Spektrum Patholinguistik 7 (2014) 123–131

Altersgruppeneffekte in childLex

Kay-Michael Würzner[1], *Julian Heister*[1], *Sascha Schroeder*[2]

[1] Universität Potsdam
[2] Max-Planck-Institut für Bildungsforschung, Berlin

1 Einleitung

Design und Analyse psycholinguistischer Experimente hängen wesentlich von der Qualität des zugrundeliegenden sprachlichen Materials ab. Die Materialauswahl erfolgt im Allgemeinen auf der Basis bestimmter Eigenschaften der zu untersuchenden sprachlichen Einheiten. Diese Eigenschaften können in sogenannten lexikalischen Datenbanken nachgeschlagen werden. Dabei wird vorausgesetzt, dass diese Datenbanken bis zu einem gewissen Grade für alle Probanden repräsentativ sind. Diese Allgemeingültigkeit gilt jedoch nicht in der Spracherwerbsforschung: Die Sprache von Kindern und Erwachsenen differiert – beispielsweise im Bereich des mentalen Lexikons (Prengel & Rickheit, 1987) – in einem solchen Maße voneinander (vgl. auch Augst, 1989; Tomasello, 2000), dass die Anwendung für den Erwachsenensprachgebrauch verfügbarer, lexikalischer Datenbanken (z. B. Brysbaert & New, 2009; Heister et al., 2011) für Studien, die sich mit der Sprache von Kindern beschäftigen, nur sehr eingeschränkt möglich ist.

An dieser Stelle setzt childLex an: Das Verbundprojekt des Max-Planck-Institutes (MPI) für Bildungsforschung, der Universität Potsdam sowie der Berlin-Brandenburgischen Akademie der Wissenschaften (BBAW) hat zum Ziel, der psycholinguistischen und der psychologischen Forschung neue linguistische Normen zur Schriftsprache von Kindern im Grundschulalter zur Verfügung zu stellen. Diese werden auf Basis eines großen Korpus an Kinderbüchern erhoben, das in drei Altersbereiche unterteilt ist: 6–8 Jahre, 9–10 Jahre und 11–12 Jahre. Während sich Schroeder, Würzner, Heister, Geyken und Kliegl (2014a) mit den Unterschieden von Kinder- und Erwachsenensprache beschäftigen und Schroeder, Würzner, Heister,

Geyken und Kliegl (2014b) childLex mit lexikalischen Datenbanken zur Kindersprache in anderen Sprachen vergleichen, konzentriert sich diese Arbeit auf einen Vergleich der drei Altersgruppen.

Zunächst fassen wir aber den aktuellen Stand der Forschung sowie Datengrundlage, Volltexterfassung und die durchgeführten linguistischen Analysen zusammen.

2 Stand der Forschung

Anders als in anderen Sprachen gibt es für das Deutsche relativ wenige Ansätze zur Erfassung und Dokumentation der Schriftsprache von Kindern im Grundschulalter (vgl. Schroeder et al., 2014b), die allesamt auf sehr kleinen Korpora beruhen (Die größte Quelle, das Braunschweiger Korpus, umfasst beispielsweise ca. 260 000 Wörter.). Das liegt vor allem daran, dass sich die bislang vorliegenden Wortlisten auf die Erfassung des produktiven Wortschatzes konzentrieren, während Kinderkorpora in anderen Sprachen stets auf den rezeptiven Wortschatz abzielen.

Auch beim Umfang der erfassten Normen fallen für das Deutsche verfügbare Quellen hinter ihren Pendants in anderen Sprachen zurück: Variablen wie Wortart, Grundform, Nachbarschaftsdichte etc. liegen im Bereich der Kindersprache derzeit nicht vor.

