

# Der Beitrag von Lernerkorpora zur Konstruktion und Validierung von Bewertungsskalen

Katrin Wisniewski

[Katrin.Wisniewski@tu.dresden.de](mailto:Katrin.Wisniewski@tu.dresden.de)



GAL-Kongress Marburg, 18.9.2014

# BEWERTUNGSSKALEN: HINTERGRUND

## Bewertungsskalen - Überblick

- weit verbreitet (NORTH 2000, ALDERSON 1991), oft in *high-stakes*-Tests verwendet → Einfluss auf wichtige Entscheidungen über das Leben von Sprachlernenden
- subjektive Einschätzung produktiver/interaktiver mündlicher/schriftlicher Leistungen in offeneren Testformaten
- sollen Aspekte sprachlicher Kompetenz hierarchisch ansteigend beschreibbar machen (HUDSON 2005)
- Vorteile:
  - betonen i.d.R. positive Eigenschaften der Lernersprache
  - bieten detaillierte **Deskriptoren**
  - zu Systemen zusammenstellbar (→ *Gemeinsamer europäischer Referenzrahmen*, GeRS)
- Problematisch:
  - **Reliabilität** (Inter-Rater- und Intra-Raterreliabilität) → Systematisierung der Subjektivität (North 1994)
  - **Validität** (→ forschungsrelevantes Konstrukt? → Bezug zu empirischer Lernersprache?)

## Bewertungsskalen – Skalierungsprozess

- **horizontale** Dimension: theorie- /modellgestützte Festlegung der zu erfassenden Aspekte der sprachlichen Kompetenz (sog. „Konstrukt“)  
BACHMAN 1990; BACHMAN/PALMER 1996/2010
- **vertikale** Dimension: unterschiedliche Ausprägung dieser Aspekte
  - Skalierungs-Prozess verläuft meist intransparent und intuitiv (Expertenurteile)  
Knoch 2011, North 2000, Europarat 1994

