University of Zurich[UZH]

Institute of Computational Linguistics

ACQDIVIZ

Visualizing Development In Longitudinal First Language Acquisition Data

Master's thesis presented to the Faculty of Arts of the University of Zurich
for the degree of Master of Arts UZH

Examiner: Prof. Dr. Sabine Stoll
Supervisor: Dr. Noah Bubenhofer

Author:
Danica Pajović
Student-ID: 07-713-548
Limmattalstrasse 242
8049 Zürich

22nd March, 2016

Abstract

The increasing availability of various software tools to collect and process textual data has simplified the work to build corpora enriched with linguistic and meta-linguistic information to a great extent. Such corpora facilitate the detection of correlations between linguistic and extra-linguistic factors which can be of interest in a multitude of research areas such as sociology, political science, discourse analysis, historical linguistics and also child language acquisition. Despite the many benefits the inclusion of more and more data brings to linguistic research, it also introduces new challanges with regards to focusing on important parts of the data. Conclusions are mostly drawn based on results and tables generated from statistical calculations, but visualizations can serve as a tool to analyse the underlying data in a visually explorative way. In this thesis, I will present the work that has been carried out in two research projects at the University of Zurich. I will, on the one hand, describe the work that has been done to build a database of longitudinal child language acquisition corpora from nine typologically maximally diverse languages and, on the other hand, I will focus on describing and using data visualizations that are specifically aimed at studying child language acquisition. I will present and discuss challenges in building visualizations for language acquisition data and I will also combine visualization techniques with a quite novel approach in child language acquisition research, namely network theory.

This work has shown that the challenges for visualizing language acquisition data lay in particular in the multi-leveled information of the data, as well as in its temporal nature. Analyses from network theory on languages from the ACQDIV project have shown that small world and scale-free properties in networks generated from lexical co-occurrences can be found cross-linguistically, with only minor differences due to typological characteristics of the analysed languages.

[1]The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 615988.

Zusammenfassung

Die zunehmende Verfügbarkeit von verschiedenen Software-Tools zur Sammlung und Verarbeitung von Textdaten hat den Arbeitsaufwand zur Erstellung von linguistischen Korpora massiv erleichtert. Solche Korpora vereinfachen das Erkennen von Korrelationen zwischen linguistischen und extra-linguistischen Faktoren, die in verschiedenen Forschungsdisziplinen von Interesse sein können, wie zum Beispiel in der Soziologie, der Politikwissenschaft, der Diskursanalyse, der historischen Linguistik und auch in der Spracherwerbsforschung. Trotz der vielen Vorteile, die das Einbinden von immer mehr Daten in linguistische Forschungsprojekte bringt, liefern mehr Daten auch neue Herausforderungen im Bezug auf das Herausfiltern von Informationen, die für eine bestimmte Fragestellung von Interesse sind. Ohne sich dabei nur auf Tabellen von statistischen Kalkulationen zu verlassen, können Datenvisualisierungen dabei helfen, einen datengeleiteten Ansatz zu verfolgen, indem die Visualisierung als Mittel zur explorativen Datenanalyse genutzt werden kann. In dieser Arbeit werde ich zwei Forschungsprojekte näher vorstellen, in deren Rahmen meine Masterarbeit entstanden ist. Ich werde einerseits ein Projekt vorstellen, in dem eine Datenbank von longitudinalen Erstspracherwerbsdaten für neun typologisch maximal verschiedene Sprachen erstellt wurde, die es nun ermöglicht, Fragestellungen aus der Erstpracherwerbsforschung mit einem sprachvergleichenden Ansatz nachzugehen. Des weiteren werde ich theoretische (aber auch praktische) Fragestellungen, die im Projekt *Visual Linguistics* herausgearbeitet wurden, an Erstspracherwerbsdaten ausführen. Ausserdem werde ich in dieser Arbeit auch versuchen aufzuzeigen, wie Fragestellungen der Netzwerktheorie und der Datenvisualisierung genutzt werden können, um sprachübergreifend die Entwicklung von lexikalischen Netzwerken zu untersuchen.

Diese Arbeit hat gezeigt, dass die hauptsächlichen Herausforderungen in der Visualisierung von Erstspracherwerbsdaten in der multidimensionalen, multimedialen und auch temporalen Natur von linguistischen Daten liegen. Des weiteren hat diese Arbeit aufgezeigt, dass sogenannte “Kleine-Welt Phänomene” und “skalenfreie” Eigenschaften aus der Netzwerktheorie auch in lexikalischen Netzwerken vorherrschen, mit nur geringen Unterschieden, die auf sprachtypologische Unterschiede zurückzuführen sind.