3 Methode

3.1 Datengrundlage

Das childLex zugrundeliegende Korpus umfasst in Version 0.06 500 Bücher, die von Kindern im Alter von 6–12 Jahren gelesen werden. Der überwiegende Teil ist fiktionalen Inhalts. Bei der Korpuszusammenstellung haben wir darauf geachtet, dass die Bücher auch wirklich von Kindern gelesen werden. Zu diesem Zweck wurden Ausleih- bzw. Verkaufsstatistiken des Berliner Bibliotheksverbundes und mehrerer Onlinebuchhändler herangezogen. Darüber hinaus haben wir

Selbstaussagen von Kindern (z. B. Fragebögen in der „KinderZEIT") und Lehrern (Ländervergleich Grundschule 2011; vgl. Stanat et al., 2012) verwendet. Bei der Einteilung in die drei Altersgruppen 6–8 Jahre (1.–2. Klasse; AG1), 9–10 Jahre (3.–4. Klasse; AG2) und 11–12 Jahre (5.–6. Klasse; AG3) haben wir uns an den Verlagsangaben orientiert. Da die einzelnen Bücher in ihrem Umfang stark differieren und Bücher für jüngere Leser oft wesentlich kürzer sind als solche für ältere Leser, unterscheidet sich die Anzahl der Bücher in den einzelnen Altersgruppen. Tabelle 1 fasst die Altersgruppeneinteilung zusammen.

Tabelle 1
Zusammenfassung der Altersgruppen in childLex (v0.06)

Altersgruppe	**Bücheranzahl**	**Tokens pro Buch (M)**
6–8 Jahre (AG1)	218	6 322,39
9–10 Jahre (AG2)	205	18 314,21
11–12 Jahre (AG3)	70	48 890,71

3.2 Volltexterfassung und linguistische Analyse

Die ausgewählten Bücher wurden zunächst eingescannt. Die anschließende Volltexterfassung erfolgt mit Hilfe der Software *FineReader* (Version 11) der Firma Abbyy. Problematische Bereiche wie Abbildungen, Inhaltsverzeichnisse, Titelblätter und Impressen wurden von der Erkennung ausgeschlossen. Die Qualität der Erfassung ist generell sehr gut (Korrektheit auf Zeichenebene > 99 %), was zum einen den guten Vorlagen, zum anderen aber auch dem vergleichsweise einfachen Textlayout (Schriftgröße, typographische Elemente, teilweise reduzierte Silbentrennung etc.) zu verdanken ist.

Die Volltexte wurden dann mit Hilfe des statistischen Tokenizers *waste* (Jurish & Würzner, 2013) in Tokens und Sätze zerlegt. Den einzelnen Tokens werden dann mit Hilfe des automatischen, morphologischen Analysesystems TAGH (Geyken & Hanneforth, 2006) Grundformen (Lemmata) und mögliche Wortarten zugewiesen, aus denen der Part-of-Speech Tagger *moot* (Jurish, 2003) die im Satzkontext wahrscheinlichste auswählt.

4 Analysen

Mit einer Reihe von deskriptiven Analysen wollen wir die Unterschiede zwischen den einzelnen Altersgruppen verdeutlichen. Tabelle 2 dokumentiert wichtige korpuslinguistische Kennwerte für die drei Gruppen.

Tabelle 2
Wichtige korpuslinguistische Kennwerte für die Altersgruppen (AG) in childLex

Kennwert	**AG1**	**AG2**	**AG3**
Tokenanzahl	1 499 505	4 055 227	3 684 402
Typeanzahl	54 455	108 306	99 572
Lemmaanzahl	37 474	73 932	65 146
Satzlänge in Tokens M (SD)	11,37 (6,35)	12,48 (7,3)	13,06 (8,15)
Tokenlänge in Zeichen M (SD)	4,01 (2,77)	4,20 (2,86)	4,23 (2,85)
Typelänge in Zeichen M (SD)	6,71 (3,02)	7,34 (3,14)	7,90 (3,18)

Auch wenn sie bezüglich der Anzahl an Büchern die größte Gruppe darstellt, bringt AG1 naturgemäß die kleinste Textmenge ein. Bücher für jüngere Kinder erreichen nicht die Komplexität von Büchern für ältere Kinder. Auch in Bezug auf Satz- und Wortlänge deutet sich an,

dass die sprachliche Komplexität von Kinderbüchern mit dem empfohlenen Lesealter steigt. Wir wollen diese Hypothese im Folgenden weiter überprüfen. Dazu betrachten wir zunächst die Type-Token-Ratios der drei Altersgruppen (Abbildung 1). AG1 hat erwartungsgemäß die geringste Anzahl verschiedener Wörter pro Stichprobe und demnach das kleinste Wortinventar. Die AGs 2 und 3 hingegen zeigen nicht das erwartete Muster: Die Anzahl der Types pro Stichprobe in AG2 übersteigt die in AG3. Der Grund für dieses abweichende Muster liegt vermutlich in der geringeren Anzahl an Büchern in AG3.