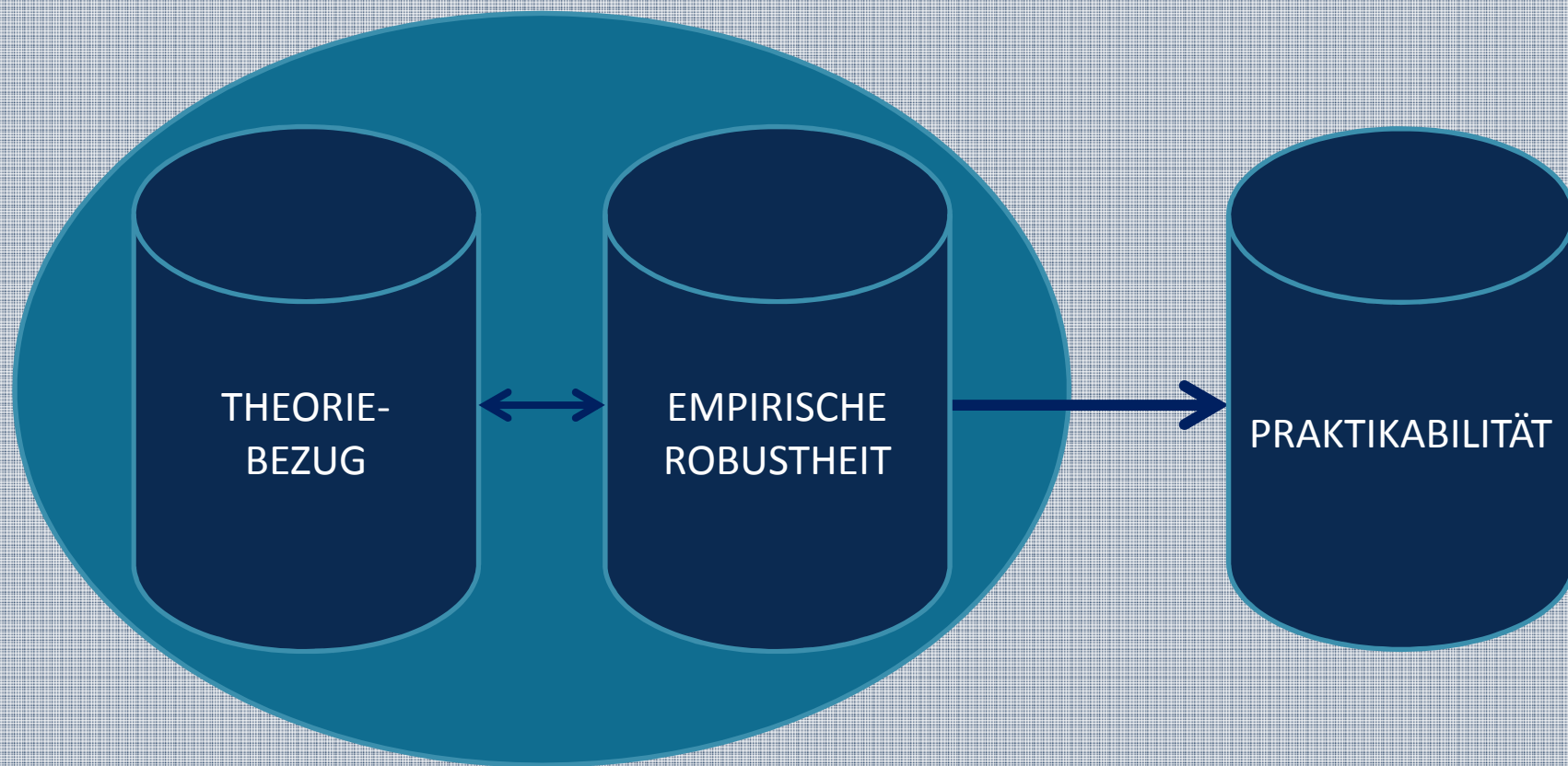
## Beispiel GeRS: Skalierungsverfahren

- vertikale Anordnung von **Expertenurteilen** mit Hilfe statistischer Verfahren („measurement-driven approach“)  
Europarat 2001, North 2000, Schneider/North 2000  
kritisch vgl. Fulcher et al. 2011, Hulstijn 2007, Wisniewski 2013, 2014
- nicht ‚empirisch‘, sondern (sattelfeste) **Konvention**
- Problem: fehlende Bezüge zur Lernaltersprache & Theorie  
→ **mangelnde Validität**
- alternative, auf Lernaltersprachenanalyse basierende Skalierungsverfahren (performance-data driven) bislang unüblich