Acknowledgement

I want to express my gratitude to my supervisors Prof. Dr. Sabine Stoll and Dr. Noah Bubenhofer, not only for giving me the chance to work on two very interesting research projects during my time as a student at the University of Zurich, but also for their support and encouragement to combine two research fields, whose interaction and potential mutual benefit have not been studied in detail yet – namely *Linguistic Data Visualization* and *Child Language Acquisition*. I would also like to thank Klaus Rothenhäusler, Katrin Affolter, Jekaterina Mažara, Robert Schikowski and Cazim Hysi for their amicable and productive team-work throughout the past year. Furthermore, I also want to thank Mirjam Zumstein for many unforgettable lectures, coffee breaks and Skype calls, as well as Florence Favre, Aysha Tresh, Sabine Dippon, Gertrud Stefanini-Götte, Raffael Aregger, Franziska Maag, Darina Frangi, Patricia Valdivieso Ávila (and Carlos), Martina Wettstein, Adamantia Karali, Paula Carvalho, Kate Barber, Sofina Begum Malik, Emily Peach, Henry and Gaby Baltes and the family Gasser for their invaluable moral support.

Above all, my sincerest gratitude goes to Dr. Steven Moran for his technical mentorship, as well as his willingness to help me with every problem I encountered, and to Olga Ignateva, my sister Ana Pajović and my parents for their patience, unconditional love and support. Thank You.

Contents

Abstract i

Acknowledgement iii

Contents iv

List of Figures vii

List of Tables x

List of Acronyms xi

1 Introduction 1

1.1 Motivation 1
1.2 Research Questions 2
1.3 Thesis Structure 4

2 The ACQDIV Project 5

2.1 CHILDES 6
2.1.1 Coding Formats: CHAT, XML, Toolbox 7
2.2 General Project Description 11
2.3 Selection of Corpora 11
2.4 ETL Pipeline 13

3 Visualizing Data 16

3.1 Historical Development in Data Visualization 16
3.2 Theoretical Concepts 24
3.2.1 Human Cognitive Perception 26
3.2.2 Gestalt Theory 27
3.2.3 Visual Information-Seeking Mantra(s) 28
3.2.4 Diagrammatik 30
3.3 Linguistic Data Visualization 31
3.4 Visualization of Linguistic Development in Child Language Acquisition 35

4 ACQDIVZ I – Visualizing Development In Russian Verbal Morphology 37
4.1 Theoretical Background . . . 37
4.1.1 Grammatical and Lexical Aspect . . . 37
4.1.2 Aspect- and Distributional Bias Hypothesis . . . 39
4.2 Data Set . . . 42
4.3 Morphological Tree Diagram . . . 44
4.4 Morphological Sunburst Visualization . . . 46
4.5 Pie-Chart Bubble-Plots . . . 53

5 ACQDIVZ II – Language Acquisition & Network Theory 56
5.1 Historical Development . . . 58
5.2 Theoretical Background . . . 61
5.2.1 Graph Properties . . . 62
5.2.1.1 Graph Directedness . . . 62
5.2.1.2 Node Degree (<k>) . . . 63
5.2.1.3 Degree Distribution (P(k)) . . . 63
5.2.1.4 Connected Components (NCC) . . . 64
5.2.1.5 Clustering Coefficient (CC) . . . 64
5.2.1.6 Average Path Length (L) . . . 65
5.2.2 Graph Types . . . 66
5.3 Research Findings: Overview . . . 68
5.3.1 Semantic Networks . . . 69
5.3.2 Syntactic Networks . . . 70
5.3.3 Phonological Networks . . . 71
5.3.4 Lexical Networks . . . 73
5.4 ACQDIV & Lexical Co-Occurrence Networks . . . 76
5.4.1 Data . . . 77
5.4.1.1 Global Networks . . . 77
5.4.1.2 Local Networks . . . 78
5.4.2 Methodology . . . 79
5.4.2.1 Global Networks . . . 79
5.4.2.2 Local Networks . . . 81
5.4.3 Analysis & Results . . . 82
5.4.3.1 Global Networks . . . 83
5.4.3.2 Local Networks . . . 85
5.4.4 Discussion . . . 99