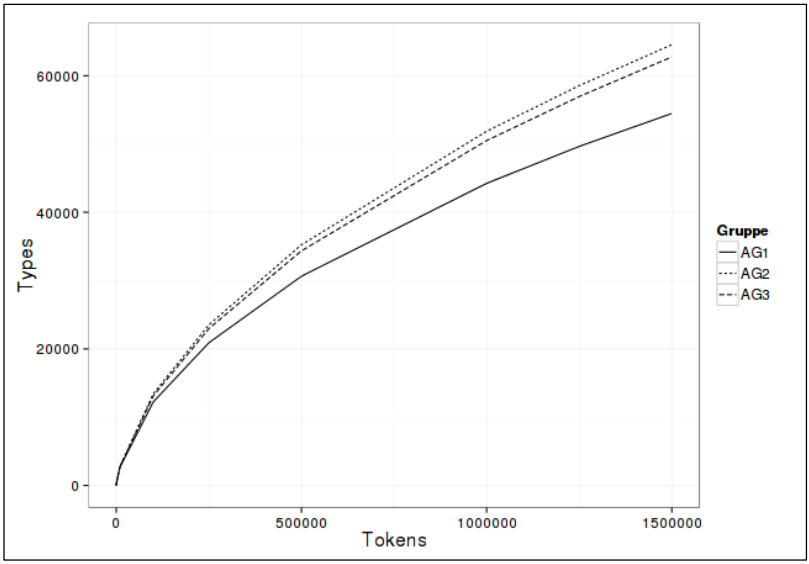

Abbildung 1. Type-Token-Ratios für die Altersgruppen (AG) in childLex

Betrachten wir nun die Verteilung der verschiedenen Wortarten in den einzelnen Altersgruppen. Tabelle 3 fasst diese für die wichtigsten Wortarten auf Type- und Tokenebene zusammen. Die Wortartenverteilung auf Typeebene bestätigt die Type-Token-Ratios: Der Anteil von Nomen und Adjektiven steigt mit zunehmendem Lesealter. Dies deckt sich mit den Beobachtungen von Ruoff (1981), der lexikalische

Varianz vor allem im nominalen Bereich verortet. Mit anderen Worten, der Bestand anderer Wortarten ist mit zunehmender Tokenanzahl ab einem bestimmten Punkt „gesättigt".

Tabbelle 3
Wortartenanteile auf Type-/Tokenebene in Prozent für die Altersgruppen (AG) in childLex

Wortart	**AG1**	**AG2**	**AG3**
Normale Nomen	42,39 /12,00	43,91 /11,98	47,92 /11,05
Eigennamen	11,61 / 4,04	12,87 / 3,78	14,04 / 3,70
Verben	23,99 /14,54	20,99 /14,71	23,23 /14,76
Adjektive	15,65 / 4,87	16,80 / 5,14	18,33 / 5,33
Präpositionen	0,30 / 5,19	0,20 / 5,44	0,21 / 5,53
Pronomen	1,37 /14,17	0,84 /14,91	0,97 /15,93
Konjunktionen	0,17 / 4,31	0,10 / 4,53	0,11 / 4,61
Kommas	0,00 / 5,06	0,00 / 5,46	0,00 / 5,85

Auf Tokenebene kann man eine Zunahme der Funktionswortkategorien erkennen, was für eine komplexere Satzstruktur mit zunehmendem Lesealter spricht. Gleichzeitig nimmt der Anteil der Nomen entsprechend ab. Der größere Anteil der Adjektive spricht für komplexere Nominalphrasen. Kontraintuitiv erscheint auf den ersten Blick der steigende Anteil der Verben zu sein. Nimmt man für AG1 eine einfache Satzstruktur mit vielen kurzen Deklarativsätzen an, so sollte der Anteil der Verben im Tokenbereich eigentlich abnehmen. Wir haben uns deshalb die unterschiedlichen Verbformen näher angesehen (Tab. 4). Der Anteil infiniter Formen und Partizipien steigt über die drei Altersgruppen hinweg an, während der Anteil finiter Verbformen sinkt. Dies unterstützt unsere Eingangshypothese.