Bachman/Cohen 1998, Fulcher 1987, 1993, 1996, Poonpon 2010

# DIE VALIDITÄT VON BEWERTUNGSSKALEN: EIN DREI-SÄULEN-ANSATZ

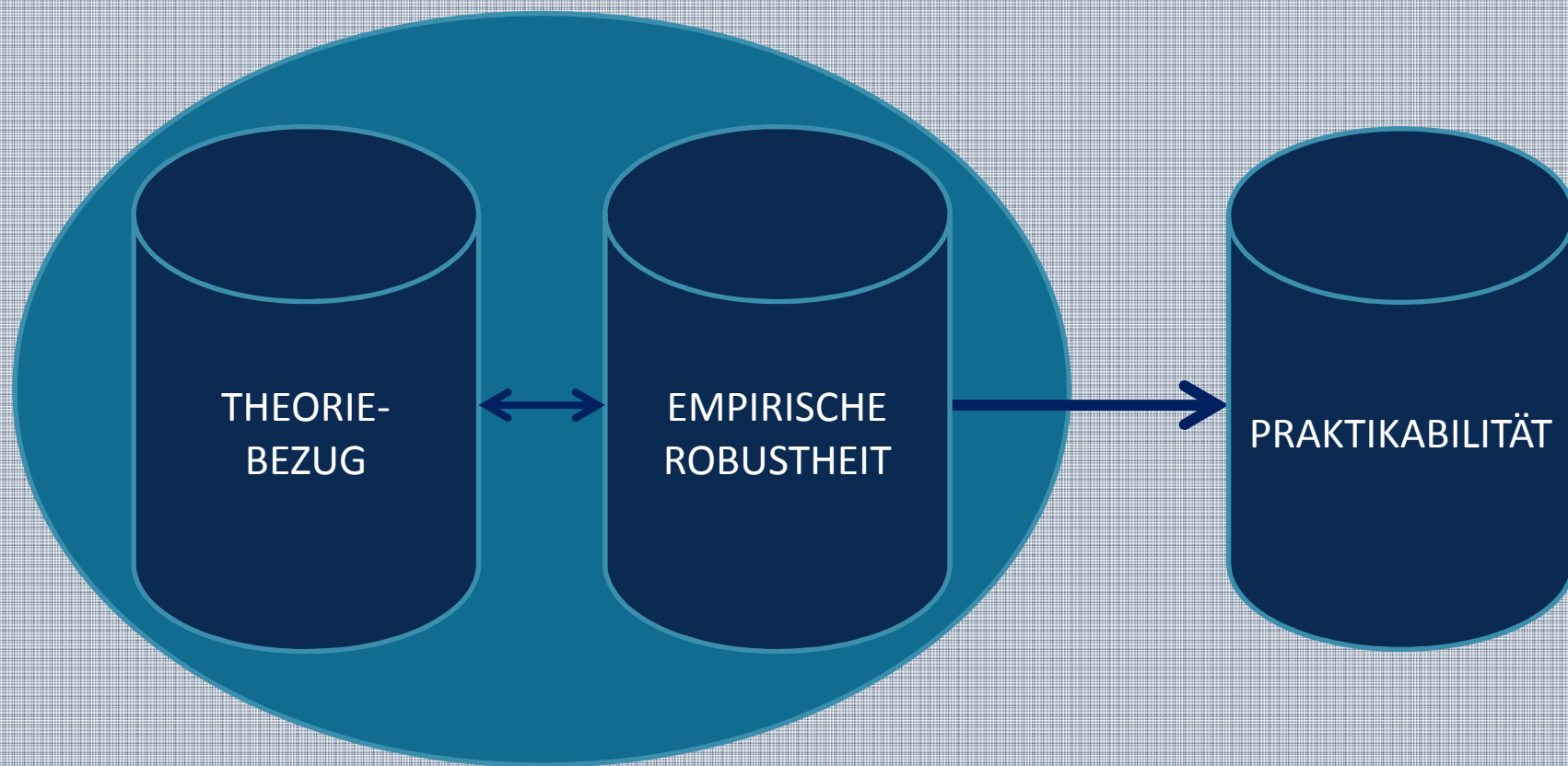
## Drei zentrale Säulen der Validität von ability-Skalen



