 2013), ohne dass entsprechend zu erwartende mündliche Eigenschaften in der Chat-Syntax nachgewiesen würden. Die erwähnten Emoticons, Asterisk-Ausdrücke und Inflektive lassen annehmen, dass die Syntax von Chats eher flach sei, denn diese Phänomene verhalten sich offensichtlich asyntaktisch: Sie weisen wenig bis keine Kombinatorik auf.<sup>15</sup> Dennoch greift die Annahme, Chatsyntax sei konzeptionell mündlich (weil flach), zu kurz: Einerseits gibt es in jedem Chat viele syntaktisch vollständig kanonische Äußerungen, was auf konzeptionelle Schriftlichkeit hinweisen könnte, und andererseits treten bestimmte Merkmale von Mündlichkeit (bestimmte Unflüssigkeiten) gar nicht auf.

Wir werden drei Aspekte der Chat-Syntax behandeln – nichtkanonische Zusammenschreibungen, syntaktische Komplexität und Unflüssigkeiten – und diese werden wir jeweils mit Daten aus dem NoSta-D-Subkorpora für gesprochene Sprache (NoSta-D-bematac) sowie mit Daten aus dem NoSta-D-Subkorpora für Zeitungssprache (NoSta-D-tuebadz) vergleichen. Wir möchten untersuchen, welche syntaktischen Aspekte in diesen drei Gebieten eher konzeptionell mündlich sind, welche eher schriftlich und wo Chat eigene syntaktische Eigenschaften ausbildet.<sup>16</sup>

Kommentiert [MR1]: In FN fehlt noch der Package-Link

### 3.1 Nichtkanonische Zusammenschreibungen

Chatbeiträge unterscheiden sich in mehreren Punkten auffällig von anderen medial geschriebenen Varietäten. Einer davon ist die häufige Zusammenschreibung von Einheiten, die in ‚Standardsprache‘ getrennt geschrieben würden. In Anhang A findet sich eine Tabelle mit allen nicht-kanonischen Zusammenschreibungen in NoSta-D-unicum. Diese lassen sich in zwei große Gruppen unterscheiden:

1. Zusammenschreibungen, die phonetisch motivierbar sind
2. Zusammenschreibungen, die nicht phonetisch motivierbar sind.

Zusammenschreibungen in der zweiten Gruppe sind oft sogenannte Inflektive oder Bestandteile von Asterisk-Ausdrücke (wobei auch Inflektive oft, aber nicht immer, als Asterisk-Ausdruck geschrieben werden, wie in (15)). Diese kommen eigentlich nur in internetbasierter Kommunikation vor. Es stellt sich die Frage, ob es sich überhaupt um Klitisierung im Sinne der Verschmelzung syntaktisch selbstständiger Formen oder um morphologische Inkorporierung syntaktisch unselbstständiger Formen handelt (vgl. für eine Analyse der internen Syntax von Inflektiven Schlobinski 2001). Die Zusammenschreibung in (16) ist zwar wie viele Inflektivkonstruktionen asterisk-markiert und subjektlos, doch statt der zu erwartenden inflektiven Verbform *sei* steht hier die flektierte Verbform *bin*. Auch hier stellt sich die Frage, welchen syntaktischen Status der Ausdruck haben kann.

(15)

Originaltext:	–	*	brust/	schwell	*	–	–
Normalisierung:	Die	–	Brust	schwillt	–	mir	.
Quelle:	NoSta-D-unicum, Segment 66						

(16)

Originaltext:	*	fett/	bin	*	–	–
Normalisierung:	–	fett	bin	–	ich	.
Quelle:	NoSta-D-unicum, Segment 293					

Für uns ist unklar, ob Inflektive (außer beim Zitieren) jemals ausgesprochen werden. Die spontane Artikulation von Inflektiven und Inflektivkonstruktionen wäre aber eine nötige Evidenz, um über syntaktische oder morphologischen Status der Inflektivkonstruktionen und damit auch über die syntaktische oder morphologische Motivierung ihrer Zusammenschreibung zu entscheiden. Viel allgemeiner aber noch wäre die spontane Artikulation von Inflektiven und Inflektivkonstruktionen die Grundvoraussetzung dafür, diesen konzeptionell mündlichen Status zuweisen zu können, wie es Schlobinski (2001:208) tut. Was nicht

<sup>15</sup> Inflektive weisen zwar eine gewisse Kombinatorik auf, allerdings ist der syntaktische Status dieser „Inflektivkonstruktionen“ (Terminus von Schlobinski 2001) noch zu klären. Möglicherweise handelt es sich bei ihnen um morphologische Inkorporationen. Und selbst wenn man Inflektivkonstruktionen „holophrasalen“ (ebd.) Status zuweist, müsste man immerhin noch erklären, warum sie (außer als direkte Rede und @-Beitrag) nicht eingebettet werden (können?), was mit finiten und auch infiniten Verben bzw. Sätzen gemacht werden kann und wird.

<sup>16</sup> Alle Daten und Skripte für die Auszählungen sind in einem Paper Package im Mind Repository [XXXURLXXX](#) hinterlegt.



ausgesprochen werden kann, kann nicht konzeptionell mündlich sein. Einzelfälle von Inflektivartikulationen reichen aber nicht aus, denn was im Mündlichen seltener vorkommt als im Schriftlichen, kann auch nicht konzeptionell mündlich sein.

Neben den Inflektiven, die für internetbasierte Kommunikation charakteristisch sind, findet man viele Fälle von nichtkanonischer, nicht phonetisch motivierter Zusammenschreibung, die vielleicht „Fehler“ sein können (wie *sichplauze*, *tombefreit*, *einemkuchen*, *[auf] jedenfall*). Ob diese einfache Fehler sind oder ob sich hier neue Regularitäten etablieren, können wir bisher aufgrund der geringen Datenmenge nicht sagen.

