Spektrum Patholinguistik 8 (2015) 183–200

Morphologische und phonologische Repräsentationen in childLex

Kay-Michael Würzner [1] *& Sascha Schroeder* [2]

[1] Berlin-Brandenburgische Akademie der Wissenschaften
[2] Max-Planck-Institut für Bildungsforschung, MPFG REaD

1 Einleitung

Lexikalische Normen, die die unterschiedliche Nutzung und Vertrautheit mit sprachlichen Materialien im Laufe der kindlichen Entwicklung dokumentieren, sind ein wichtiges Instrument, das Forschung und Praxis erlaubt, altersadäquate Stimuli und linguistische Kontexte, z. B. für empirische Studien oder die Konzeption von Trainings- und Therapiematerialien, auszuwählen. Für das Deutsche lagen bislang nur Normen für Erwachsene vor (CELEX: Baayen, Piepenbrock & Guilkers, 1996; DWDS: Heister et al., 2011; SubtlexDE: Brysbaert et al., 2011).

Dieses Angebot wird nun durch childLex (Schroeder, Würzner, Heister, Geyken & Kliegl, im Druck a) ergänzt, das lexikalische Normen für die Schriftsprache zur Verfügung stellt, die Kinder im Alter zwischen sechs und zwölf Jahren lesen. Bisherige Analysen haben gezeigt, dass sich die linguistischen Merkmale von Kinder- und Erwachsenen-(Schrift-)Sprache sowohl quantitativ als auch qualitativ voneinander unterscheiden (Schroeder, Würzner, Heister, Geyken & Kliegl, im Druck b) und insbesondere Sprache für jüngere Kinder sich von älteren Altersgruppen unterscheidet (Würzner, Heister & Schroeder, 2014). Dabei zeigt sich folgendes charakteristisches Muster: Unterschiede im lexikalischen und superlexikalischen, d. h. Wortverbindungen betreffenden Bereich, sind besonders groß und werden immer geringer, je basaler das linguistische Beschreibungsniveau wird. So finden sich z. B. auf der sublexikalischen Ebene, d. h. bei einzelnen Buchstabenverbindungen, kaum Unterschiede zwischen Kinder- und Erwachsenensprache.

Obwohl childLex allein auf schriftsprachlichen Materialien basiert, ist der Anteil konzeptueller Mündlichkeit in den Materialien hoch, d. h. es wird häufig auf quasi-orale Elemente wie direkte Rede oder die Imitation mündlicher Sprache zurückgegriffen. Das schlägt sich z. B. im Anteil dialektaler Sprache („Ick ben een Berlina, wa!"), Klitisierungen („Was'n los?") oder Aspekten der Artikulation („W-w-w-wirklich?", „Isch nuschel schon scheit meiner Kindheit.") nieder. Gerade bei Leseanfängern wird also versucht, sich an Charakteristika der den Kindern bereits vertrauten mündlichen Sprache zu orientieren. Darüber hinaus werden diese Materialien den Kindern bereits im Kindergartenalter häufig vorgelesen und stellen damit einen wichtigen Inputanteil ihrer sprachlichen Umgebung dar. Es ist deswegen zu erwarten, dass sich verschiedene Aspekte von childLex auch auf den prä-literalen Bereich übertragen lassen.

Allgemein kann festgestellt werden, dass die phonologische und morphologische Beschreibungsebene im vorschulischen Bereich besonders wichtig ist (Szagun, 2013). Im Schulanfangsalter besteht die entscheidende Lernleistung gerade in der Rekonstruktion der Beziehung zwischen Laut und Schrift, die besser gelingt, je differenzierter phonologische Aspekte repräsentiert sind. Sowohl für den Kindergarten- als auch für den Schulbereich werden deswegen dringend bessere Informationen über Vorkommen und Häufigkeit verschiedener morphologischer und phonologischer Verarbeitungseinheiten wie Morpheme, Silben und Phoneme benötigt, um die Schwierigkeit von Test-, Trainings- und Unterrichtsmaterialien möglichst optimal zu gestalten.

Leider liegen solche Informationen im Deutschen selbst für den Erwachsenenbereich bislang nicht oder nur unzureichend vor. Selbst die DGD-Korpora (Datenbank für gesprochenes Deutsch) des IDS enthalten lediglich Wortform- und Lemma-Frequenzen, jedoch keine Transkription auf phonologischer Ebene. Die einzige Datenbank, die sowohl morphologische als auch phonologische Informationen zur Verfügung stellt, ist CELEX. Die dortigen Analysen sind jedoch aus

methodischen Gründen stark fehlerbehaftet und enthalten darüber hinaus kein kindersprachspezifisches Material.