6 Conclusion 101

References 103

A Age Ranges Russian Children **111**

B Networks: Degree Distributions **112**

B.1 Degree Distributions CDS All Languages 112

B.1.1 Cummulative Degree Distribution of CDS Networks 112

B.1.2 Cummulative Degree Distribution of RU Networks (children) . . 113

B.1.3 Cummulative Degree Distribution of CTN Networks (children) . 113

C Networks: ACQDIVIZ Package **114**

List of Figures

2.1 CHAT format 8
2.2 TalkBank XML format 9
2.3 Toolbox format 10
2.4 Entity Relationship Diagram of the ACQDIV Database 15
3.1 Milestones in data visualization (source: Chen et al. [2007, 18]) . . . 17
3.2 Planetary Movement Over Time (source: Chen et al. [2007, 19]) . . . 18
3.3 Changes in Sunspots 18
3.4 Distances: Toledo to Rome (source: Tufte and Robins [1997, 15]) . . 19
3.5 Comparison of Population and Taxes (source: Chen et al. [2007, 24]) 20
3.6 Minard: Map of the fate of the French army during the Russian Campaign (1812-1813) (source: Kraak [2014, 17]) 21
3.7 F. Nightingale (1858): "Diagram of Causes of Mortality in the Army in the East" (source: Meirelles [2013, 95]) 22
3.8 Gestalt Principles (source:Taylor [2014, 11]) 28
3.9 Visualization Pipeline (Source:Keim et al. [2008, 156]) 30
3.10 Wordcloud Visualization 32
4.1 Screenshot ACQDIVIZ: Main page 41
4.2 Screenshot ACQDIVIZ: Visualizations 42
4.3 RUChild1: Tree-Diagram of Verbal Morphology 45
4.4 RUChild1: Tree-Diagram Zoom 45
4.5 Sunburst Visualization: Legend 47
4.6 Sunburst Visualization: Visual Clues 48
4.7 Sunburst Visualization: Hover and Zoom 48
4.8 Sunburst Small-Multiples: RUChild2 50
4.9 Sunburst Small-Multiples: RUChild2 adults 50
4.10 Sunburst Small-Multiples: RUChild5 51
4.11 Sunburst Small-Multiples: RUChild5 adults 52
4.12 VerbsPie-Scatter-Plot: Schema 53
4.13 VerbsPie: Legend 54
4.14 VerbsPie: Hover 54
4.15 VerbsPie: Summary Column 55
5.1 Euler 1936: 7 Bridges of Königsberg (source: Meirelles [2013, 49]) 57

5.2 Semantic network derived from dictionary entries (source: Mihalcea and Radev [2011, 2]) . . . 59
5.3 Visuwords Semantic Network . . . 59
5.4 Syntactic dependency graph (source: Mihalcea and Radev [2011, 142]) 60
5.5 Examples of Graph Structures (source: Mihalcea and Radev [2011, 12]) 62
5.6 Normal Degree Distrib. . . . 64
5.7 Power-Law Degree Distrib. . . . 64
5.8 Example graph for two sentences . . . 80
5.9 Example for dynamic .gexf file . . . 82
5.10 RU degree distribution . . . 84
5.11 RU random degr. distrib. . . . 84
5.12 RU fitted power-law distribution . . . 84
5.13 MLU for Russian children . . . 86
5.14 MLU for Russian Mothers . . . 86
5.15 N RU children . . . 87
5.16 <k>RU children . . . 87
5.17 N RU mothers . . . 87
5.18 <k>RU mothers . . . 87
5.19 Out-degree centralization for Russian children . . . 88
5.20 Out-degree centralization for Russian mothers . . . 88
5.21 Hubs development RU children . . . 90
5.22 Hubs development RU mmothers . . . 91
5.23 Hubs development stage 1 child RUChild2 . . . 92
5.24 Development hub *ne* child RUChild2 . . . 92
5.25 MLU for Chintang children . . . 93
5.26 MLU for Chintang adults . . . 93
5.27 N CTN children . . . 94
5.28 <k>CTN children . . . 94
5.29 N CTN kids . . . 94
5.30 <k>CTN adults . . . 94
5.31 Out-degree for Chintang children . . . 95
5.32 Out-degree for Chintang adults . . . 95
5.33 Hubs development CTN children . . . 97
5.34 Hubs development CTN adults . . . 98
5.35 Hubs stage 1 child LDCh1 . . . 99
5.36 Development hub *ba* LDCh1 . . . 99
B.1 Fitted power-law distribution for all languages . . . 112
B.2 Fitted power-law distribution for Russian children . . . 113
B.3 Fitted power-law distribution for Chintang children . . . 113

C.1 Selbstständigkeitserklärung 115

List of Tables

2.1 Language Sample 12
2.2 Corpora Contained in the ACQDIV database 13
3.1 Cognitive Benefits of Good Information Visualizations 26
3.2 Description of Gestalt Principles (Taylor [2014]) 28
4.1 Lexical Aspects According to Vendler (1957) 38
4.2 Basic Statistics: RU data 43
4.3 Distribution Aspect-Tense-Mode 52
5.1 Corpus Size 78
5.2 Russian Data 79
5.3 Chintang Data 79
5.4 Network parameters comparison 83
5.5 Networks Parameters Comparison (incl. Children) 85
A.1 Age Ranges Russian Children 111