Die nichtkanonischen Zusammenschreibungen, die wir phonetisch motivierbar genannt haben, sind für die Fragen nach Mündlichkeit und linguistischen Annahmen der Sprecher interessant. Wenn wir davon ausgehen, dass die Sprecher mit solchen Schreibungen Mündlichkeit nachahmen wollen, zeigen diese Schreibungen, welche phonologischen Wörter oder Klitisierungen für sie salient sind (siehe ein Beispiel in (17a) und eine Liste weiterer solcher Schreibungen in (17b)).

(17a)

Originaltext:	–	so/	n	bierbauch	–	–	?
Normalisierung:	–	So	ein	Bierbauch	ist	das	?
Quelle:		unicum, Segment 286					

(17b) *sparste, stimmts, weils, heisstes, kommste, issen, bisse, isses, hatter*

Die Zusammenschreibungen (17b) betreffen Elemente, die in der linken Satzklammer stehen können (Verben oder Subjunktionen), und klitisierte unbetonte Pronomina. Man wird abwarten müssen, ob solche Fälle zu syntaktischem Wandel in der linken Satzperipherie führen.

### 3.2 Syntaktische Komplexität

Nachdem wir uns in Abschnitt 3.1 mit einem zwar graphisch auffälligen, aber nur bedingt syntaktischen Merkmal von Chat-Sprache beschäftigt haben, wenden wir uns in diesem und dem folgenden Abschnitt genuin syntaktischen Merkmalen zu. Zunächst werden die Länge von Chat-Beiträgen und die Tiefe von Satzeinbettungen in Chat-Beiträgen ermittelt, denn diese geben Aufschluss über das „syntaktische Potenzial“. Je länger Sätze sind und je mehr Einbettungen sie haben, desto mehr komplexe Strukturen sind in ihnen zu erwarten.

Dass Chatbeiträge kürzer sind als Zeitungssätze oder Sätze in literarischer Prosa dürfte wohl zu erwarten sein. Ob aber Chatbeiträge kürzer sind als Äußerungen in der gesprochenen Sprache, ist intuitiv weniger offensichtlich. Eine Auszählung der Segmentlängen (Tabelle 1) zeigt, dass die Segmente in NoSta-D-unicum im Vergleich zu den Segmenten in NoSta-D-bematac und NoSta-D-tuebadz signifikant kürzer sind (ca. 4 Tokens pro Segment)<sup>17</sup>.

Subkorpus	Segmente	Originaltext-Dependenten	Segmentlänge (in Originaltext-Dependenten)		
			$\bar{x}_{arithm}$	$\sigma$	$\bar{x}_{med}$
NoSta-D-unicum	787	3182	4,04	3,10	3
NoSta-D-bematac	1791	7996	4,46	4,65	3
NoSta-D-tuebadz	295	4245	14,39	9,39	13

Tabelle 1: Segmentlängen im NoSta-D-Korpus. Der Längenunterschied von lediglich 0,42 Tokens pro Segment zwischen NoSta-D-unicum und NoSta-D-bematac ist aufgrund der Datenmenge (787 vs. 1791 Segmente) signifikant ( $p_{binom}=0,007$ ).

<sup>17</sup>Man könnte einwenden, dass in den ca. 4,5-Token-langen Segmenten aus NoSta-D-bematac viele Selbstkorrekturen – also ‚eigentlich nicht gemeinte‘ und durch ‚eigentlich gemeinte‘ überschriebene Tokens – enthalten sind, in den ca. 4-Token-langen Chatbeiträgen aber nicht (s. 3.3.1). Reduziert man die faktisch ausgesprochenen Äußerungen aus NoSta-D-bematac auf die ‚eigentlich gemeinten‘, dürfte der Längenvergleich zwischen NoSta-D-unicum und NoSta-D-bematac wohl anders ausfallen. Allerdings gehen wir davon aus, dass einige Unflüssigkeiten wie gefüllte Pausen und einige Selbstkorrekturen integraler Bestandteil von gesprochener Sprache sind und deswegen beim Segmentlängenvergleich unbedingt mit zu berücksichtigen sind (Eklund 2004, Belz 2013). Außerdem müsste man, wenn man schon nur die ‚eigentlich gemeinte‘ Segmentlänge berücksichtigen wollte, konsequenterweise auch alle nichtrealisierten, aber ‚mitgemeinten‘ Tokens (Ellipsen und andere Auslassungen) ergänzen, wodurch die Segmentlänge doch wieder länger würde.

In kürzeren Äußerungen – könnte man meinen – gibt es weniger komplexe syntaktische Strukturen, zum Beispiel weniger Satzeinbettungen. In NoSta-D-unicum und in NoSta-D-bematac sollte man also weniger und weniger tiefe Satzeinbettungen erwarten, als in NoSta-D-tuebadz. In NoSta-D-unicum sollte man zudem auch weniger und weniger tiefe Satzeinbettungen erwarten als in NoSta-D-bematac, weil die Segmente kürzer sind.

Eine Auszählung der Satzeinbettungen zeigt, dass in NoSta-D-unicum und in NoSta-D-bematac Satzeinbettungen tatsächlich signifikant seltener und flacher sind als in NoSta-D-tuebadz. Jedoch zeigt die Auszählung nicht, dass in NoSta-D-unicum Satzeinbettungen signifikant seltener oder flacher sind als in NoSta-D-bematac. Im Gegenteil, die arithmetischen Mittelwerte in Tabelle 2 legen eine umgekehrte Tendenz nahe: in NoSta-D-unicum gibt es trotz kürzerer Beiträge mehr und tiefere Satzeinbettungen als in NoSta-D-bematac.<sup>18</sup>

Subkorpus	Segmente	Tiefe der Satzeinbettung			
		1	2	3	4
NoSta-D-unicum	787	46 (=0,06 pro Segment)	2 (=0,003 pro Segment)	0	0
NoSta-D-bematac	1791	88 (=0,05 pro Segment)	4 (=0,002 pro Segment)	0	0
NoSta-D-tuebadz	295	93 (=0,32 pro Segment)	15 (=0,05 pro Segment)	5	1

Tabelle 2: Satzeinbettungen in NoSta-D.

