Grammatikschreibung im digitalen Zeitalter

SEBASTIAN DRUDE

Frankfurt, 2012-06-20

Habilitationsvortrag

:: Digitale Grammatikschreibung ::

1. Einleitung: Grammatikschreibung
2. Sprachenvielfalt und -dokumentation
3. Digitale Klartextgrammatiken
4. Implementierte Grammatiken
5. Über- und Ausblick: Integration

1. Einleitung: Grammatikschreibung

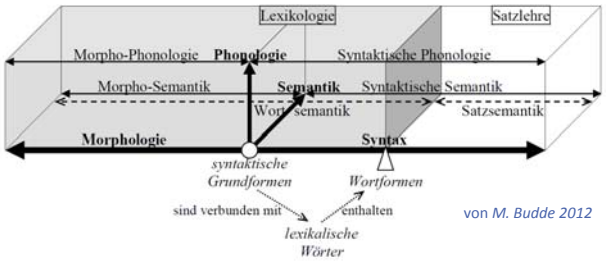
- ***Grammatik***: Beschreibung oder Theorie der Struktur einer Sprache
- Kernbereich seit Beginn der Sprachwissenschaft:
  - PĀṆINI *Aṣṭādhyāyī* (Sanskrit)
  - DIONYSIUS THRAX *Tékhnē grammatiké* (Griechisch)
  - AELIUS DONATUS *Ars Grammatica* (Latein)
  - Europäische u. Missionarsgrammatiken ab 16. Jhd
  - Moderne deskriptive Linguistik seit Boas / Sapir
- Typologie u. Sprachtheorien beruhen auf Gr.

1. Einleitung: Grammatikschreibung

- „Boas'scher Dreiklang":
  Grammatik, Wörterbuch, Textsammlung
- Inhalt:
  - Phonetik / Phonologie (manchmal)
  - Morphologie *(nach Wortarten)* (immer)
  - Syntax *(nach Konstituenten- und Satzarten)* (oft)
  - Semantik (implizit bei Morphologie & Syntax)
  - Pragmatik, Textstruktur, Satzintonation,... (selten)

1. Einleitung: Grammatikschreibung

Disziplinäre Untergliederung der Sprachwissenschaft im Bereich der bedeutungstragenden Ausdrücke



von *M. Budde 2012*

1. Einleitung: Grammatikschreibung

- Etabliertes akademisches Genre
- Typische / spezielle Elemente in Grammatiken:
  - *Exemplare* von objektsprachlichen Ausdrücken
  - Tabellen (Paradigmen...), Strukturgraphen (Baumdiagramme,...), Formale Regeln, …
  - Spezifische Termini (theorieabhängig)
- Jede Gr. ist „formuliert *in terms of*" einer Theorie
- Oft wird in Grammatiken eine (Teil)Theorie sowohl erläutert als auch angewendet

2. Sprachenvielfalt und -dokumentation

Sprachenvielfalt

- Es gibt weltweit 5.000–7.000 Sprachen
- Exakte Zahlen sind schwierig (fehlendes Wissen, Sprache-Dialekt-Problem)
- Ca. 85 Sprachen mit mehr als 10 Mio. Sprechern
- Nur zu wenigen Dutzend Spr. gibt es mehrere umfassende Beschreibungen und Materialien
- Bis zu über 90% der Sprachen sind „bedroht“ (möglicherweise keine Sprecher in einer bis fünf Generationen)

2. Sprachenvielfalt und -dokumentation

Feldforschung

- Die große Mehrzahl der Sprachen kann nur in der Sprachgemeinschaft untersucht werden
- Feldforschung im Schnittfeld der Disziplinen:
  - Ethnolinguistik / *anthropological linguistics*
  - Sprachtheorie
  - Typologie & Universalienforschung
  - Seit ca. 2000: Sprachdokumentation -> digitale Sprachtechnologie

2. Sprachenvielfalt und -dokumentation

Sprachdokumentation

- Neues Arbeitsgebiet (Himmelmann 1998)
- Zusätzlich zu schriftlichen Textsammlungen: **Korpora mit annotierten Multimedia-Daten**
- Fokus auf tatsächlichem Sprachgebrauch
- Korpora für viele Anwendungen (*multi-purpose*)
- Ein Zweck ist die Analyse zur Erstellung einer Grammatik im traditionellen Sinne
- Digitale Daten und Werkzeuge -> neue Möglichkeiten für die Grammatikschreibung