Im Rahmen des dlexDB-Projektes (Heister et al., 2011) haben wir ein grundständiges und flexibel einsetzbares Analysetool erstellt, mit dem sich die Aussprache von Wörtern aus ihrer Schreibform rekonstruieren lässt (*gramophone*, vgl. Würzner, 2014). Da viele phonologische Regeln für den Bereich der Silben definiert sind (Wiese, 1996), ist es hierfür notwendig, den schriftsprachlichen Input zunächst verlässlich in Silben zu segmentieren. Zur Lösung dieses Problems ist es wiederum sinnvoll, die Wortformen zunächst morphologisch zu zergliedern, da im Deutschen alle starken Morphemgrenzen mit einer Silbengrenze zusammenfallen und unplausible Segmentierungen ausgeschlossen werden („ver-arbei-ten" vs. „ve-rar-bei-ten").

Ziel der hier berichteten Analysen ist es, die Qualität dieses Ansatzes zu berichten und die Verteilung verschiedener morphologischer und phonologischer Einheiten in childLex zu untersuchen. Dafür beschreiben wir zunächst die verwendeten Werkzeuge genauer und gehen auf die Datengrundlage der Analysen ein. Danach berichten wir erste Ergebnisse für drei verschiedene Analyseebenen: Morpheme, Silben und Phoneme sowie Zuordnungen von Graphemen und Phonemen, d. h. die Konsistenz der orthographischen und phonologischen Beschreibungsebene.

2 Methode

2.1 Korpus

childLex ist eine lexikalische Datenbank zur Schriftsprache, die von Kindern im Alter zwischen 6 und 12 Jahren gelesen wird (Schroeder et al., im Druck a). childLex basiert auf 500 Kinder- und Jugendbüchern, die auf der Grundlage aktueller Verkaufs- und Ausleihstatistiken von Online-Buchhändlern und öffentlichen Bibliotheken ausgewählt wurden. Insgesamt umfasst childLex ca. acht Millionen Wörter (Token), die sich auf ca. 180.000 verschiedene Wortformen (Types)

und ca. 120.000 Grundformen (Lemmata) verteilen. childLex ist für drei verschiedene Altersgruppen verfügbar (6–8, 9–10 und 11–12 Jahre) und kann unter www.childlex.de abgerufen werden. Die berichteten Normen umfassen lexikalische (Worthäufigkeit, Nachbarn, syntaktische Funktion), sublexikalische (Zeichen-Uni/Bi/Trigramme) und superlexikalische (Type-Bi/Trigramme) Merkmale, die bislang jedoch allein auf der orthographischen Ebene basieren.

2.2 Linguistische Analysen

Die Texte in childLex werden im Zuge der Aufbereitung für die Datenbank zunächst tokenisiert (i. e. Wort- und Satzgrenzenbestimmung) und morphologisch klassifiziert (i. e. Grundform- und Wortartenbestimmung). Die konkrete Analysekette ist ausführlich in Schroeder et al. (im Druck a) beschrieben. Im Folgenden beschreiben wir zwei weitergehende automatische Analyseschritte, die für die nachfolgenden Analysen zentral sind: die *morphologische Segmentierung* und die *phonologische Transkription*.

2.2.1 Morphologische Segmentierung

Als morphologische Segmentierung betrachten wir die Zerlegung eines Wortes in seine Morpheme. Wir unterscheiden dabei zwischen der Oberflächen- und der Tiefenzerlegung eines Wortes. Diese Unterscheidung lässt sich mit Hilfe des Wortes „Ärztekammern" anschaulich illustrieren: „Ärztekammern" besteht zunächst aus den Wörtern „Ärzte" und „Kammern". Diese sind jedoch wieder morphologisch komplex und bestehen jeweils aus einem Stamm und einer Endung („Arzt" + „e" und „Kammer" + „n"). Im Falle des Erstglieds findet jedoch zusätzlich zur Suffigierung auch eine Umlautung des Stammvokals statt: Aus „Arzt" wird „Ärzt". Beide werden *Allomorphe* des lexikalischen Morphems {Arzt} genannt. Als Oberflächenzerlegung bezeichnen wir nun die Zerlegung eines Wortes in seine tatsächlich beobachteten Bestandteile, also im Beispiel „Ärzt" + „e" +

„kammer" + „n". Als Tiefenzerlegung bezeichnen wir die Zerlegung eines Wortes in seine lexikalischen Komponenten, also im Beispiel {Arzt} und {Kammer}. Neben der Lokalisierung von Morphemgrenzen wird in diesem Analyseschritt auch die Art der beteiligten Wortbildungsprozesse bestimmt. Wir unterscheiden hier zwischen *Präfigierung*, *Suffigierung* und *Komposition*.