Nach den Merkmalen Segmentlänge und Einbettungstiefe ergibt sich eine klare Trennung zwischen NoSta-D-unicum und NoSta-D-bematac auf der einen Seite und NoSta-D-tuebadz auf der anderen. Daraus könnte man schließen, dass sich Chat-Sprache und gesprochene Sprache sehr ähneln. Betrachtet man aber ein weiteres syntaktisches Phänomen – die Parenthese –, ergibt sich ein anderes Bild. Parenthesen sind – ganz allgemein – Syntagmen, die linear in einem anderen Syntagma vorkommen, aber zu diesem in keiner syntaktischen Relation stehen. Parenthesen sind prinzipiell überall möglich, ihr Vorkommen zeugt nicht von syntaktischer Komplexität, sondern von Einfachheit. Tabelle 3 zeigt, dass es sehr viele Parenthesen in NoSta-D-bematac, aber kaum welche in NoSta-D-unicum gibt. Die Syntax in NoSta-D-unicum ist also integrierter, die Syntagmen seltener durch Uneingebettetes unterbrochen.

Subkorpus	PAR-Dependenz		Dependenzen insgesamt
	Ja	Nein	
NoSta-D-unicum	2	3180	3182
NoSta-D-bematac	21	7975	7996
NoSta-D-tuebadz	12	4233	4245

Tabelle 3: Parenthesen im NoSta-D-Korpus.

### 3.3 Syntaktische Wohlgeformtheit

Nachdem wir in Abschnitt 3.1 mit den nichtkanonischen Zusammenschreibungen ein Phänomen von Chat-Sprache besprochen hatten, das unmittelbar am Originaltext – auch ohne Vergleich mit anderen Varietäten oder mit einer Normalisierung – zugänglich und als Chat-Spezifikum erkennbar ist, haben wir in Abschnitt 3.2 Eigenschaften von Chat-Sprache besprochen, die erst im Vergleich mit anderen Varietäten zugänglich und als chat-spezifische Verteilungen von auch in anderen Varietäten vorkommenden syntaktischen Phänomenen erkennbar werden. Ein Vergleich mit einer Normalisierung des Originaltexts war auch dafür noch nicht notwendig. In Abschnitt 3.3 wollen wir nun zwei syntaktische Phänomene besprechen, die prinzipiell nur durch einen Vergleich mit einer Normalisierung zugänglich werden und deren chat-spezifische

<sup>18</sup> Ähnliche Ergebnisse für das Englische werden auf der Internetseite der University of Texas berichtet, nur dass der Längenunterschied zwischen Sätzen der Chatsprache und der gesprochenen Sprachen dort nicht formuliert wird und für die Komplexität der Chatsprache kein Komplexitätsmaß angegeben wird:  
<http://coerll.utexas.edu/methods/modules/writing/01/cmc.php>

Verteilung nur im Vergleich mit anderen Varietäten erkennbar wird. Phänomene also, deren Erforschung erst dank des NoSta-D-Korpus möglich geworden ist.

### 3.3.1 Selbstkorrekturen und Wiederholungen

Die Annahme, Chat-Syntax sei wie gesprochene Syntax, impliziert auch, dass Chat-Syntax ebenso wie gesprochene Syntax einen ausgeprägten „on line“-Charakter haben sollte (vgl. Auer 2000). Gesprochene Sprache entsteht spontan und enthält viele Selbstkorrekturen und Wiederholungen, in denen ein Sprecher einmal geäußerte Wörter oder Syntagmen durch andere ersetzt bzw. in denen ein Sprecher Silben, Wörter und Syntagmen wiederholt, um eine bereits angedeutete Selbstkorrektur zurückzunehmen oder das Parsing einer komplexen Struktur zu erleichtern (vgl. z. B. Belz 2013). An der Spontaneität der Chat-Beiträge ist nicht zu zweifeln. Sie entstehen in schneller Abfolge. Mitunter kann die am Absendezeitpunkt festgemachte chronologische Abfolge der Beiträge sogar ihre inhaltliche Abfolge überholen: Chat-Teilnehmer setzen ihre Beitragssequenzen bereits fort, während andere noch Antworten auf frühere Beiträge tippen (vgl. Storrer 2001:3f.). Dieses Verhalten führt dazu, dass ganze Chat-Beiträge wiederholt und gegebenenfalls korrigiert werden müssen, wie Beispiel (18) zeigt.

(18)

Segment	Chat-Teilnehmer	Chat-Beitrag
128	Quaki	„und wo is ein apfel für mich??“
...		
146	Zora	„zora bewirbt sich damit mal um stipendien an den unis“
...		
154	Emon	„zora bewirbt sich womit um stipendien?“
155	Emon	„nen apfel?“
...		
157	Zora	„ne mit meinem zeugniss, ich bin ja nich doof!“

Eine genaue Unterscheidung von Selbstkorrekturen und Wiederholungen und eine dadurch motivierte unterschiedliche Dependenzannotation derselben wurde im NoSta-D-Korpus nicht vorgenommen. Es wäre eine äußerst diffizile und im Grunde auch keine syntaktische Unterscheidung, sondern eine pragmatische. Das syntaktisch Gemeinsame von Selbstkorrekturen und Wiederholungen, womit die einheitliche Dependenzannotation derselben in NoSta-D motiviert wird, ist der durch sie bedingte Überschuss an syntaktischer Struktur: Selbstkorrekturen sind ein „Zuviel“ an Struktur, Wiederholungen ebenfalls. Ein Hörer bzw. Leser muss erkennen, dass nicht alle Teile einer Äußerung tatsächlich gemeint sind, sondern dass er bestimmte vorangehende, korrigierte bzw. Erstvorkommen von Wörtern/Syntagmen von der Interpretation ausschließen und nur nachfolgende, korrigierende bzw. wiederholte Vorkommen in die Interpretation einschließen soll. Darin, dass sie nur partiell (in ihrer jeweils letzten Instanziierung) zu interpretieren sind, unterscheiden sich Selbstkorrekturen und Wiederholungen fundamental von anderen Reihungen wie Koordination (*dann rechts, dann links, dann gradeaus*) und Framing (*hier, oben, bei dem Bild, so links davon, genau da*), die holistisch zu interpretieren sind.