List of Acronyms

ACL	Association for Computational Linguistics
CDS	Child-Directed Speech
CHILDES	Child Language Data Exchange System
MLU	Mean Length of Utterance
NLP	Natural Language Processing
POS	Part-Of-Speech
TEI	Text Encoding Initiative
UTF-8	Unicode Transformation Format (8-bit)
XML	eXtensible Markup Language

Language Codes (ISO 639-3):

ctn	Chintang
cre	Cree
chp	Denë
ind	Indonesian
ike	Inuktitut
jpn	Japanese
rus	Russian
sot	Sesotho
tur	Turkish
yua	Yucatec

1. Introduction

1.1. Motivation

The technical advantages that came along with the so-called "Digital Revolution" have lead to the development of new and powerful ways to process, manipulate and store data not only for business applications or research questions in the fields of finance, biology and physics (where the study of large datasets always has been part of the scientific research), but also in (traditionally) more humanist fields such as political science, sociology, as well as linguistics and literature. Especially the last two fields are nowadays studied under a research field commonly referred to as "Digital Humanities", where the development and application of large databases consisting of linguistically annotated data allow us to study linguistic phenomena on a much larger scale than it was possible before. However, the inclusion of more and more data also led to the need of additional tools to help find, extract and analyze the ever growing data set. A common way to show the results of conducted studies is to use various forms of visualizations to make the data "visible" in a form other than by only using letters, numbers and tables. Often, the data sets are so large, that tables alone cannot help to see the important data points at one glance. This is also a reason why, nowadays, we are used to being confronted with numerous forms of diagrams showing the correlations between different variables. But also in the field of data visualization, the state-of-the art techniques have become far more versatile and allow to take exploratory approaches to research questions where the visualizations themselves are not anymore exclusively used to *present* the findings of a study. Rather, they are used as research tools to help the researchers detect interesting patterns and/or outlayers within the data sets. The development of JavaScript libraries such as Mike Bostock's *Data Driven Documents* – `d3.js` (Bostock [2012]) or Lauren McCarthy's `p5.js` library (McCarthy [2015]), as well as the R (R Development Core Team [2008]) package *rCharts*[1] – just to name a few – have facilitated the development and use of interactive visualizations that can be used in web browsers, thus opening the field of data visualization not only to

[1] Which principally also uses the d3 JavaScript library in the background.

data scientists who want to visualize the results of their statistical analyses, but also to people coming from other fields such as data journalism, arts, sociology and information design. Furthermore, the growing number of digital humanists who are able to write code to process, analyze and visualize data, has led to the fact that visualization techniques are being applied and tested in a wide range of research domains. Two projects at the University of Zurich, which use corpus linguistic approaches, as well as approaches from data visualization are the following:

- The project ***Visual Linguistics***[2] funded by the Swiss National Science Foundation and led by Dr. Noah Bubenhofer at the Institute of Computational Linguistics. This project aims at, on the one hand, establishing a more fundamental theoretical framework for analyzing visualizations of underlying linguistic data and, on the other hand, at exploring various ways of applying visualization techniques to specific linguistic research questions (cf. Bubenhofer [2016]).
- The project ***ACQDIV – Acquisition processes in maximally diverse languages: min(d)ing the ambient language***[3] funded by the European Research Council and led by Prof. Dr. Sabine Stoll at the Institute of Comparative Linguistics. The goal of this project is to conduct child language acquisition research on nine typologically maximally diverse languages, in order to find cross-lingual developmental patterns in child language acquisition.

As both projects allowed me to gain insight into analyzing and building frameworks for linguistic analyses based on underlying multilingual datasets, I decided that this Master's thesis in Multilingual Text Analysis would be an ideal opportunity to combine insights from these two currently ongoing research projects. Therefore, the goal of this Master's thesis is to describe the process of building the ACQDIV database, as well as to use this database in order to follow a data-driven approach to specific research questions in child language acquisition by creating visualizations to display patterns in the development of verbal morphology in Russian child language, as well as the development of lexical networks in Russian and Chintang child language.