# LERNERKORPORA ZUR KONSTRUKTION & VALIDIERUNG V. BEWERTUNGSSKALEN



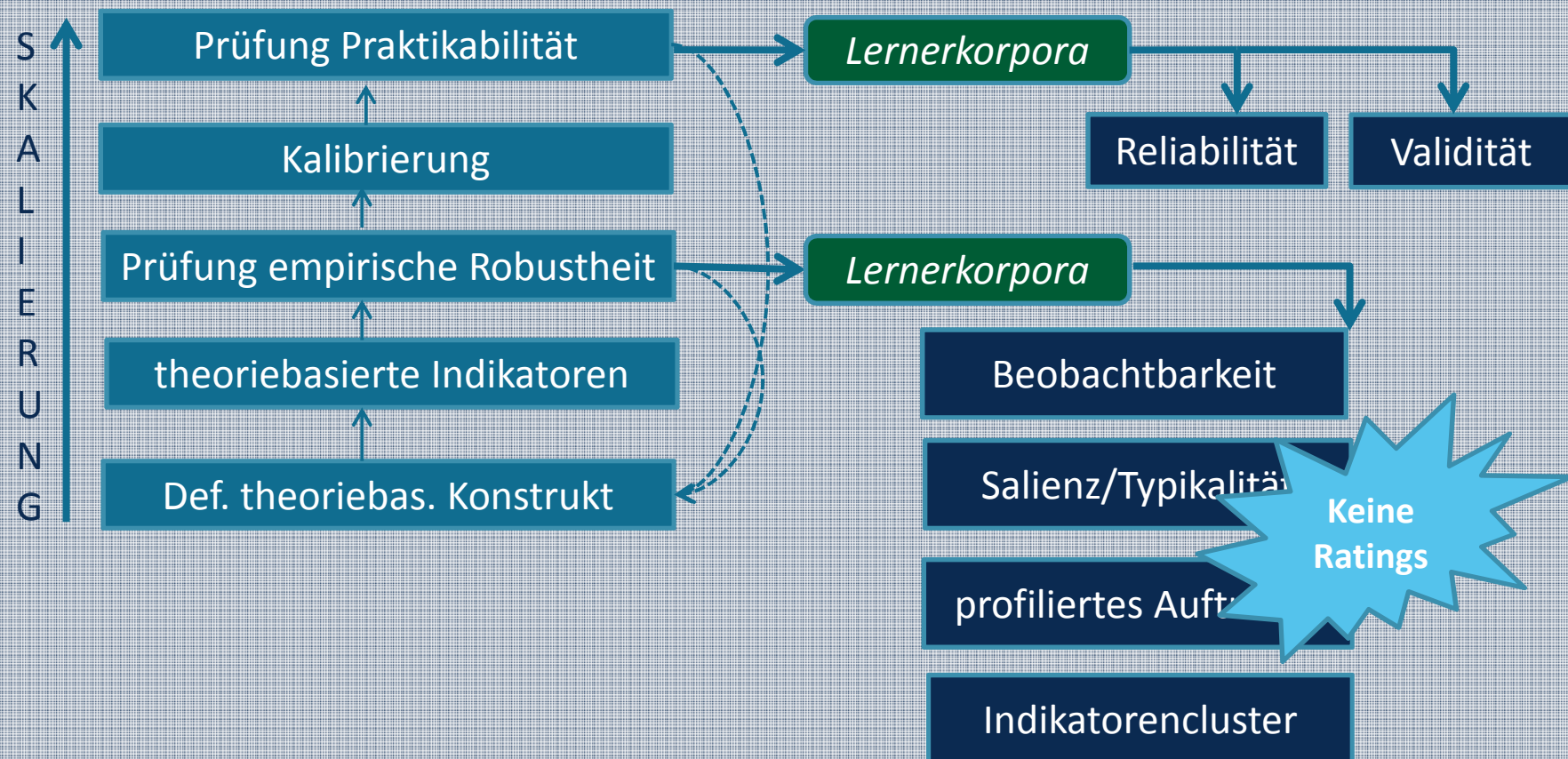
## Drei zentrale Säulen der Validität von ability-Skalen



## Lernerkorpora & Bewertungsskalen

- Lernerkorpora stützen die empirische Rückkoppelung existierender Bewertungsskalen an Lernaltersprache (*empirische Validierung*)
- Lernerkorpora können auch bei der *Erstellung neuer Skalen* verwendet werden
- noch nicht erkanntes Potenzial, aber in Sprachtestforschung zunehmender Rekurs auf (Lerner-)Korpora  
Alderson 1996, Ball 2001, Barker 2004, 2006, 2010, Flowerdew 2012
- beginnende Nutzung von Lernerkorpora zur *Illustration* bewerteter GeRS-Stufen (≠ *Validierung* von Skalen) vgl. English Profile Project: 'criterial features' (Hawkins/Filipović 2012)

# Lernerkorpora zur Skalenkonstruktion & -validierung



# STUDIE: EMPIRISCHE VALIDIERUNG VON SKALEN DES GERS (WISNIEWSKI 2014)

## Die Studie

- Pilotstudie (N=19)
- Skalen für Flüssigkeit, Wortschatzspektrum und Wortschatzbeherrschung des GeRS
- Niveaustufen A2-B2
- Zielsprachen: ITA/GER (Südtiroler Lernende)
- mündlicher Test, 2 Aufgaben (Monolog/Dialog)
- Qualitätssicherung streng

AERA/APA/NCME 1999; Bachman/Palmer 1996/2010; Europarat 2004, 2009

## Methoden

- Operationalisierung der GeRS-Skalen (‚Skalenvariablen‘, SV)
- Transkription (CHAT/Elan), Reliabilitätschecks
- Annotation (Skalenvariablen + forschungsbasierte Maße), Reliabilitätschecks
- Analysen (deskriptive Statistik, Korrelationen, Cluster- und Diskriminanzanalysen, Signifikanztests); **keine Generalisierungen möglich**

# Ergebnisse (empirische Robustheit)

## Einzelne Skalenvariablen

- *Beobachtbarkeit* eingeschränkt („unverständliche Aussagen“)
- teils keine *Profilierung*, („Grundwortschatz“ >90% bei ALLEN Lernenden) → Beurteilungsdilemma
- Skaleninhalte *spekulativ* („Pausenursachen“ bei 55.45% unklar; Kodierer-Reliabilität beim Rest:  $C=.936$  aber  $\kappa=.51$ )