Eine Auszählung der Selbstkorrekturen und Wiederholungen (Tabelle 4) zeigt, dass sich Chat-Teilnehmer innerhalb von Posts nicht selbst korrigieren und wiederholen, was aber durchaus typisch für gesprochene Sprache ist, vgl. Beispiel (18) aus NoSta-D-bematac.

Subkorpus	COR-Dependenz		Dependenzen insgesamt
	Ja	Nein	
NoSta-D-unicum	0	3182	3182
NoSta-D-bematac	139	7857	7996
NoSta-D-tuebadz	0	4245	4245

Tabelle 4: COR-kreuzgelabelte Abhängigkeiten in NoSta-D. Ein COR-Label wird vergeben für korrigierte Abhängigkeiten oder Erstvorkommen von wiederholten Elementen.<sup>19</sup>

(18)

Originaltext:	<i>hinter dem Toaster muss rechts Entschuldigung links abgebogen werden</i>
Quelle:	NoSta-D-bematac_2012-10-31-A, Segment 25

Die Sprache in NoSta-D-unicum erscheint auf der Ebene der Posts also „editiert“, insofern dass die Chat-Teilnehmer syntaktisch wohlgeformte(re) Posts abschicken, ohne Selbstkorrekturen und Wiederholungen. Auf Diskurs-Ebene erscheint die Sprache in NoSta-D-unicum allerdings nicht editiert, insofern dass die entstehenden Texte nicht kohärent sind. Die Beitragsabfolge weicht von der inhaltlichen Abfolge ab, ohne dass dafür die sonst in der Schriftsprache vorhandenen Kohärenzsichernden Mittel genutzt werden. Erstere Eigenschaft markiert Chat-Sprache als „nicht mündlich“, die zweite Eigenschaft hingegen als „mündlich“, denn auch in mündlichen Gruppengesprächen wird gleichzeitig gesprochen, gegenseitig unterbrochen, spontan das Thema gewechselt, zum vorherigen Thema zurückgekehrt usw.

### 3.3.2 Fragmente

Während uns im Abschnitt 3.3.1 die syntaktische Wohlgeformtheit im Sinne der Abwesenheit überschüssiger syntaktischer Struktur interessierte („nicht zu viel Struktur“), kommen wir in diesem Abschnitt nun zur syntaktischen Wohlgeformtheit im komplementären Sinne, also im Sinne der Abwesenheit fehlender syntaktischer Struktur („nicht zu wenig Struktur“).

Fehlende syntaktische Struktur ist in der Linguistik unter Bezeichnungen wie *Ellipse*, *Auslassung*, *Abbrüche* bekannt. Es handelt sich um oberflächlich nicht realisierte, aber mitgemeinte und potentiell realisierbare Elemente – ganz im Gegensatz zu den realisierten, aber nicht mitgemeinten Elementen in Selbstkorrekturen und Wiederholungen.<sup>20</sup> Was genau nicht realisiert, aber mitgemeint ist, wird aus dem sprachlichen oder außersprachlichen Kontext hergeleitet. Zahlreiche Studien beschäftigen sich damit, fehlende syntaktische Strukturen nach Art und Umfang des jeweils zur Identifizierung heranzuziehenden Kontexts zu klassifizieren (für einen Überblick siehe z.B. Reich 2011). Wir kennen allerdings keine Studien, die sich spezifisch mit dem sprachlichen Material beschäftigen, das in elliptischen Strukturen realisiert wurde. Wir wählen hierfür die Bezeichnung „Fragmente“.

Im NoSta-D-Korpus können wir die objektsprachlichen Fragmente und die nicht objektsprachlichen Ellipsen/Auslassungen – diese werden ja nur vom Hörer/Leser bzw. Linguisten hinzugedacht – sauber trennen: Fragmente stehen unverändert im Originaltext und fehlende syntaktische Strukturen werden in der Normalisierung aufgefüllt. Ein Vergleich des Originaltexts mit der Normalisierung ermöglicht nun auch eine qualitative Auswertung der Fragmente (neben der weiterhin möglichen Auswertung der Ellipsen, Auslassungen etc.).

Eine Auszählung der Fragmente im NoSta-D-Korpus (Tabelle 5) zeigt, dass sich die Fragmente in NoSta-D-unicum qualitativ von den Fragmenten in NoSta-D-bematac unterscheiden. In (dependenzfähigen) Token gezählt gibt es in beiden Subkorpora ungefähr gleich viel fragmentarisches Material, denn in beiden Subkorpora mussten gleich viele (dependenzfähige) Token hinzunormalisiert werden, um kanonische Texte zu erhalten. Das fragmentarische Material in NoSta-D-unicum verteilt sich aber im Gegensatz zu NoSta-D-bematac auf eine kleinere Anzahl von Fragmenten, die intern parallel zur Normalisierung, aber nach außen hin abweichend von der Normalisierung annotiert werden. Zudem befinden sich die Fragmente in NoSta-D-