1.2. Research Questions

By combining insights from the above mentioned projects, the main research questions in this thesis will be the following:

[2]Project website: `http://www.cl.uzh.ch/de/research/visuallinguistics.html`

[3]Project website: `http://www.acqdiv.uzh.ch/en.html`

1. **What steps are necessary to build a database of child language data from typologically maximally diverse languages?**
   Most of the research in child language acquisition has focused on analyzing language development in Indo-European languages[4], which, considering the fact that there exist many more language families, does not capture the processes involved in child language acquisition while also considering typologically very different languages. The ACQDIV project seeks to bridge this gap between research in language acquisition and language typology by building a framework allowing researchers to raise questions that can be applied to study developmental patterns systematically and cross-linguistically. With regards to the steps necessary to build such a framework, I will describe the various data collection, extraction and processing tasks that were necessary to build a database that allows cross-linguistic comparison with regards to specific questions in child language acquisition research.

2. **What are the challenges in visualizing linguistic data and what already existing theoretical frameworks can be applied to analyze such visualizations?**
   There are many so-called "best practices" (Bubenhofer et al. [2016]) for creating "good" data visualizations. However, within the field of linguistics, very few attempts have been made so far in developing theoretical frameworks to analyze and discuss the usage and potential benefits of using data visualizations. With regards to the above research question, I will present historical developments, as well as current approaches in establishing a theoretical framework for analyzing linguistic data visualizations.

3. **To what extent can visualizations be helpful to approach specific research questions in child language acquisition?**
   Based on the theories that will be presented and discussed for developing and analyzing linguistic data visualizations, I will present two use-cases in child language acquisition research where a visual approach to specific research questions was taken in order to detect patterns when analyzing longitudinal language acquisition data. I will present and discuss factors where the visualizations helped to gain new insights from the underlying data, but also where expected benefits did not occur.

4. **How can we apply analysis techniques from network theory to look**

[4] A great amount of research has also been carried out for Chinese (cf. for example Zhang et al. [2008], Huang [2006]). For a cross-linguistic study on child language acquisition that also includes non-Indo-European languages, see Stoll et al. [2014] and Slobin [2014].

for cross-linguistic developmental patterns in longitudinal child language acquisition corpora? During the last two decades, network theory has seen a tremendous growth in different applications in various research fields. It has already been applied in various studies on child language acquisition, but never on such a broad cross-linguistic scale as applicable with the ACQDIV database. Therefore, we want to investigate how network theory can be applied to child language acquisition and language typology: will our corpora confirm already existing findings, or will they lead to new insights which can be explained due to the typological differences in our languages? We will test certain hypotheses on networks generated from lexical co-occurrences.

1.3. Thesis Structure

In the first part of this chapter I introduced my personal motivation for this thesis as well as the research questions that will be treated throughout this work. In **chapter 2**, I will describe the procedures required to build a multilingual database consisting of longitudinal language acquisition data of nine maximally diverse languages. In this part, I will treat in further detail the data formats which are used to store language acquisition data as well as the work that was done in order to build the ACQDIV database. **Chapter 3** introduces the theoretical background for analyzing visualizations of linguistic data. Here I will focus on work that has already been carried out to try to define parameters to analyze visualizations of linguistic data and I will also introduce the most commonly used guidelines in information visualization theory. Furthermore, in **chapter 3.3**, I will discuss the current theoretical background in linguistic data visualization before focusing on potential challenges in visualizing child language acquisition data. In **chapter 4 and 5**, I will apply the theoretical framework presented in chapter 3 and use the data presented in chapter 2 in two use-cases where we tried to visualize linguistic development based on two different research directions: The first use-case discussed in **4.1** will focus on the development of Russian verbal morphology in five target children and their adult peers. In the second use-case, discussed in **5**, we applied statistical analyses from network theory in order to measure lexical development in five Russian- and four Chintang-speaking children and their main caregivers. Additionally, I will show how the graph analysis toolkit *Gephi* (Bastian et al. [2009]) can be used to visually analyze lexical co-occurrence networks based on the calculation of various statistical parameters.