## Niveaubeschreibungen

- große Heterogenität bei Ausprägungen der Skalenvariablen, keine *Typikalität*
- *Clusterbildung* (Passung von Lernerproduktionen auf gesamte Niveaubeschreibungen) nur teilweise überhaupt möglich. Außerdem:
  - auf viele Produktionen passt keine Niveaubeschreibung  $N=43 / 114$ )
  - Mehrdeutigkeit vieler Zuordnungen ( $N=24 / 114$ )
- Allgemein: Aufgaben- und Textlängenabhängigkeit



## Ergebnisse (Praktikabilität)

- Kein einziges Urteil kann vollständig durch SV erklärt werden (Wilks Lambda, Diskriminanzanalysen)
- Plausible Ausweichstrategien der Beurteilerinnen:  
(Fragebogen-Triangulierung & statist. Analysen)
  - Verallgemeinerungen anderer Skalenniveaus ('Weitersprechen nach Formulierungsproblem')
  - Uminterpretationen von Skaleninhalten
  - skalen-externe Einflussfaktoren, konstruktrelevant (*speech rate, mean length of runs*, Lexikalische Dichte, Lexikalische Variation [Advanced Guiraud/Guiraud])
  - griffige Konzepte (gefüllte Pausen, Pausenpositionen, Abbrüche, Verständlichkeit, Fehlerzahl), teilweise konstruktrelevant



# DAS MERLIN-LERNERKORPUS ZUR VALIDIERUNG VON GERS- SKALEN

# MERLIN – ein dreisprachiges, GeRS-bezogenes Lernerkorpus

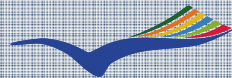
MERLIN: „**M**ultilingual Platform for the **E**uropean **R**eference **L**evels: **I**nterlanguage Exploration in Context“: [www.merlin-platform.eu](http://www.merlin-platform.eu)

Förderung im Programm Lebenslanges Lernen der EU (LLP 518989-LLP-1-2011-1-DE-KA2-KA2MP), Laufzeit 01/2012 – 12/2014

**Partner:** Technische Universität Dresden (DE) (Koordination), Europäische Akademie Bozen (IT), Karls-Universität Prag (CZ), telc GmbH (DE), Berufsförderungsinstitut Oberösterreich (AT), Eberhard-Karls-Universität Tübingen (DE); European Center of Modern Languages - Council of Europe (AT) (assoziiierter Partner))

## Ziele:

- frei verfügbare Online-Plattform zur **Illustrierung** der GeRS-Stufen für Deutsch – Italienisch – Tschechisch (basierend auf qualitätsgeprüften Ratings & auditierten Tests)
- Beitrag zur **Validierung** ausgewählter GeRS-Skalen



Education and Culture DG

Lifelong Learning Programme  
06.03.2015

K. Wisniewski: Lernerkorpora & Bewertungsskalen  
GAL-Kongress, Marburg 2014

Folie Nr. 18 von XYZ

## MERLIN-Korpus

- 2.286 Texte (ITA: 812, DE: 1.033, CZ: 441) aus standardisierten Sprachtests, Niveaus A1-C1 (telc Frankfurt, UJOP Prag)
- Nachbewertungen → xml-basierte Transkriptionen → (man./autom.) Annotationen
- Annotationsebenen:
  - Zielhypothesen 1 u. teilweise 2 (→ Falko-Projekt)
  - Merkmale der Lernaltersprache (Fehler u.a.)
- momentan Pilotierung der Plattform
- Quellen für Annotationen: **GeRS**, Forschung, Nutzer, Textanalysen

## Annotations-Tags aus GeRS-Deskriptoren in MERLIN

- **Skalen** aus Kapitel 5, GeRS (→ kommunikative L2-Kompetenz)
  - **Grammatische Korrektheit**
  - **Orthographie (orthogr. Fehler)**
  - Wortschatzspektrum (z.B. ‚Umschreibungen‘)
  - Wortschatzbeherrschung (z.B. ‚gute Beherrschung des Grundwortschatzes‘)
  - Soziolinguistische Angemessenheit (z.B. ‚Anrede‘)
  - Kohärenz/ Kohäsion (z.B. ‚einfache Konnektoren‘, ‚sprunghaft‘)
- **Grenzen:**
  - Subjektivität, Uneindeutigkeit der Skalen
  - mangelnde Passung auf das Korpus, Aufwand