<sup>19</sup> Ob eine Selbstkorrektur oder eine asyndetische Koordination vorliegt, ist (ohne Kontext) nicht immer entscheidbar, denn beides sind Reihungen gleichwertiger (aber nicht identischer) Elemente. In der NoSta-D-Normalisierung wird das Problem durch Betrachtung des Kontexts entschieden und die Entscheidung durch Nichteinfügen bzw. Einfügen einer Konjunktion fixiert. Bei Reihungen ungleichwertiger (und auch nicht identischer) Elemente werden keine Konjunktionen o. Ä. hinzunormalisiert, es gilt: nur was nicht als Framing interpretiert werden kann, wird als Selbstkorrektur annotiert. (Wiederholungen, also Reihungen gleichwertiger und identischer Elemente, sind nicht mit Koordinationen oder Framings verwechselbar. Vereinzelt treten auch (idiomatisierte) Reihungen gleichwertiger und (scheinbar) identischer Elemente auf, diese werden den Koordinationen subsumiert, z.B. *Er ist durch und durch Lehrer*.)

<sup>20</sup> Als „Null-Elemente“ werden demgegenüber eher Elemente verstanden, die nicht realisiert und potentiell auch gar nicht realisierbar, aber dennoch mitgemeint sind, z.B. Null-Subjekte in Infinitivsätzen mit *zu*.

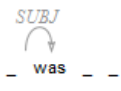
unicum häufiger unten im Dependenzgraphen und fehlenden syntaktischen Strukturen oben – mit der Konsequenz, dass die Fragmente selbst an die Segmentwurzel gebunden werden – während die Fragmente in NoSta-D-bematac häufiger diffus im Dependenzgraphen verteilt und mit fehlenden syntaktischen Strukturen verzahnt sind.<sup>21</sup>

Subkorpus	Originaltext- dependenten	hinzunormalisierte Dependenten (≈aufgefüllte Ellipsen/Auslassungen)	von Normalisierung abweichend angebundene Originaltextdependenten (≈Regenten (!) von Fragmenten)		
			gesamt	an Segmentwurzel angebunden (Fragment unterhalb von Ellipse/Auslassung)	X- Dependenz (Fragment um Ellipse/Auslassung drumherum)
NoSta-D- unicum	3182	461 (=14%)	248	202	16
NoSta-D- bematac	7996	1142 (=14%)	1336	699	178
NoSta-D- tuebadz	4245	122 (=3%)	202	60	27

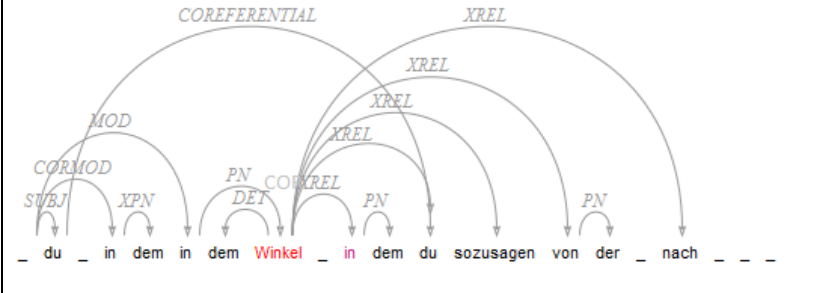
Tabelle 5: Fragmente in NoSta-D.

Beispiel (19) zeigt ein für NoSta-D-unicum typischeres Fragment. Beispiel (20) zeigt ein für NoSta-D-bematac typischeres Fragment.

(19)

ORIG:	
Quelle:	NoSta-D-unicum, Segment 3

(20)

ORIG:	
Quelle:	NoSta-D-bematac_2011-12-14-B, Segment 178

Fragmente in NoSta-D-unicum sind innerhalb eines Segments eher dependentiell gleichwertig – sie sind gleichermaßen Dependenten von nichtrealisierten Regenten, nicht Dependenten voneinander. Die Fragmente in NoSta-D-bematac hingegen sind innerhalb eines Segments eher nicht dependentiell gleichwertig – die einen Fragmente sind (mittelbare) Dependenten der anderen Fragmente.

<sup>21</sup> Einer gesonderten Betrachtung bedürfen Fragmente unterhalb von Konjunktionen und Verba dicendi, auf welche hier aus Platzgründen verzichtet wird. Das hier Dargestellte bleibt davon jedoch unbeeinträchtigt.

Auch hier sehen wir wieder, dass sich NoSta-D-unicum und NoSta-D-bematac, die beide auf den ersten Blick ähnlich scheinen, sich im Detail in interessanter Weise unterscheiden. Die Fragmente in NoSta-D-unicum sind zwar häufig kein Sätze, aber immerhin wohlgeformte Konstituenten. Die Fragmente in NoSta-D-bematac dagegen sind oft weder Sätze noch wohlgeformte Konstituenten.

#### 4. Zusammenfassung

In diesem Beitrag haben wir einen Weg aufgezeigt, wie man syntaktische Strukturen in Plauderchat-Daten beschreiben und mit analogen Strukturen in anderen Varietäten qualitativ und quantitativ vergleichen kann. Wir haben anhand des NoSta-D-Korpus gezeigt, dass eine konsistente, gut beschriebene Normalisierung Vergleiche auf mehreren sprachlichen Ebenen zulässt. Wir analysieren und vergleichen einen Plauderchat mit gesprochenen MapTask-Daten und Zeitungsdaten.