2. The ACQDIV Project

Despite the fact that there are approximately 7000 languages in this world, children can learn any language, and seem to do so quite effortlessly. Interest in child language acquisition can principally be traced back until Ancient Egypt times, where King Psammetichus ordered to raise two children in complete isolation, in order to see which language they would start to speak when growing up on their own. This language would then be the langauge of the "original" people, he thought (Hoff [2013, 10]). At the age of two, one child seemingly said a word which sounded like *becos*, the Phrygian word for 'bread'. Therefore, King Psammetichus concluded that the Phrygians were the oldest people with the oldest language (ibid.).
By the 18th century, the interest in using language development in children to account for proto-languages expanded to philosophical debates about the nature of humankind in general, where philosophers such as Descartes claimed that human nature was an innate characteristic. On the other hand, philosophers such as Locke argued that humans only become human as a result of social interaction (ibid.). As language was considered to be one of the main defining properties of humanity, an approximately 12 year old boy, who apparently lived his whole life in the French woods near Aveyron, and who was found in 1800, served as a new research object to study the origin of human language acquisition. Because the boy was capable of making sounds, but did not speak any (known) language, the people working with him favoured the idea that human language can only be learned within a human society. Also the fact that he never learned more than only a few words, made people suggest that there is a critical period in childhood for language acquisition, which, once passed, makes normal development in language acquisition nearly impossible (Hoff [2013, 11]).
With respect to the first *corpus linguistic* approaches to child language acquisition, they can be traced back to the end of the 19th and the beginning of the 20th century, where people started to simply observe how language emerges when children develop in a normal way. These so-called "baby-biographies" (Hoff [2013, 11], MacWhinney [2000a, 6]) contained transcripts of the language as well as the accompanied gesture of the children. Some of the first people to write such baby-biographies were Charles Darwin (who is better known for his theory of evolution), C. and W. Stern (Stern

and Stern [1907]), A. N. Gvozdev (Gvozdev [1949]) and Werner F. Leopold (Leopold [1939], Leopold [1947]), just to name a few. The limitations of writing down these diaries are rather obvious today: transcribing child-speech in real time is nearly impossible and many details are not captured at all. This changed radically in the late 1950s with the introduction of the tape recorder. Suddenly, people were able to capture human speech in real-time in all its facettes. However, transcripts of these newly generated corpora increased drastically in size, which prevented researchers from publishing their entire corpora (MacWhinney [2000a]). The research group around Prof. Roger Brown at Harvard University was the first one to start producing multiple copies of the transcripts of their language corpus of three children (Adam, Eve and Sarah) in 1962, enabling other researchers to also work with their data. Even though the first step in creating reproducible language acquisition data was done by making the data available to other researchers, an important part was still not made interoperable – the addition of metalinguistic and linguistic annotation to the corpus (ibid).

In the 1980s, the increasing availability of more powerful computers, able to process not only textual, but also audio-visual information in massive amounts of data, as well as more and more sophisticated coding schemes for transcribing and annotating speech data, have led to the need of standardized, interoperable data repositories which enable researchers to use various data sets in order to reproduce studies for different languages and/or use the same data set for different research questions. A first approach in creating a standardized framework for processing and storing child language acquisition data was the *Child Language Data Exchange System* (CHILDES) project (MacWhinney [2000b]). As the CHILDES project serves as precursor of the ACQDIV project, and because most of the corpora used in the ACQDIV database have been coded in formats developed or used within CHILDES, I will, in a first section, present the main milestones of CHILDES project, as well as briefly introduce the three data formats present in the corpora of the ACQDIV database.

2.1. CHILDES

As mentioned above, the CHILDES project was the first to provide tools for processing child language data, as well as to curate a publicly available repository of longitudinal child language corpora collected from multiple languages. The CHILDES database was established in 1984 by Brian MacWhinney and Catherine Snow and is now curated by Brian MacWhinney at the Carnegie Mellon University (MacWhin-

ney [2000a]). In a first stage, the corpora for this database were collected by using optical scanning and various computer programs to bring earlier corpora (with the earliest dating back to the 1960s) (ibid.) into a newly introduced coding standard called CHAT. After 1987, when other researchers also mainly produced digitized corpora, new methods to directly transform transcribed corpora into CHAT have been built (ibid). The CHILDES database now includes child language acquisition data in 35 languages (Stoll and Bickel [2013]) from 130 different corpora, and has recently also been incorporated into the TalkBank corpus, which also contains data from second language acquisition, language learning, conversation analysis and aphasics (MacWhinney [2000a]).

2.1.1. Coding Formats: CHAT, XML, Toolbox

As stated by (MacWhinney [2000a, 11]), by providing a computerized exchange system for transcripts of linguistic data the following goals were set by the CHILDES project:

- Automate the process of data analysis.
- Enhance the quality of the data by storing it in a consistent and fully-documented transcription system.
- "Provide more data for more children from more ages, speaking more languages" (ibid.).

Within the CHILDES project, three tools were developed in order to reach these goals:

- **The CHAT transcription and coding format**
  Which is used as a consistent annotation tool.
- **The CLAN analysis editor**
  Which is used to analyze data coded in the CHAT format.
- **The CHILDES database**
  Which contains language acquisition data in a standardized format.

An example excerpt of a file coded in **CHAT** and opened in the CLAN editor can be seen in Figure 2.1 (Moran et al. [to appear]). The first lines of each session file starting with an `@` contain metadata information for the session as well as the involved speakers. The first line in an utterance block is its transcription (on the same line as the speaker's label) and the following lines are the annotations belonging to the transcribed utterance (e.g. gloss and morpheme information in the `%tm` tier

and situation information in the %sit tier).