Die Oberflächenzerlegung dient der Verbesserung der nachfolgenden phonologischen Transkription (Abschnitt 2.2.2). Auf Basis der Tiefenzerlegung analysieren wir das Morpheminventar in childLex (Abschnitt 3.1).

Die allgemein verfügbaren Werkzeuge zur automatischen, morphologischen Analyse (z. B. SMOR: Schmid, Fitschen & Heid, 2004 oder TAGH: Geyken & Hanneforth, 2006) kommen für die beschriebene Art der Zerlegung leider nicht in Frage, da sie viele morphologisch komplexe Wörter wegen ihrer Häufigkeit bzw. aus Gründen semantischer „Intransparenz" nicht segmentieren (in SMOR z. B. „Achtung", in TAGH z. B. „Schneemann"). Wir bedienen uns daher zur Zerlegung eines überwachten statistischen Lernverfahrens. Das zugrunde liegende Modell, ein *Conditional Random Field* (CRF: Lafferty, McCallum & Pereira, 2001), wurde anhand von 15.000 manuell segmentierten Wortformen trainiert und für alle Zeichenbigramme innerhalb eines Wortes bestimmt, ob sich zwischen ihnen eine Morphemgrenze befindet (vgl. Klenk & Langer, 1989). Wir illustrieren dieses Vorgehen in Abbildung 1. Die Methode erreicht eine Präzision von 93,8 % bei einem Recall von 96,8 % (10-fache Kreuzvalidierung). Zur Implementierung des CRF benutzen wir CRF++ (Kudo, 2005).

```
Eingabewort: Abbauvereinbarung
Bigramrepräsentation und Morphemgrenzen:
_A   0
Ab   0
bb   +
ba   0
au   0
uv   #
ve   0
er   0
re   +
ei   0
in   0
nb   +
ba   0
ar   0
ru   ~
un   0
ng   0
Zerlegung:
Ab+bau#ver+ein+bar~ung
```

```
Erläuterung:
+ Präfix
# Kompositionsgrenze
~ Suffixgrenze
```

Abbildung 1. Zerlegung des Wortes „Abbauvereinbarung" in Zeichenbigramme und vom CRF geschätzte Morphemgrenzen

2.2.2 Phonologische Transkription

Mit phonologischer Transkription bezeichnen wir die lautgetreue Darstellung eines Wortes im internationalen phonetischen Alphabet (IPA). Zusätzlich zu den lautlichen Entsprechungen der Graphemebene kodieren wir auch Silbengrenzen[1].

Die Generierung der Transkriptionen erfolgt vollautomatisch auf Basis der vorangegangenen morphologischen Segmentierung mit

[1] Hier besteht ein wesentlicher Unterschied zu Celex, wo Silbenrepräsentationen auf graphematischer Ebene angeboten werden.

Hilfe von *gramophone* (Version 0.1; Würzner, 2014)[2]. *gramophone* hat zwei grundsätzliche Verarbeitungsphasen: In Phase 1 werden auf Basis manuell erstellter Ersetzungsregeln Kandidaten für die Aussprache einer Morphemsequenz generiert. Die Herausforderung bei der Modellierung liegt vor allem in der korrekten Realisierung der Vokalphonologie. In sequentiell abzuarbeitenden Schritten werden dabei zunächst Graphemsequenzen markiert, die Konsonanten enthalten, aber rein vokalisch ausgesprochen werden (können), wie im Falle von „onds" in „Fonds" [fõː] oder „ail" in „Detail" [deˈtaɪ]. Danach werden die Vokale transkribiert. Hierbei liefert die Graphemebene oft klare Hinweise durch explizite Längenmarkierung oder Doppelkonsonanten. Morphemgrenzen werden dabei beachtet, so dass „ll" in „Tallage" nicht als Doppelkonsonant behandelt wird. Im nächsten Schritt erfolgt die Markierung der Silbengrenzen. Wir folgen dabei im Wesentlichen einem Ansatz von Bouma (2003), der das für die Silbenstruktur verantwortliche *Maximum Onset Principle* (MOP) mit Hilfe linearer Ersetzungsregeln implementiert. Auch in diesem Fall liefert die Morphemstruktur wertvolle Anhaltspunkte, da Präfix- und Kompositionsgrenzen im Deutschen unabhängig vom MOP stets Silbengrenzen konstituieren (vgl. „ver-ir-ren" und „ve-ri-fi-zie-ren"). Zuletzt werden dann die Konsonanten transkribiert. Diese Reihenfolge hat unter anderem den Vorteil, dass ambisyllabische Konsonanten korrekt behandelt werden können: Im Wort „machen" liegt die Silbengrenze auf dem stimmlosen, uvularen Frikativ, so dass weder [maˈxən] noch [maxˈən], sondern [maχən] dessen korrekte Silbifizierung widerspiegelt. Viele der Ersetzungen in Phase 1 erfolgen op-