Tabelle 4
Anteil unterschiedlicher Verbrealisierungen auf Tokenebene für die Altersgruppen (AG) in childLex in Prozent

Verbform	**AG1**	**AG2**	**AG3**
finite Verbform	11,47	11,25	11,16
infinite Verbform	2,19	2,51	2,65
Partizipform	1,04	1,31	1,41
Imperativ	0,15	0,13	0,13

5 Diskussion

Wir haben in dieser Arbeit einen kurzen Überblick über die drei Altersgruppen in childLex, einer lexikalischen Datenbank für die Schriftsprache in Kinderbüchern, gegeben. Die beobachteten Unterschiede zwischen den Altersbereichen sind im Vergleich zu den Unterschieden, die wir zur Erwachsenensprache verzeichnen (Schroeder et al. 2014a), eher gering, bestätigen uns aber in der Annahme, dass die von Kindern in Kinderbüchern rezipierte Sprache mit steigendem empfohlenen Lesealter komplexer wird. Die verwendeten Daten werden im Laufe des Jahres 2014 über die lexikalische Auskunftsplattform www.dlexdb.de verfügbar gemacht.

6 Literatur

Augst, G. (1989). *Schriftwortschatz. Untersuchungen und Wortlisten zum orthographischen Lexikon bei Schülern und Erwachsenen*. Frankfurt a. M.: Peter Lang.

Brysbaert, M. & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behaviour Research Methods*, *41*, 977–990.

Geyken, A. & Hanneforth, T. (2006). TAGH: A complete morphology for German based on weighted finite state automata. In A. Yli-Jyrä, L. Karttunen & J. Karhumäki (Hrsg.), *Finite State Methods and Natural Language Processing* (55–66). Berlin: Springer.

Heister, J., Würzner, K.-M., Bubenzer, J., Pohl, E., Henneforth, T., Geyken, A. & Kliegl, R. (2011). dlexDB: Eine lexikalische Datenbank für die psychologische Forschung. *Psychologische Rundschau*, *62*, 10–20.

Jurish, B. & Würzner, K.-M. (2013). Word and sentence tokenization width hidden Markov models. *Journal of Language Technology and Computational Linguistics*, *28*, 61–83.

Jurish, B. (2003). *Part-of-Speech tagging with finite state morphology*. Poster präsentiert auf der Konferenz Collocations and Idioms: Linguistic, Computational, and Psycholinguistic Perspectives, Berlin.

Pregel, D. & Rickheit, G. (1987). *Der Wortschatz im Grundschulalter*. Hildesheim: Olm.

Ruoff, A. (1981). *Häufigkeitswörterbuch gesprochener Sprache*. Tübingen: Narr.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A. & Kliegl, R. (2014a). *childLex: Eine lexikalische Datenbank zur Sprache in deutschen Kinderbüchern*. Manuskript, Max-Planck-Institut für Bildungsforschung, Berlin.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A. & Kliegl, R. (2014b). childLex: A lexical database of German read by children. Manuskript, zur Veröffentlichung eingereicht. Max Planck Institute for Human Development, Berlin.

Stanat, P., Pant, H. A., Böhme, K. & Richter, D. (Hrsg.). (2012). *Kompetenzen von Schülerinnen und Schülern am Ende der vierten Jahrgangsstufe in den Fächern Mathematik und Deutsch*. Münster: Waxmann.

Tomasello, M. (2000). Do children have adult syntactic competence? *Cognition*, *74*, 209–253.

Kontakt

Kay-Michael Würzner
wuerzner@bbaw.de