Bisher werden Chats oft als eine eigene Varietät untersucht und ihre Eigenschaften werden nur cursorisch denen anderer Varietäten verglichen. Weil Chats interaktional ablaufen, wird oft angenommen, dass sie auch sprachlich irgendwie ‚mündlich‘ seien. Als Evidenz dafür wird angeführt, dass es nichtkanonische Zusammenschreibungen gibt und die Beiträge eher kurz und nicht komplex seien. Chats werden nicht detailliert mit Gesprächen oder auch mit anderen schriftlichen Daten verglichen. In unserem Beitrag haben wir gezeigt, dass Chatdaten auf den ersten Blick tatsächlich aussehen wie Gesprächsdaten und dass sich beide Varietäten von konzeptionell geschriebenen Daten unterscheiden. Bei genauerem Hinsehen ist das Bild komplexer. Nur einige der nichtkanonischen Zusammenschreibungen ahmen phonetische Klitisierungen nach. Andere folgen ganz eigenen Regeln. Chatbeiträge sind kürzer als Beiträge in Gesprächen, aber syntaktisch etwas integrierter. Im Unterschied zu Gesprächen finden wir in Chatdaten kaum Parenthesen. Chatdaten und gesprochene Daten enthalten viele Fragmente, im Detail unterscheiden sich diese aber: Chatfragmente sind meist kanonische Phrasen, während gesprochene Fragmente oft nicht einmal Phrasen bilden.

Unsere Daten stammen aus einem bestimmten Plauderchat. Plauderchats (egal, wie man sie definiert) unterscheiden sich erheblich voneinander. Es wäre daher nötig, unsere Befunde anhand von anderen Daten (Plauderchats aus anderen Bereichen oder mit anderen Teilnehmern aus anderen sozio-ökonomischen Schichten) zu replizieren, bevor man Aussagen über ‚Sprache in Chats‘ machen kann. Ohne eine gute Datengrundlage kann man keine Untersuchungen zu den möglichen Einflüssen wie Medium, Gleichzeitigkeit, Anonymität etc. auf syntaktische Strukturen machen.

#### 5. Quellen

##### Forschungsliteratur

Auer, Peter (2000): On line-Syntax – oder: Was es bedeuten könnte, die Zeitlichkeit der mündlichen Sprache ernst zu nehmen. In: Sprache und Literatur 85, 43-56.

Albert, Stefanie, Jan Anderssen, Regine Bader, Stephanie Becker, Tobias Bracht, Thorsten Brants, Vera Demberg, Stefanie Dipper, Peter Eisenberg, Silvia Hansen, Hagen Hirschmann, Juliane Janitzek, Carolin Kirstein, Robert Langner, Lukas Michelbacher, Oliver Plaehn, Cordula Preis, Marcus Pußel, Marco Rower, Bettina Schrader, Anne Schwartz, George Smith und Hans Uszkoreit (2003): TIGER Annotationsschema. Online unter: <https://files.ifi.uzh.ch/cl/siclemat/lehre/papers/tiger-annot.pdf> (15.11.2014).

Bartz, Thomas, Michael Beißwenger und Angelika Storrer (2014): Optimierung des Stuttgart-Tübingen-Tagset für die linguistische Annotation von Korpora zur internetbasierten Kommunikation. Phänomene, Herausforderungen, Erweiterungsvorschläge. In: Journal for Language Technology and Computational Linguistics 28 (1), 157–198.

Behr, Irmtraud und Hervé Quintin (1996): Verblöde Sätze im Deutschen. Zur syntaktischen und semantischen Einbindung verblöder Konstruktionen in Textstrukturen. Tübingen: Stauffenburg.

Beißwenger, Michael (2013): Das Dortmunder Chat-Korpus. In: Zeitschrift für germanistische Linguistik 41(1), 161–164.

Eingereicht | 26.11.2014 | Kommentare willkommen: [burkhard.dietterle@hu-berlin.de](mailto:burkhard.dietterle@hu-berlin.de)

Beißwenger, Michael, Maria Ermakova, Alexander Geyken, Lothar Lemnitzer und Angelika Storrer (2012): A TEI Schema for the Representation of Computer-mediated Communication. In: Journal of the Text Encoding Initiative 3. Online unter: <http://jtei.revues.org/pdf/476> (15.11.2014).

Belz, Malte (2013) Disfluencies und Reparaturen bei Muttersprachlern und Lernern - eine kontrastive Analyse. Masterarbeit. Humboldt-Universität zu Berlin. Online unter: <https://edoc.hu-berlin.de/docviews/abstract.php?lang=ger&id=40482> (25.11.2014).

Bley-Vroman, Robert (1983): The comparative fallacy in interlanguage studies: The case of systematicity. In: Language Learning 33(1), 1-17.

Bohnet, Bernd (2010): Top Accuracy and Fast Dependency Parsing is not a Contradiction. In: Huang, Churen und Dan Jurafsky (Hrsg.): Proceedings of the 23rd International Conference on Computational Linguistics. Peking: Tsinghua University Press, 89-97.

Dipper, Stefanie, Anke Lüdeling und Marc Reznicek (2013): NoSta-D: A Corpus of German Non-Standard Varieties. In: Zampieri, Marcos und Sascha Diwersy (Hrsg.): Non-Standard Data Sources in Corpus-Based Research 5. Köln: Shaker (ZSM Studien), 69-76.

Duden 1 (2013): Die deutsche Rechtschreibung. 26. Aufl., Mannheim: Dudenverlag.

Duden 4 (2005): Die Grammatik. Mannheim: Dudenverlag.

Eisenberg, Peter (2007): Sprachliches Wissen im Wörterbuch der Zweifelsfälle. Über die Rekonstruktion einer Gebrauchsnorm. In: Zeitschrift für Sprachkritik und Sprachkultur 3, 209-228.

Eklund, Robert (2004): Disfluency in Swedish Human-human and Human-machine Travel Booking Dialogues. PhD thesis, Linköping Studies in Science and Technology, Dissertation No. 882, Department of Computer and Information Science, Linköping University, Sweden.

Forst, Martin, Nuria Bertomeu, Berthold Crysmann, Frederik Fouvry, Silvia Hansen-Schirra und Valia Kordoni (2004): Towards a Dependency-based Gold Standard for German Parsers — The TiGer Dependency Bank. In: COLING Workshop on Linguistically Interpreted Corpora. Online unter: <http://www.coli.uni-saarland.de/conf/linc-04/forst.pdf> (15.11.2014).