[2] Die hier beschriebene Prozedur spiegelt den Stand von November 2014 wider. Inzwischen liegt eine neue Version von *gramophone* vor, die einen grundsätzlich anderen methodischen Ansatz verfolgt. Dabei wird die Umschreibung der Graphemsequenzen in passende Phonemsequenzen auf Basis der Entscheidungen eines *Conditional Random Fields* vorgenommen (für Details vgl. Würzner & Jurish, eingereicht). Durch das geänderte Vorgehen ergeben sich u. U. andere phonologische Transkriptionen. Wir gehen jedoch davon aus, dass sich das berichtete Gesamtbild dadurch nicht grundsätzlich ändert.

tional, sodass an deren Ende eine Menge an möglichen Transkriptionen vorliegt. Die technische Umsetzung erfolgt mit Hilfe endlicher Automaten unter Verwendung von *foma* (Hulden, 2009).

In Phase 2 werden die Kandidaten unter Verwendung eines Sprachmodells bewertet. Die Bewertung erfolgt auf Basis von 5-Grammen über Paare von Graphem-Phonem-Sequenzen. Die Parameter des Modells werden auf Basis der 148.279 manuellen Transkriptionen des Deutschen Teils des Wiktionary-Projekts[3] geschätzt. Nach der Beseitigung von offensichtlichen Fehlern und der Korrektur von Inkonsistenzen enthält unser Trainingsmaterial 147.421 transkribierte Wörter. Um ein Sprachmodell über Graphem-Phonem-Abbildungen trainieren zu können, wurden die orthographische und die phonologische Repräsentation aligniert, d. h. aneinander ausgerichtet. Die Basis der Alignierung sind 541 manuell erstellte Graphem-Phonem-Abbildungen von Graphemsequenzen auf mögliche Phonemrepräsentationen. Enthalten sind beispielsweise die verschiedenen Möglichkeiten, „r" ([ɾ], [ʀ], [ʁ], [ɐ]) und „ch" ([ç], [k], [χ]) zu realisieren. Die Abbildungen werden als endlicher Transduktor repräsentiert. Für die eigentliche Ausrichtung werden mit Hilfe dieses Transduktors alle theoretisch möglichen paarweisen Graphem-Phonem-Sequenzen für ein Wort generiert (darunter auch viele praktisch unmögliche) und danach die korrekte Sequenz herausgefiltert. Die nachfolgende Kompilierung eines geglätteten 5-Gramm-Modells wird beispielsweise in Hanneforth und Würzner (2009) ausführlich beschrieben und in unserem Fall mit Hilfe der OpenGRM-Bibliothek (Roark et al., 2012) umgesetzt.

gramophone generiert in Phase 1 für das Wort „Computer" beispielsweise 200 Aussprachevarianten. Diese große Zahl erklärt sich durch die hohe Varianz bei der Aussprache der Vokale „o" und „u" aber auch durch die vielen möglichen Realisierungen des Graphems

[3] http://de.wiktionary.org, Dump vom 7. April 2014. Für die Transkriptionsrichtlinien vgl. http://de.wiktionary.org/wiki/Hilfe:IPA.

„c" (neben dem hier korrekten [k] sind das auch [tʃ] wie in „Cipollinosäulen", [ts] wie in „Vertices" und [s] wie in „Cineast"). In Phase 2 wird aus diesen Varianten dann die korrekte Transkription [kɔmpɻu:tɐ] ausgewählt.

2.3 Analysen

Um die Qualität der Analysen zu beurteilen und erste Ergebnisse zur Verteilung morphologischer und phonologischer Merkmale in childLex zu berichten, wurden handtranskribierte Wörter aus der deutschen Wiktionary-Datenbank herangezogen, die ca. 150.000 Wortformen umfasst. Von diesen kommen ca. 24.000 auch in childLex vor und werden dort ca. 720.000 Mal verwendet. Die Abdeckungsrate liegt damit sowohl auf Type- als auch auf Token-Ebene bei ca. 15 %. Die folgenden Analysen beziehen sich auf diese Schnittmenge, wobei alle Formen zunächst in Kleinbuchstaben konvertiert wurden. Um die Auftretenswahrscheinlichkeit in childLex zu bestimmen, wurde auf die korrespondierenden Type- bzw. Lemma-Frequenzen des Gesamtkorpus zurückgegriffen.