3. Digitale Klartextgrammatiken

a. Hypertextgrammatiken

b. Language Archiving Technology

c. Wiki-Grammatiken

d. Die Text-Encoding Initiative (TEI)

e. Versionierung und Publikation

3. Digitale Klartextgrammatiken

a. Hypertextgrammatiken

- Beschreibungen als digitale Dokumente
- Über Office-Programme und PDFs hinaus:
- Hypertextdokumente mit einzelnen aber miteinander verknüpften (‘verlinkten’) Seiten
- Notwendigkeit besteht schon, wenn Boas’scher Dreiklang als integriertes Werk verstanden wird
- Beispiel: JEFFREY HEATH on *Nunggubuyu*: 1980: *N. Myths and Ethnographic Texts*; 1982: *N. Dictionary*; 1984: *Functional Grammar of Nunggubuyu*

Zitiert nach: **Musgrave und Thieberger** (to appear) “Language description and hyper-text: Nunggubuyu as a case study”, in S. Nordhoff (ed.) *Grammaticography*, LD&C, special issue.

Text:

43.4 wu=wayama-n^{g}i-yaj mari dhan^{g}gid! ajaba ∅=lhi-n^{y},
as it proceeded$_c$ and chop then it chopped it$_p$
2 ana-ra:n^{g}ag, ∅=madhari-n^{y} ∅=madhari-n^{y} ∅=madhari-n^{y}, yin^{g}ga
wood it chopped it$_p$ nearly
wu-ragar=bayama-n^{g}i mari n^{g}ijan^{g} wurugu gulmurg!,
it went along forcefully$_c$ and more later run
4 wini=wilbili-n^{y} arwagarwar-ala-aj,
they (MDu) flew$_p$ around on top
It (devil) came along and began to chop down the tree. It was chopping and chopping. It (tree) was about to crash down, but then they (two) flew away. (They flew) around up high.

Wörterbuch:

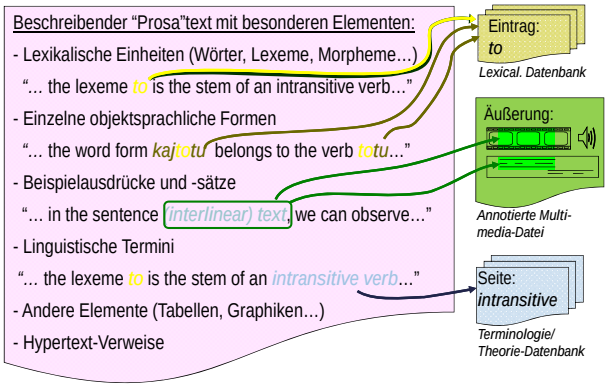
dhan^{g}gid! Rf to chop. 16.14.3, 43.4.1, 43.6.4.
Associated with verb =lha- 'to chop'.

Grammatik:

This particle can combine with other particles. We mentioned /mari wurugu/ and /wurugu n^{g}a/ in the previous section (it is likely that /n^{g}a wurugu/ also occurs). We can cite /n^{g}ijan^{g} wurugu/ (cf. next section) 'again later' or 'more later' 21.9.1, 21.10.1, 33.1.2, 43.4.3 (with preceding /mari/), 43.5.2/4, 52.5.2/3, 163.19.2/3, showing this order to be consistent. There is also an ex. of /wurugu yin^{g}ga/ (cf. §12.7) 'later' (with anticipation nuance) 71.2.4.
Additional exx. of /wurugu/ are 7.6.1/2, 13.13.4, 37.2.4 (if not mistranscribed), 47.12.7, 55.9.2, 69.5.1, 69.7.4/6, 71.18.1, 73.5.5, 106.3.1/2, 116.8.2, 143.10.3, 157.7.2, 161.1.4, 161.3.4, 161.20.2, 161.32.4, 162.7.5, 162.14.1, 163.14.2, 165.1.1.
A competing form (not a particle) is /an-uba-ni:-'la-wala/ 'after that' (§7.8, §7.31).