# Beispiel Operationalisierung Grammatikskala

## Grammatische Korrektheit (GeRS 2001)

C2	Zeigt auch bei der Verwendung komplexer Sprachmittel eine durchgehende Beherrschung der Grammatik, selbst wenn die <u>Aufmerksamkeit anderweitig beansprucht wird (z. B. durch vorausblickendes Planen oder Konzentration auf die Reaktionen anderer).</u>
C1	Kann beständig ein hohes Maß an grammatischer Korrektheit beibehalten; Fehler sind selten und fallen kaum auf.
B2	Gute Beherrschung der Grammatik; gelegentliche Ausrutscher oder nichtsystematische Fehler und kleinere Mängel im Satzbau können vorkommen, sind aber selten und können oft rückblickend korrigiert werden.
B1	Gute Beherrschung der Grammatik; macht keine Fehler, die zu Missverständnissen führen. Kann sich <u>in vertrauten Situationen</u> ausreichend korrekt verständigen; im Allgemeinen gute Beherrschung der grammatischen Strukturen <u>trotz deutlicher Einflüsse der Muttersprache</u> . Zwar kommen Fehler vor, aber es bleibt klar, was ausgedrückt werden soll. Kann ein Repertoire von häufig verwendeten Redefloskeln und von Wendungen, die an <u>eher vorhersehbare Situationen</u> gebunden sind, ausreichend korrekt verwenden.
A2	Kann einige einfache Strukturen korrekt verwenden, macht aber noch systematisch elementare Fehler, hat z.B. die Tendenz, Zeitformen zu vermischen oder zu vergessen, die Subjekt-Verb-Kongruenz zu markieren; trotzdem wird in der Regel klar, was er/ sie ausdrücken möchte.
A1	Zeigt nur eine begrenzte Beherrschung einiger weniger einfacher grammatischer Strukturen und Satzmuster in einem <u>auswendig gelernten Repertoire</u> .

## Beispiel Operationalisierung Grammatikskala

- Annahmen in der Skala (*allgemein*):
  - „schwerere“ Fehler auf niedrigeren Niveaustufen
  - Zusammenhang von Korrektheit & Komplexität
  - ansteigendes Niveau, weniger Fehler
- Annahmen in der Skala (*niveaubezogen*):
  - Tempus-Verwechslungen (bis A2), Kongruenzfehler (bis A2), Selbstkorrektur-Fähigkeit (B2+), Verständlichkeit (A2, B1+)
- Operationalisierung
  - Tags:
    - *Fehler, die die Verständlichkeit beeinträchtigen (Teilaspekt Fehlerschwere), Tempusfehler, Kongruenzfehler, Selbstkorrekturen, verschiedene Grammatikfehler*
  - Maße:
    - *Komplexitäts- und Korrektheitsmaße (z.B. mittl. Länge T-Units, Anzahl gramm. Fehler pro T-Unit/Satz/Token, vgl. etwa Lu 2010) % fehlerfreie T-Units, Fehler/T-Unit*

# DISKUSSION/AUSBLICK



## Diskussion der Einsatzmöglichkeiten von Lernerkorpora

- nötig zur Skalierung & Validierung: passende **Annotationen**
  - bislang nicht sehr **viele**, nicht sehr **zugängliche** und nicht sehr **große** Lernerkorpora in nicht sehr vielen **Sprachen**, mit sehr wenigen **Aufgabenformaten**
- eigene Annotationen sind sehr zeitintensiv