3 Ergebnisse

3.1 Morpheme

Alle Komposita wurden zunächst auf orthographischer Ebene in ihre verschiedenen Teilwörter zerlegt („lesebuch" = „lese" + „buch"). In einem nachfolgenden Schritt wurden die Teilwörter weiter dekomponiert in die ihnen zugrunde liegenden Stämme, Präfixe und Suffixe/Flexionsendungen („vorlesen" = „vor" [Präfix] + „les" [Stamm] + „en" [Flexion]). Eine Zusammenfassung dieser Analyse findet sich in Tabelle 1.

Von den insgesamt ca. 24.000 Wortformen sind nur 9,6 % Simplizia, d. h. bestehen aus lediglich einer morphologischen Komponente. Über 90 % aller Types sind hingegen morphologisch komplex, wobei Drei-Morphem-Kombinationen („zahn+bürst+en") mit ca.

40 % besonders häufig vorkommen, Zwei- und Vier-Morphem-Kombinationen sind jedoch ebenfalls sehr zahlreich (ca. 30 % und 17 %). Das durchschnittliche Wort hat 2,8 (SD = 1,0) Morpheme, wobei eine Kombination von Komposition und Flexionsendung besonders typisch ist.

Tabelle 1
Anzahl und Häufigkeit von Teilwörtern, Stämmen, Präfixen und Suffixen/ Flexionsmarkern in childLex

Einheit	Typ	Anzahl	*F* = 1	*F* < 5	Rang 1–10	Rang 1–100
Wörter	Types	23.935	28,3 %	55,1 %	< 0,1 %	0,4 %
	Token	721.876	0,9 %	3,4 %	10,9 %	30,7 %
Teilwörter	Types	21.992	24,5 %	49,9 %	< 0,1 %	0,5 %
	Token	803.710	0,7 %	2,6 %	9,8 %	28,9 %
Stämme	Types	8.231	11,4 %	26,1 %	0,1 %	1,2 %
	Token	593.261	<0,1 %	0,1 %	23,8 %	51,2 %
Präfixe	Types	236	18,6 %	39,8 %	4,2 %	42,4 %
	Token	183.805	<0,1 %	0,1 %	87,0 %	99,7 %
Suffixe/ Flexion	Types	313	20,1 %	42,5 %	3,2 %	32,0 %
	Token	593.261	<0,1 %	<0,1 %	89,2 %	99,8 %

In den Wörtern werden insgesamt ca. 22.000 verschiedene Teilwörter verwendet. Von diesen kommen ca. 25 % nur ein einziges Mal in childLex vor (d. h. $F = 1$), stellen jedoch nur 1 % aller Token. Die zehn häufigsten Teilwörter decken ca. 10 % aller Token ab, die 100 häufigsten Teilwörter 29 %. Ein Vergleich der Gesamt- und Teilwortformen zeigt, dass ihre Verteilung sehr ähnlich ist, d. h. die Verteilungseigenschaften von Komposita und ihren Teilkonstituenten unterscheidet sich nicht grundlegend.

In den Teilwörtern werden nur ca. 8.000 verschiedene Stämme verwendet. Von den Stämmen kommen ca. 11 % nur einmal in childLex vor, was weniger als 0,1 % aller Token entspricht. Die zehn häufigsten Stämme sind (Hilfs)Verben („haben", „wollen", „werden", „können", „stehen", „gehen", „fragen") oder Adjektive („klein",

„gut") und decken 24 % aller Token ab, die 100 häufigsten Stämme sogar über 50 %. Das heißt, auch wenn die meisten Wörter in der Stichprobe morphologisch komplex sind, sind Einfachformen auf der Token-Ebene der Regelfall.

In den Teilwörtern werden 236 verschiedene Präfixe verwendet. 19 % kamen dabei lediglich einmal in childLex vor. Dieser Anteil ist auf Segmentierungsfehler zurückzuführen. Die zehn häufigsten Präfixe („ge", „be", „ver", „ab", „an", „er", „aus", „ein", „auf", „un") decken ungefähr 80 % aller Token ab, die hundert häufigsten Präfixe beschreiben die Verwendung in childLex fast vollständig.