3. Digitale Klartextgrammatiken

a. Hypertextgrammatiken



3. Digitale Klartextgrammatiken

a. Hypertextgrammatiken

Hauptfunktionalitäten

- *Links* von *Exemplaren* zu Äußerungen in einem Multimedia-Korpus
- Automatische Generierung von Konkordanzen mit weiteren Beispielen
- Links von lexikalischen Einheiten zu einem Online-Lexikon
- Trennung der Beschreibung von der Erläuterung analytischer Begriffe (diese in einer terminologischen / Theoriedatenbank)

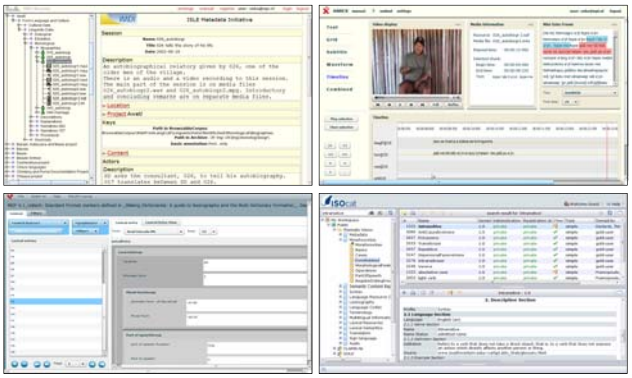
3. Digitale Klartextgrammatiken

b. Language Archiving Technology

- Entwickelt am Max-Planck-Institut für Psycholinguistik seit ca. 2000
- Context u.A.: Sprachdokumentation (DOBES)
- Sprachkorpora mit IMDI-Metadaten
- ELAN und ANNEX für Bearbeitung / Ansicht von annotierten Multimedia-Daten
- LEXUS für lexikalische Online-Datenbanken
- ISOcat Datenkategorieregister für Fachtermini
- Keine Komponente für wissensch. Meta-Texte wie z.B. typologische Arbeiten oder Grammatiken

3. Digitale Klartextgrammatiken

b. Language Archiving Technology



3. Digitale Klartextgrammatiken

c. Wiki-Grammatiken

Content Management Systeme & 'Wikis'

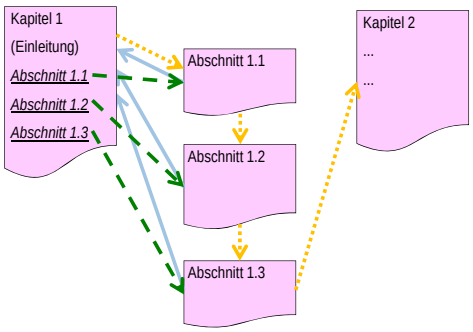
- Online-Zusammenarbeit
- Benutzerverwaltung
- Versionskontrolle
- Updates etc. werden von anderen bereitgestellt

Herausforderungen

- Spezielle Formatierung (*markup*, erweiterbar)
- Spezielle Funktionalitäten
- Serielle & hierarchische Anordnung von Seiten

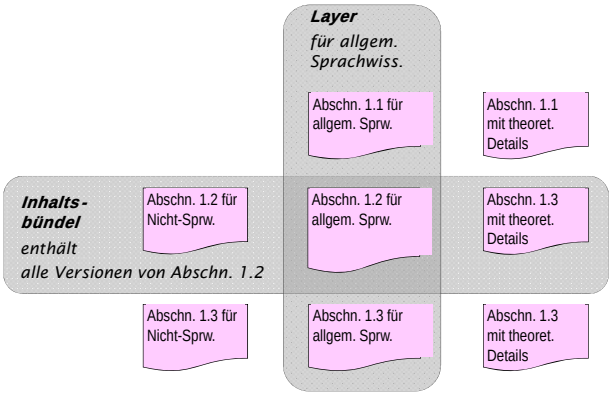
3. Digitale Klartextgrammatiken

c. Wiki-Grammatiken



3. Digitale Klartextgrammatiken

c. Wiki-Grammatiken



3. Digitale Klartextgrammatiken

c. Wiki-Grammatiken

Mehrdimensionale Anordnung der Seiten

- Lineare Reihenfolge (didaktischer Aufbau, wie Buch)
- Hierarchische Gliederung (Abschnitte, Kapitel, Teile)
- Verschiedene chronologische Versionen desselben Textes (Nachvollzug der Edierung der Beschreibung)
- Verschiedene *'Layer'* desselben Inhaltes für unterschiedliche Leserschaften (statt Fußnoten):
  - allg. Sprachwissenschaftler, Typologen;
  - Laien;
  - Spezialisten einer gewissen Theorie