## Diskussion der Einsatzmöglichkeiten von Lernerkorpora

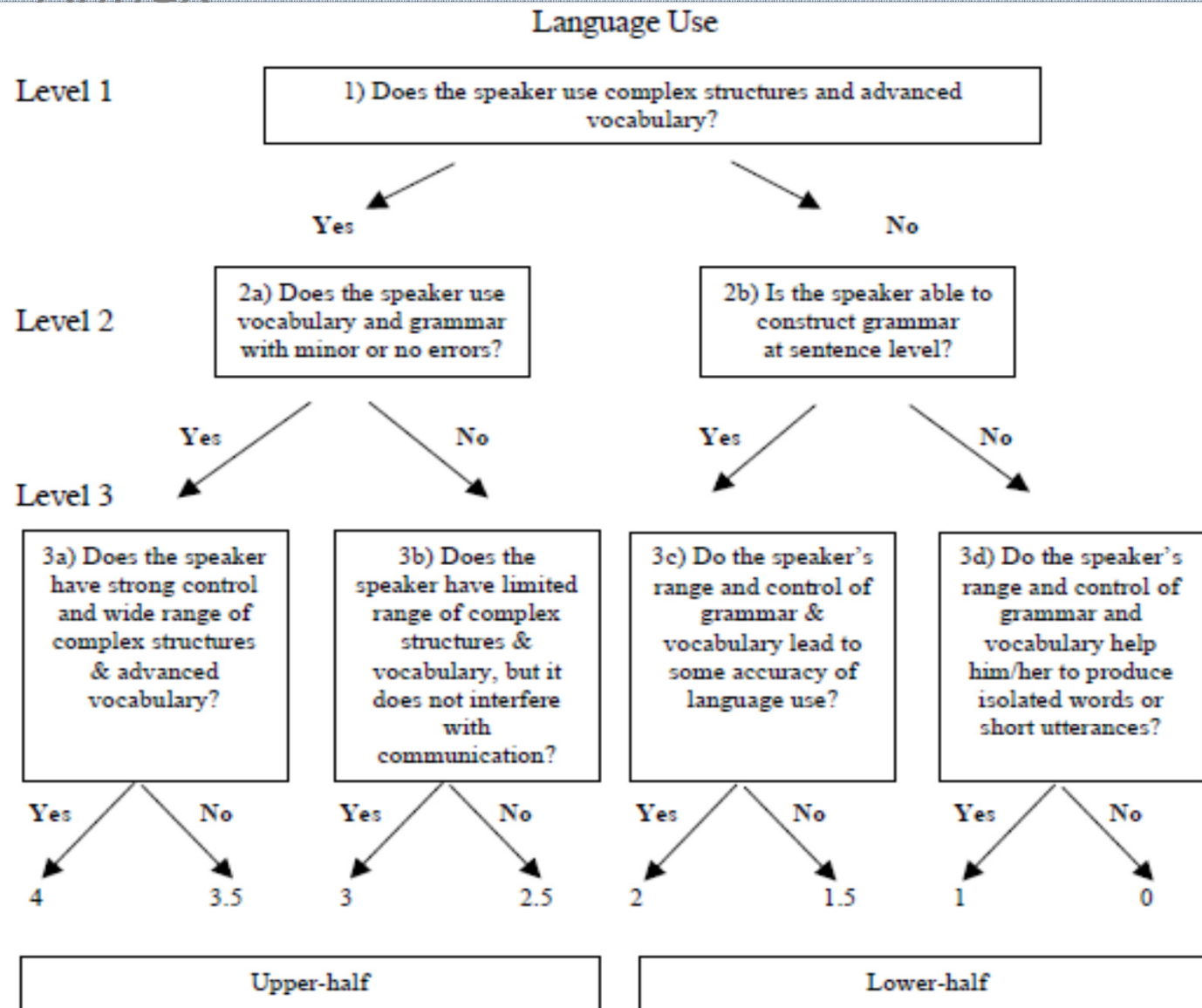
- **Ziel: Fairness gegenüber Lernenden!**
- nötiger Aufwand für empirische Validierung abhängig von Testtyp (**Konsequenzen**)
- Möglich: qualitative Erprobung besonders kritischer **Teilaspekte von Skalen** in **kleinen Stichproben** (Kane 2006, 2013, Messick 1989)
- auch nicht annotierte Lernerkorpora erleichtern dabei die Arbeit
- **Konsequenzen**: gegebenenfalls Revision/ Neuskalierung

“If descriptors are to be meaningful characterizations of ability, then they should be able to be related to actual performance”  
(Alderson 1991:74)

**Vielen Dank für Ihre Aufmerksamkeit!**

Katrin.Wisniewski@tu-dresden.de

# Annex



EBB,  
Poonpon  
2010

Figure 2. A proposed rating guide for language use.