In den Teilwörtern finden sich 313 verschiedene Suffixe und Flexionsendungen, wobei ca. 20 % nur einmal verwendet werden und ebenfalls Segmentierungsfehler sind. Die zehn häufigsten Formen bestehen hauptsächlich aus Flexionsendungen („en", „e", „t", „te", „er", „n", „s" „et") sowie frequenten Suffixen („ig", „ung") und stellen ca. 80 % aller Token. Die 100 häufigsten Suffixe und Flexionsendungen decken über 99 % aller Token ab. Insgesamt ist die Verteilung bei den Präfixen und Suffixen/Flexionsendungen sehr ähnlich.

Zusammenfassend ist festzustellen, dass die Qualität der morphologischen Analyse bereits gut, aber durchaus noch verbesserungsfähig ist, insbesondere im Prä- und Suffixbereich. Insgesamt werden jedoch ca. 80 bis 90 % aller Morpheme korrekt erkannt und Fehlklassifikationen sind durch ihr seltenes Vorkommen leicht zu identifizieren.

Bei den betrachteten Einheiten scheint eine Zweiteilung zu bestehen: Auf der einen Seite stehen Wörter und Komposita-Teilwörter, die sich sehr ähnlich verteilen. Auf der anderen Seite stehen Flexions- und Derivationsendungen, die ebenfalls sehr ähnlich verteilt sind und als geschlossene Klasse unterschiedlichen Gesetzmäßigkeiten zu unterliegen scheinen. Eine interessante Zwischenstellung nehmen die Stämme ein, die in gewisser Weise zwischen den beiden anderen Elementen liegen: Ihr Einsatz ist nicht ganz so variabel wie der von

lexikalischen Vollformen, aber auch nicht so regelmäßig wie die Derivations- und Flexionsmorphologie. Sie nehmen damit eine mittlere Abstraktionsebene ein, die ohne eine morphologische Analyse nicht möglich wäre.

3.2 Silben

In einem nächsten Schritt wurden die Wörter auf der Grundlage der morphologischen Analyse in Silben und Phoneme zerlegt. Tabelle 2 zeigt die Ergebnisse getrennt für die Silben-, Phonem- und Graphemebene.

Insgesamt gibt es ca. 7.000 unterschiedliche Silben. Ungefähr 13 % werden nur ein einziges Mal in childLex verwendet, was weniger als 0,1 % aller Token entspricht. Die zehn häufigsten Silben ([tə], [gə], [bə], [tn̩], [nn̩], [gn̩], [ɛe̯], [baɪ̯], [fɛe̯], [te]) decken ca. 20 % aller Vorkommen ab, die zwanzig häufigsten 26 %. Insgesamt zeigen die Silben damit ein ähnliches Verteilungsmuster wie die Stämme, und in der Tat fallen viele Silben und Stämme ja zusammen („Nuss", „Baum" etc.). Die häufigsten Silben basieren auf Flexionsmorphemen, umfassen jedoch Teile des Stamms (z. B. [tn̩] in [fʏʁç.tn̩]). Das ist bedeutsam, da in diesem Fall die Silbe mit keinem der beiden Teilmorpheme zusammenfällt, d. h. an dieser Stelle legen Silben- und Morphemebene unterschiedliche Segmentierungen nahe (Silben: „fürchten" = „fürch+ten" vs. Morpheme: „fürcht+en").

Teilt man die Silben anhand ihres Vokal-Konsonanten-Musters in verschiedene Typen ein, so zeigte sich eine große Variationsbreite (167 verschiedene Silbenarten). Die zehn häufigsten („CVC", „CV", „CVV", „CVCC", „VC", „CVVC", „CCVC", „CCVV", „VVC", „VV") decken ca. 89 % aller Vorkommen ab. Die zwanzig häufigsten Silbentypen sogar 97 %.

Insgesamt bestätigen die Ergebnisse, dass das Silbeninventar des Deutschen ausgesprochen umfangreich ist, was an der im Vergleich zu anderen Sprachen reichhaltigeren Silbenstruktur mit komplexem

An- und Ablaut liegt. Die Silbe ist dabei eine interessante Gliederungseinheit, die ungefähr zwischen dem Stamm und den Derivations- und Flexionsmorphemen anzusiedeln ist und Aspekte beider Ebenen miteinander vereinigt.