3. Digitale Klartextgrammatiken

d. Die Text-Encoding Initiative (TEI)

- XML verspricht, ein dauerhafter Standard zu sein
- Lesbar für Menschen und Maschinen
- Die TEI-"Empfehlungen" bauen auf XML auf
- TEI ist ein weitverbreiteter *de-facto*-Standard
- TEI-XML sollte ein Format (zur Archivierung, zum Austausch) einer Digitalen Grammatik sein

Herausforderungen

- Wie wird der Text (XML) und Markup bearbeitet?
- Es gibt noch kein spezifisches TEI-Modul für linguistische beschreibende / typologische Texte

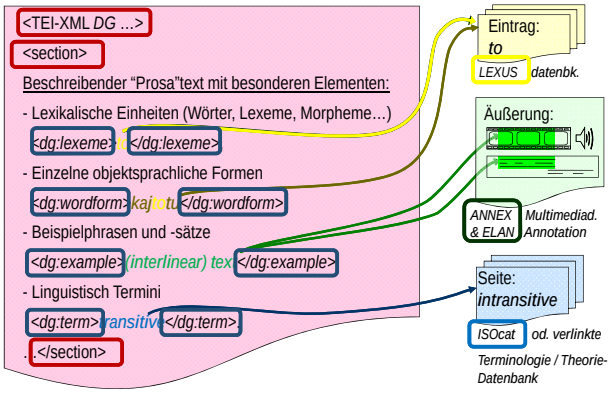
3. Digitale Klartextgrammatiken

d. Die Text-Encoding Initiative (TEI)

Linguistisch-ontologischer Typ	Tags und Eigenschaft.	Formatierg.	Haupt-funktionalität	Mögliche weitere Funktionalitätens (Tooltips und ähnlich)
syntaktische Einheit (Wortfolge)	<dg:SUnit > he goes </dg:SUnit >	*he goes* kursiv, Serifen	Abspielen Media-Datei	• zu Interlinearglossen • zum syntakt. Baum • Links z. Worteinträgen für einzelneWörter
Einzelnes Wort	<dg:word> goes </dg:word>	*goes* kursiv, Serifen	Link zum Worteintrag	• zu Interlinearglossen • ggf. Mediadatei absp.
Lexikalisches Wort	<dg:lexwd homn.nb =1 > go </dg:lexwd >	*go*$_1^W$ kursiv, Serifen Hochgest. W, Index 1	Springe zum Worteintrag (Auswahl bei Homonymen)	• zeige Bedeutung • zeige Wortart

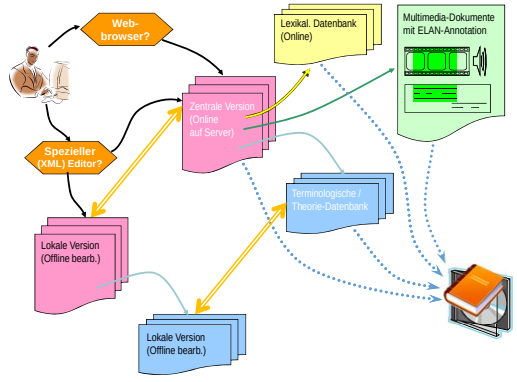
3. Digitale Klartextgrammatiken

d. Die Text-Encoding Initiative (TEI)



3. Digitale Klartextgrammatiken

e. Versionierung und Publikation



3. Digitale Klartextgrammatiken

e. Versionierung und Publikation

- Keine statischen, sondern „lebende Dokumente“
- Versionskontrolle (automatisch bei Wiki/CMS)
- Ausdruck der Grammatik in Buchform
- Digitale „*Snapshot*“-Distributionen erstellen
- Zitierbare Versionen und Distributionen
- Idealerweise ist möglich, an einer Offline-Kopie zu arbeiten (z.B. im Feld)
- Complexe Fragen der Synchronisierung
- und, noch einmal, die eines geeigneten Editors