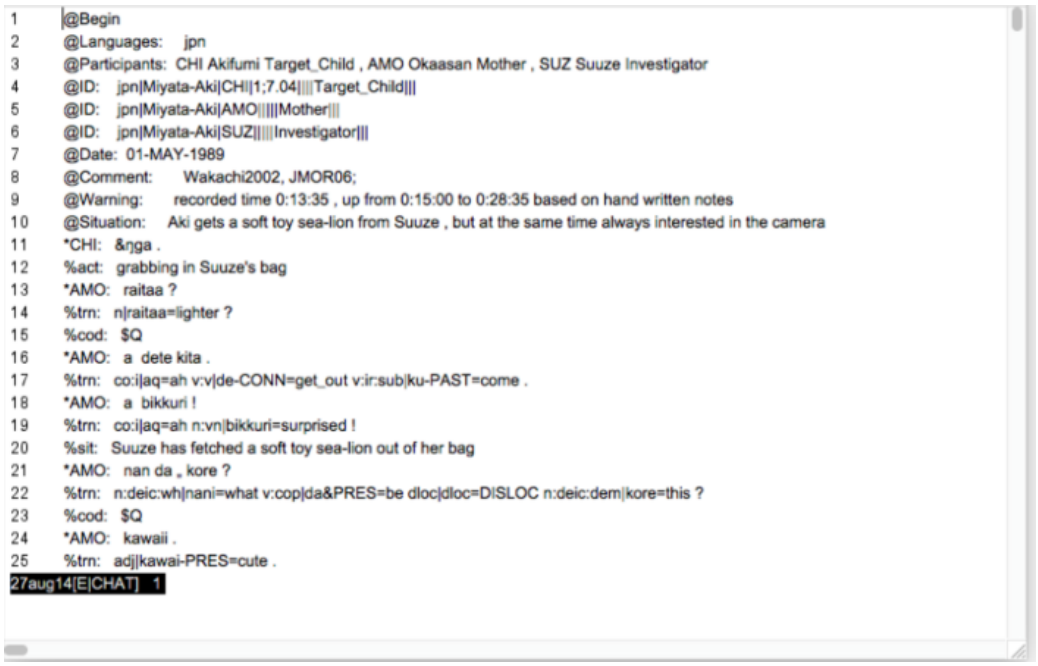

Figure 2.1.: CHAT format

Even though the CHAT format was developed with the intention to guarantee standardized compatibility across different corpora, it allows also for a great deal of encoding flexibility on part of the corpus compiler, which lead to challenges in unifying the CHAT corpora in our work (Moran et al. [to appear]).

Another file format that was used in some of the initial corpora is **TalkBank XML**. TalkBank XML is an XML closely associated with CHILDES and CHAT.[1] TalkBank XML has the same basic structure as CHAT, where data and metadata belonging to a session are coded in a single XML file, also with a head and body section. Nested XML tags are then used in the body section to further integrate utterance (marked by <u>) and word (marked by <w>) levels. Unlike in CHAT, where the line containing the utterance transcription is split up into words, in TalkBank XML all tiers other than <u> are grouped toghether in the <a> tag directly under the utterance level, and contain various attributes which mark the type of the tier. One of the main difficulties in processing TalkBank XML files in order to unify them with the other file formats was the very frequent annotation mismatches which appear when transforming former CHAT files to TalkBank XML (which was done in some of the initial corpora). Furthermore, only one of the TalkBank XML corpora used in the ACQDIV database had explicit XML coding for morphemes, in the other

[1]Further information about the TalkBank XML format can be found unter: http://talkbank.org/talkbank.xsd and https://talkbank.org/software/talkbank.xsd.

corpora the morphological information was coded less explicitly in often idiosyncratic formats, which again demanded coding inference strategies when parsing this format for the information we wanted to have in the ACQDIV database. Figure 2.2 shows an excerpt from a file coded in TalkBank XML.

```
<u who="MOT" uID="u1">
  <w>yoisho</w>
  <t type="p"></t>
  <media
    start="7.152"
    end="8.231"
    unit="s"
  />
  <a type="extension" flavor="trn">co:i|yoisho .</a>
  <a type="orthography">よいっしょ。</a>
</u>
<u who="XXX" uID="u2">
  <w>yoichotto<replacement><w>yoishotto</w></replacement></w>
  <t type="p"></t>
  <media
    start="9.857"
    end="11.335"
    unit="s"
  />
  <a type="extension" flavor="trn">co:i|yoishotto .</a>
  <a type="orthography">よいちょっと。</a>
</u>
```