Tabelle 2
Anzahl und Häufigkeit von Silben, Phonemen und Graphemen in childLex

Einheit	Typ	Anzahl	*F* = 1	*F* < 5	Rang 1–10	Rang 1–20
Silben	Types	7.046	12,6 %	28,8 %	0,3 %	1,4 %
	Token	1.689.085	<0,1 %	0,2 %	19,1 %	26,3 %
Phoneme	Types	80	0,0 %	0,0 %	12,5 %	25,0 %
	Token	4.225.138	0,0 %	0,0 %	49,9 %	74,3 %
Grapheme	Types	124	4,0 %	8,9 %	8,1 %	16,1 %
	Token	4.225.138	<0,1 %	<0,1 %	55,5 %	77,6 %

3.3 Phoneme und Grapheme

Tabelle 2 zeigt ebenfalls das Phonem- und Graphem-Inventar, das von *gramophone* angenommen wird. Dabei ist wichtig festzustellen, dass es sich dabei nicht um Grapheme und Phoneme im eigentlichen Sinne handelt, sondern um orthographische und phonologische Einheiten-Cluster, die konsistent aufeinander abgebildet werden. Auf orthographischer Ebene sind dabei häufig Buchstaben-Verbindungen involviert, die ein bis sechs Buchstaben umfassen (z. B. „ailles" in „Versailles", das auf die Phonemkombination [aɪ] abgebildet wird). Das ist auch gleichzeitig ein Beispiel dafür, dass auch auf phonologischer Seite häufig Mehrfachsegmente involviert sind, die ein bis drei Phoneme umfassen (z. B. „au" in „Zwickau" auf die Phonemkombination [ʔaʊ]). Allgemein besteht also eine n:m Beziehung zwischen Buchstaben und Phonemen. Viele dieser komplexen Beziehungen werden jedoch nur gebraucht, um fremdsprachliches Material oder Lehnwörter korrekt zu kodieren.

Auf der phonologischen Seite werden dabei insgesamt 80 Phonemkombinationen benötigt. Keine der Phonemkombinationen

kommt dabei nur ein einziges Mal in childLex vor, allerdings ergeben sich große Unterschiede in der Vorkommenshäufigkeit. Die zehn häufigsten Phoneme ([t], [ə], [n̩], [n], [l], [a], [ʁ], [b], [g], [ɛ]) decken fast 50 % aller Vorkommen in childLex ab, die zwanzig häufigsten sogar 75 %.

Auf der Graphem-Seite wird zur Generierung der Aussprache ein etwas größeres Inventar von 124 Buchstabenkombinationen benötigt, was bestätigt, dass das Deutsche eine höhere Vorwärts- als Rückwärts-Konsistenz aufweist. Einige Grapheme (ca. 10 %) kommen jedoch nur selten in childLex vor und werden für die Aussprache von Fremdwörtern benötigt. Die zehn häufigsten Grapheme („e", „t", „r", „en", „n", „a", „g", „b", „s", „i") entsprechen fast vollständig den häufigsten Phonemen und decken zusammen 56 % aller Vorkommen ab. Lediglich das komplexe Graphem „en", das für die Aussprache des silbischen [n̩] bei Schwa-Deletion benötigt wird, ist komplex. Auch unter den zehn nächsthäufigen Graphemen befinden sich lediglich drei komplexe Segmente („ei", „er" und „ch").

Zur Abbildung der Grapheme auf die Phoneme werden dabei insgesamt 212 der in Abschnitt 2.2.2 eingeführten Zuordnungskombinationen benötigt. Nicht alle kommen jedoch gleich häufig vor und sie unterscheiden sich stark in ihrer Konsistenz. Das Phonem mit den meisten zugeordneten Graphemen ist das [s] (stimmloses s), für das es neun verschiedene Verschriftlichungen gibt („c", „ce", „es", „s", „se", „ß", „ss", „sse", „ts"). Allerdings sind nur vier von ihnen („s", „se", „ß" und „ss") hinreichend frequent, um für den Schriftspracherwerb bedeutsam zu sein. Die mit Abstand häufigste Verschriftlichung erfolgt dabei durch das einfache „s". Überraschenderweise ist es nicht so, dass es für Vokale, denen im Deutschen eine größere Variabilität zugeschrieben wird, mehr Verschriftlichungsvarianten gibt (Tab. 3). Das lange o ([oː]) weist mit sechs Graphemen die meisten Verschriftlichungen auf („au", „eau", „o", „oa", „oh", „oo"), von denen jedoch wiederum nur zwei („o" und „oh") hinreichend frequent

sind. Selbst das Graphem „oo" liegt hinter dem fremdsprachlichen „au" zurück.