4. Implementierte Grammatiken

a. Grammatiken als Programme

b. Das Grammatik-Matrix-Projekt

c. Interoperable Grammatiken

d. Andere implementierte Grammatiken

4. Implementierte Grammatiken

a. Grammatiken als Programme

- Computerlinguistik und NLP arbeiten seit langem an digitalen Repräsentationen von Grammatiken
- Verschiedene Systemtypen:
  - Regelbasiert
  - Statistisch
  - durch Analogiebildung
  - *Machine learning (unsupervised & supervised)*
- Verschiedene Ziele, meist morphologische und/oder syntaktische Analyse (Zuschreibung einer Struktur)

4. Implementierte Grammatiken

b. Das Grammatik-Matrix-Projekt

“*LinGO Grammar Matrix*” (E. Bender & al.)

- Regelbasiert: grammatische Regeln werden „kodiert“ und vom System angewendet
- Es analysiert („parst“) Sätze und ordnet ihnen syntaktische Baum-Strukturen zu
- *Treebank* (Datenbanken für Strukturbäume)
- Linguistisch präzise Grammatiken auf Grundlage des HPSG Rahmens
- Systeme mit anderen Theorien sind möglich

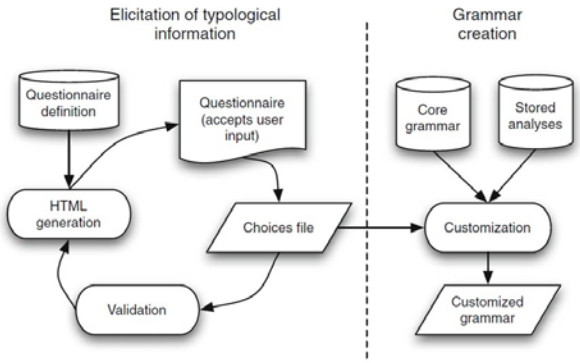
4. Implementierte Grammatiken

b. Das Grammatik-Matrix-Projekt

- Matrix-E-Grammatiken beruhen auf einer traditionelllen Beschreibung der Sprachstruktur
- Einige *Features* (Wortarten, -reihenfolge, usw.) werden übersetzt in „Zeichen“ (techn. Regeln)
- Lexikalische Einheiten sind auch „Zeichen“
- *Machine learning*: Korrekte Bäume wdn. erinnert
- Derzeit werden Grammatiken kompiliert von typologischen *Features*, erhoben m. Fragebogen
- Zukünftig sollen Grammatikfragmente extrahiert werden von Texten mit Interlinearglossen

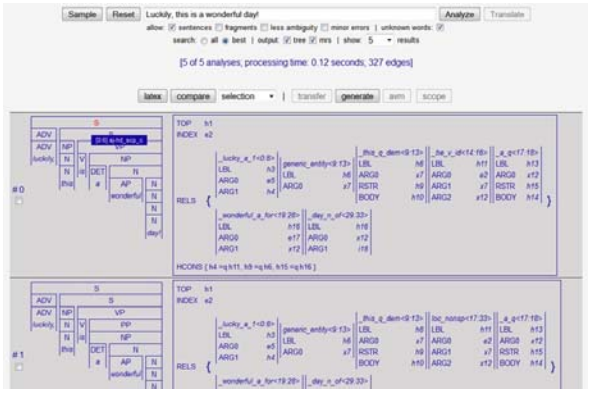
4. Implementierte Grammatiken

b. Das Grammatik-Matrix-Projekt



4. Implementierte Grammatiken

b. Das Grammatik-Matrix-Projekt



4. Implementierte Grammatiken

c. Interoperable Grammatiken

Interoperable Grammatiken (Mike Maxwell & al.)