Foth, Kilian A. (2006): Eine umfassende Constraint-Dependenz-Grammatik des Deutschen. Technischer Report. Hamburg: Universität Hamburg. Online unter: [http://edoc.sub.uni-hamburg.de/informatik/volltexte/2014/204/pdf/foth\\_eine\\_umfassende\\_.pdf](http://edoc.sub.uni-hamburg.de/informatik/volltexte/2014/204/pdf/foth_eine_umfassende_.pdf) (15.11.2014).

Frank, Anette (2013): A Tour of Grammar Formalisms. In: Holloway King, Tracy und Valeria de Paiva (Hrsg.): From Quirky Case to Representing Space. Papers in Honor of Annie Zaenen. Stanford: CSLI Publications, 75-93.

Hennig, Mathilde (2006): Grammatik der gesprochenen Sprache in Theorie und Praxis. Kassel: Kassel University Press.

Herring, Susan C., Dieter Stein und Tuija Virtanen (Hrsg.) (2013): Handbook of pragmatics of computer-mediated communication. Berlin: Mouton de Gruyter.

Hirschmann, Hagen, Seanna Doolittle und Anke Lüdeling (2007): Syntactic Annotation of Non-canonical Linguistic Structures. In: Proceedings of Corpus Linguistics 2007, Birmingham. Online unter: [http://ucrel.lancs.ac.uk/publications/CL2007/paper/128\\_Paper.pdf](http://ucrel.lancs.ac.uk/publications/CL2007/paper/128_Paper.pdf) (15.11.2014).

Kübler, Sandra und Jelena Prokic (2006): Why is German Dependency Parsing More Reliable than Constituent Parsing? Proceedings of the Fifth Workshop on Treebanks and Linguistic Theories. Prague, Czech Republic, 7-18. Online unter: <http://cl.indiana.edu/~skuebler/papers/german.pdf> (15.11.2014).

Lüdeling, Anke (2011): Corpora in Linguistics. Sampling and Annotation. In: Grandin, Karl (Hrsg.) Going Digital. Evolutionary and Revolutionary Aspects of Digitization. Nobel Symposium 147. New York: Science History Publications, 220-243.

Eingereicht | 26.11.2014 | Kommentare willkommen: [burkhard.dietterle@hu-berlin.de](mailto:burkhard.dietterle@hu-berlin.de)

Maas, Utz (2010): Literat und orat. Grundbegriffe der Analyse gesprochener und geschriebener Sprache. Graz: Grazer linguistische Studien 73, 21–150.

Myslin, Mark und Stefan Th. Gries (2010): k dixez? A corpus study of Spanish Internet orthography. In: Literary and Linguistic Computing 25(1), 85–104.

Nivre, Joakim, Jens Nilsson, Johan Hall, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov und Erwin Marsi (2007): MaltParser. A Language-Independent System for Data-Driven Dependency Parsing. In: Natural Language Engineering 13 (1), 1–41.

Rafferty, Anna and Christopher D. Manning (2008): Parsing Three German Treebanks. Lexicalized and Unlexicalized Baselines. In: Proceedings of the Workshop on Parsing German. Stroudsburg: Association for Computational Linguistics, 40–46. Online unter: <http://dl.acm.org/citation.cfm?id=1621401.1621407> (15.11.2014).

Reich, Ingo (2011): Ellipsis In: Maienborn, Claudia, Klaus von Heusinger und Paul Portner (Hrsg.): Semantics: An International Handbook of Natural Language Meaning. Berlin/New York: de Gruyter, 1849–1874.

Reznicek, Marc, Anke Lüdeling und Hagen Hirschmann (2013): Competing Target Hypotheses in the Falko Corpus. A Flexible Multi-Layer Corpus Architecture. In: Díaz-Negrillo, Ana, Nicolas Ballier und Paul Thompson (Hrsg.): Automatic Treatment and Analysis of Learner Corpus Data. Amsterdam: John Benjamins, 101–124.

Sauer, Simon und Anke Lüdeling (erscheint): Flexible Multi-Layer Spoken Dialogue Corpora. In International Journal of Corpus Linguistics.

Schalowski, Sören (2009): Über Topik-Drop im Deutschen. Untersuchung zum Einfluss der grammatischen Funktion und des Merkmals 'Person'. Magisterarbeit, Humboldt-Universität zu Berlin.

Schiller, Anne, Simone Teufel, Christine Stöckert und Christine Thielen (1999): Guidelines für das Tagging deutscher Textkorpora mit STTS. Technischer Bericht. Stuttgart/Tübingen: Universität Stuttgart/ Universität Tübingen. Online unter: <http://www.sfs.uni-tuebingen.de/resources/stts-1999.pdf> ( 15.11.2014).

Schlobinski, Peter (2001): \*knuddel zurueckknuddel dich ganzdolkknuddel\*. Inflektive und Inflektivkonstruktionen im Deutschen. In: Zeitschrift für Germanistische Linguistik 29(2), 192–218.

Seeker, Wolfgang und Jonas Kuhn (2012): Making Ellipses Explicit in Dependency Conversion for a German Treebank. In: Proceedings of the 8th International Conference on Language Resources and Evaluation. Istanbul, Turkey: ELRA, 3132–3139.

Storrer, Angelika (2001) Getippte Gespräche oder dialogische Texte? Zur kommunikationstheoretischen Einordnung der Chat-Kommunikation. In: Lehr, Andrea; Matthias Kammerer, Klaus-Peter Konerding, Angelika Storrer, Caja Thimm und Werner Wolski (Hrsg.): Sprache im Alltag. Beiträge zu neuen Perspektiven in der Linguistik. Berlin u.a.: de Gruyter, 439–465.