Tabelle 3
Verschriftlichungsvarianten und -häufigkeit des Phonems [oː] in childLex

Phonem	Graphem	Häufigkeit
oː	au	2612
	eau	1
	o	19712
	oa	107
	oh	3551
	oo	994

Umgekehrt ist jedoch das Phonem [oː] die häufigste Aussprache für das Graphem „oo" (für das es insgesamt drei Aussprachen gibt), während für „au" (für das es vier Aussprachevarianten gibt) der Diphthong [aʊ] die präferierte Aussprache ist (Tab. 4).

Tabelle 4
Aussprachevarianten und -häufigkeit der Grapheme „oo" und „au" in childLex

Graphem	Phonem	Häufigkeit
oo	oː	994
	ʊ	24
	uː	207
au	aʊ	45582
	ɔ	24
	oː	2612
	ʔaʊ	2111

4 Diskussion

Wir haben einen kurzen Überblick über den aktuellen Stand der sublexikalischen Repräsentationen in childLex gegeben und dazu einerseits die Transkriptionsprozedur *gramophone* und anderseits die resultierenden Korpusstatistiken besprochen. Die ersten Ergebnisse sind vielversprechend und spiegeln an den meisten Stellen sprachwissenschaftliche Erkenntnisse wider. Ein besonderes Augenmerk

muss der Verbesserung der morphologischen Segmentierung gelten, da dort gemachte Fehler direkte Auswirkungen auf alle folgenden Analyseschritte haben. Es ist geplant, die beschriebene Architektur auch auf die anderen Korpusgrundlagen in dlexDB anzuwenden, um so vergleichende Analysen zwischen Kinder- und Erwachsenensprache zu ermöglichen.

5 Literatur

Baayen, R. H., Piepenbrock, R. & Guilkers, L. (1996). *CELEX2 [CD-ROM]*. Philadelphia, PA: Linguistic Data Consortium.

Bouma, G. (2003). Finite-state methods for hyphenation. *Natural Language Engineering*, *9*, 5–20.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J. & Böhl, A. (2011). The word frequency effect: A review of recent developments and implications for the choice of frequency estimates in German. *Experimental Psychology*, *58*, 412–424.

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.

Hanneforth, T. & Würzner, K.-M. (2009). Statistical language models within the algebra of weighted rational languages. *Acta Cybernetica*, 19, 313–356.

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

Hulden, M. (2009). Forma: A finite-state compiler and library. In *Proceedings of the EACL 2009 Demonstration Session*, 29–32.

Klenk, U. & Langer, H. (1989). Morphological segmentation without a lexicon. *Literary and Linguistic Computing*, 4, 247–253.

Kudo, T. (2005). *CRF++: Yet another CRF tool kit*. Zugriff am 16.03.2015: http://taku910.github.io/crfpp

Lafferty, J., McCallum, A. & Pereira, F. C. N. (2001). Conditional random fields: Probabilistic models for segmenting and labeling data. In *Proceedings of the 18th International Conference on Machine Learning*, 282–289.

Roark, B., Sproat, R., Allauzen, C., Riley, M., Sorensen, J. & Tai, T. (2012). The OpenGrm open-source finite-state grammar software libraries. In *Proceedings of the ACL 2012 System Demonstrations*, 61–66.

Schmid, H., Fitschen, A. & Heid, U. (2004). SMOR: A German Computational Morphology Covering Derivation, Composition, and Inflection. In *Proceedings of the IVth International Conference on Language Resources and Evaluation*, 1263–1266.

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A. & Kliegl, R. (im Druck a). childLex: A lexical database of German read by children. *Behavior Research Methods.*

Schroeder, S., Würzner, K.-M., Heister, J., Geyken, A. & Kliegl, R. (im Druck b). childLex: Eine lexikalische Datenbank zur Schriftsprache von Kindern im Deutschen. *Psychologische Rundschau.*

Szagun, G. (2013). *Spracherwerb beim Kind: Ein Lehrbuch* (5., vollst. überarb. Aufl.). Weinheim: Beltz.

Wiese, R. (1996). *The Phonology of German.* Oxford University Press.

Würzner, K.-M. (2014). *gramophone – A finite-state letter-to-sound system for German*. Talk given at the 7th International Workshop on Weighted Automata: Theory and Applications.

Würzner, K.-M., Heister, J. & Schroeder, S. (2014). Altersgruppeneffekte in childLex. In A. Adelt, T. Fritzsche, J. Roß & S. Düsterhöft (Hrsg.), *Spektrum Patholinguistik Band 7* (123–131). Universitätsverlag Potsdam.

Würzner, K.-M. & Jurish, B. (eingereicht). *A hybrid approach to grapheme-phonem conversion.*

Kontakt

Kay-Michael Würzner
wuerzner@bbaw.de