- Idee: Verweben von Computercode und Klartext (“*literate programming*”)
- Auch hier: Implementierte Grammatiken, hier mit einem Fokus auf morphologischen Parsern
- Regeln kodiert in einem XML-Format, das verständlich für Linguisten und unabhängig vom letztendlich verwendeten Parser ist

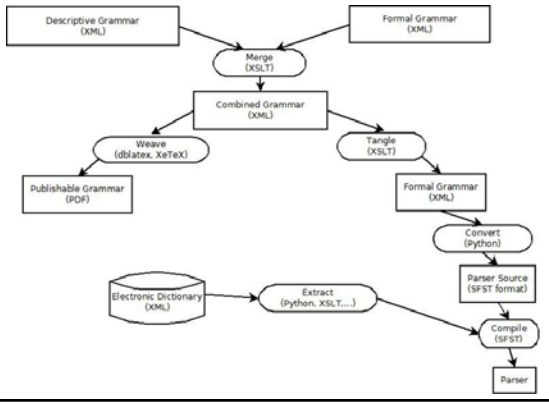
4. Implementierte Grammatiken

c. Interoperable Grammatiken

- Vorteil: Parser kommen und gehen, aber die Beschreibung bleibt immer technisch nutzbar
- Ein “Prä-Prozessor” konvertiert die XML-Regeln in das vom jeweiligen Parser gebrauchte Format
- Das Hauptergebnis der Anwendung der Grammatik ist ein morphologisch analysierter und „getaggter“ Text

4. Implementierte Grammatiken

c. Interoperable Grammatiken



4. Implementierte Grammatiken

c. Interoperable Grammatiken

Beispiel-Enkodierung eines Morphems

```
<Mo:InflectionalAffix gloss="-1Fut" id="af1Fut">
  <!--The two "allomorphs" are in fact allographs-->
  <Mo:Allomorph form="ো◌বা">
    <!--Spelled 'bo'; usually (not always) after a C-stem -->
  </Mo:Allomorph>
  <Mo:Allomorph form="ব">
    <!--Spelled 'b'; usually (not always) after a vowel stem -->
  </Mo:Allomorph>
  <Mo:inflectionFeatures>
    <Fs:f name="Tense"><Fs:symbol value="Future"/></Fs:f>
    <Fs:f name="Mood"><Fs:symbol value="Indicative"/></Fs:f>
    <Fs:f name="Person"><Fs:symbol value="1"/></Fs:f>
  </Mo:inflectionFeatures>
</Mo:InflectionalAffix>
```

4. Implementierte Grammatiken

d. Andere implementierte Grammatiken

Es gibt verschiedene weitere Projekte zu implementierten Grammatiken:

- ParGram (Butt , King et al.)
- Meta-Grammar (Kinyon et al.)
- KPML (Bateman et al.)
- Grammix (Müller)
- OpenCCG (Baldridge et al.)

Viele Systeme sind regelbasiert; statistische u. ä. Systeme hauptsächlich für resourcenreiche Spr.

5. Über- und Ausblick: Integration

Auswertung: Implementierte Grammatiken

- Beide computerlinguistischen Projekte zielen auf das *Parsen* von Sätzen (eines Korpus)
- Beider Architektur hat eine allgemeine *Parsing Engine*, die mit Regeln konfiguriert & angepasst und mit Daten von Einzelsprachen trainiert wird
- Vorteile: a) Gewinn von reicheren Daten und b) Erkennen von Lücken in Analyse u. Beschrbg.

5. Über- und Ausblick: Integration

Auswertung: Hypertextgrammatiken

- Weniger technisch innovativ als impl. Gram.
- Linguistik braucht ein allgem. Werkzeug für die Präsentation linguistischer Arbeit verknüpft mit Corpora, Lexika und terminolog. Definitionen
- Pionierarbeiten von Christian Lehmann, Sebastian Nordhoff, Nick Thieberger u. a.
- Jede Arbeit deckt einige, aber keine auch nur die Mehrzahl der notwendigen Aspekte ab

5. Über- und Ausblick: Integration

- Die drei Beispielprojekte überlappen und komplementieren einander
- Ideal wäre, sie miteinander und ähnlichen Lösungen modular zu kombinieren
- Interop.Gr.: Morphology, *literate programming*
- GrammarMatrix: Syntax, mögl. Erweiterung zu Semantik, typologische Verallgemeinerungen
- Hypertext Gr.: Präsentation als verlinkte Resourc.
- Andere Module: Statistische Ansätze

5. Über- und Ausblick: Integration

Durch so eine Integration gewännen wir:

- Reich annotierte Korpora
- Umfassende, empirisch begründete und nachprüfbare grammatische Beschreibung
- Ein tieferes Verständnis von Sprachvariabilität
- Die konzeptuelle und technische Grundlegung für das Verstehen von Sprachstruktur, als Basis für machinelles Verstehen und Übersetzung