Storrer, Angelika (2013): Sprachstil und Sprachvariation in sozialen Netzwerken. In: Frank-Job, Barbara, Alexander Mehler & Tilmann Sutter (Hrsg.): Die Dynamik sozialer und sprachlicher Netzwerke. Konzepte, Methoden und empirische Untersuchungen an Beispielen des WWW. Wiesbaden: VS Verlag für Sozialwissenschaften, 331–366.

Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler und Heike Zinsmeister (2005): Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z). Online unter: [www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-0911.pdf](http://www.sfs.uni-tuebingen.de/fileadmin/static/ascl/resources/tuebadz-stylebook-0911.pdf) (15.11.2014).

Teuber, Oliver (1998): fasel beschreib erwähn — Der Inflektiv als Wortform des Deutschen. In: Germanistische Linguistik 141/142, 7–26.



Eingereicht | 26.11.2014 | Kommentare willkommen: [burkhard.dietterle@hu-berlin.de](mailto:burkhard.dietterle@hu-berlin.de)

Wiese, Heike, Ulrike Freywald und Katharina Mayr (2009): Kiezdeutsch as a Test Case for the Interaction between Grammar and Information Structure. Potsdam: Universitätsverlag (= Interdisciplinary Studies on Information Structure 12).

#### **Internetquellen:**

##### **NoSta-D Projekt-Webseite:**

Korpus und Vorverarbeitungs- und Annotationsrichtlinien

<https://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/research/nosta-d>

Korpus ist durchsuchbar in ANNIS3

<https://korpling.german.hu-berlin.de/annis3/>

##### **Bematac**

Korpus und Vorverarbeitungs- und Annotationsrichtlinien

[www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/bematacKorpus](http://www.linguistik.hu-berlin.de/institut/professuren/korpuslinguistik/forschung/bematacKorpus)

Korpus ist durchsuchbar in ANNIS3

<https://korpling.german.hu-berlin.de/annis3/>

##### **Dortmunder Chat Korpus**

Korpus und Suche

<http://www.chatkorpus.tu-dortmund.de>

zitierter Plauderchat

[http://www.chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/unicum\\_21-02-2003\\_1.html](http://www.chatkorpus.tu-dortmund.de/files/releasehtml/html-korpus/unicum_21-02-2003_1.html)

## Anhang A

Tokens	In Asterisk-Ausdruck?	Phonetisch motivierbar?
„erleichtert/ guck“	ja	nein
„[wat] für/ n“	nein	ja
„brust/ schwell“	ja	nein
„[eine] Zeit/ lang“	nein	nein, aber nach alter Rechtschreibung erlaubt
„stimmt/ s“	nein	ja
„so/ ne“	nein	ja
„[auf] jeden/ fall“	nein	nein (allerdings feste Phrase)
„Na/ gut/ 50/ cm/ Laufaufleine“	ja	nein
„51/ cm“	ja	nein
„neue/ such“	ja	nein
„k/ A“	nein	nein
„so/ was“	nein	ja (und auch orthographisch möglich)
„[auf] jeden/ fall“	nein	nein
„so/ was“	nein	ja
„weil/ s“	nein	ja
„wür/ s“	nein	ja
„heisst/ es [... @ ...]“	nein	ja
„an/ ne [Stirm]“	nein	ja
„[aber wie] kommst/ e“	nein	ja
„[nennt] sich/ plauze“	nein	nein
„kommst/ e“	nein	ja
„so/ n“	nein	ja
„kopfl kratz“	ja	nein
„fett/ bin“	ja	nein
„korsett/ such“	ja	nein
„so/ was“	nein	ja
„sparst/ e“	nein	ja
„augen/ reib“	ja	nein
„zwickizwacki/ marc30/ quaki“	nein	nein
„auf/ n [baum]“	nein	ja
„hinterher/ kann“		
„So/ n“	nein	ja
„[wat] für/ n“	nein	Ja
„iss/ en“	nein	ja
„kopfl kratz“	ja	nein
„bis/ se“	nein	ja
„[is] tom/ befreit	nein	nein
„so/ was“	nein	ja
„auf/ n [baum]“	nein	ja
„gegen/ den/ chat/ und/ emine/ compi/ tret“	ja	nein
„empört/ guck“	ja	nein
„macht/ s“	nein	ja
„wink/ und/ weg“	nein	nein
„[hab] ich/ s“	nein	ja
„macht/ es“	nein	ja
„einem/ kuchen“	nein	nein
„zunge/ raushäng“	ja	nein
„skeptisch/ zuhör“	ja	nein
„iss/ es“	nein	ja
„Lena/ anschau“	ja	nein
„wink/ und/ wech“	nein	nein
„hatt/ er“	nein	ja
„hab/ s [... gesehen]“	nein	ja

Alle nichtkanonischen Zusammenschreibungen in NoSta-D mit einer Einschätzung über ihre Motiviertheit.

## Anhang B

Syntaktische Darstellung der beiden Satzeinbettungen zweiten Grades in NoSta-D-unicum.

ORIG:	<p>Diagram illustrating the syntactic structure of the sentence: "Die eins Minus hast du bekommen, weil du eine Zeit lang häufig gesagt hast, was ist Benehmen?". The diagram shows hierarchical structure with labels like S, OBJA, COREFERENTIAL, NEB, KONJ, SUBJ, OBJC, DET, MOD, AUX, PREL, and SUBJ.</p>
Quelle:	NoSta-D-unicum, Segment 92

ORIG:	<p>Diagram illustrating the syntactic structure of the sentence: "bochum mit kastanien bewerf, damit es den baum auf dem ich bin in ruhe lässt". The diagram shows hierarchical structure with labels like SINFL, OBJA, MOD, PN, REL, MOD, SUBJ, and PN.</p>
Quelle:	NoSta-D-unicum, Segment 614