Figure 2.2.: TalkBank XML format

A third file format which was used in the initial format in some of our corpora is SIL's **Toolbox** format.[2] Like in CHAT, corpora stored in the Toolbox format have sessions containing three central levels: *utterance, word* and *morpheme*. However, unlike in CHAT, the syntactic coding of this structure is much more implicit in Toolbox. For example, the syntactic unit which corresponds to the utterance level is the *record*. The way in which records are delimited also differs from the CHAT format and each record may have several tiers consisting of a so-called field marker, which starts with a backslash and indicates the type of content (e.g. `\ps` for parts of speech, `\gw` for word and `\eng` for the English translation). Furthermore, processing Toolbox files was especially complicated due to the fact that the association of annotations with the above mentioned three levels (utterance, word, morpheme) is not explicitly coded, but has to be inferred in various ways. Figure 2.3 shows a typical Toolbox file.

As already mentioned earlier in this chapter, the driving force behind the CHILDES project was to provide a framework for processing, analyzing, storing and sharing child language acquisition data. However, as we have seen in the introduction of the various coding formats, the processing and inclusion of cross-lingual child language

[2] Additional information about the format can be found under `http://www-01.sil.org/computing/toolbox/`

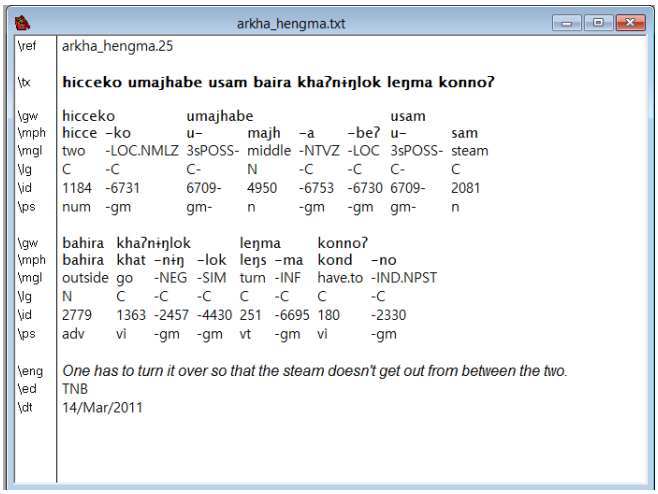

Figure 2.3.: Toolbox format

acquisition data remains problematic, even when the corpora are available in the same data format.[3] Also, in CHILDES, each corpus can only be analyzed individually with the CLAN tool, which in fact limits cross-linguistic research (Moran et al. [to appear]). What is more, most of the data in the CHILDES database is concentrated on either English or other Indo-European languages, which also led to the fact that research in child language acquisition has so far been heavily biased towards European language (which are typologically very similar to each other). Two other problematic factors with the data in the CHILDES database are that morphologically glossed corpora are available for only half of the languages, and only very few include translations which would allow for cross-linguistic comparisons (Stoll and Bickel [2013, 4]). Therefore, further concentrating on these languages does not help in answering the question of how children cope with linguistic diversity. Instead, research in child language acquisition has to broaden its linguistic horizon. This is where the ACQDIV project comes into play, by trying to bridge the gap between research in child language acquisition and language typology. In the following section, I will present the ACQDIV project in more detail, as well as the corpora contained in the database which we have built during the first year of the project's timeline. Based on Moran et al. [to appear], I will also briefly illustrate the various steps which were necessary to clean and process the various corpora, in order to combine them in an interoperable, cross-linguistic database of longitudinal child language acquisition data.

[3]The technological incompatibility can, to some degree, also be explained by the compilation dates of the corpora, which range from 1984-2005 (Moran et al. [to appear].

2.2. General Project Description

ACQDIV is a project founded by the European Research Council and led by Prof. Sabine Stoll from the Department of Comparative Linguistics and the Psycholinguistics Lab at the University of Zurich. The full title of the project is *ACQDIV – **Acq**uisition processes in maximally **div**erse languages: min(d)ing the ambient language*. During the first year of the project, we have assembled ten electronic corpora of longitudinal child language acquisition data into one cross-linguistic database. During this process, the following issues were encountered and had to be solved (Moran et al. [to appear]):

- Compiling disparately formatted and annotated corpora, cleaning and processing them in order to make them technologically and linguistically interoperable.
- Creating workflows for extracting, transforming and loading the data into standardized formats and interfaces for further cross-linguistic analysis.
- Producing a single unified database structure that allows us to mine patterns in naturally spoken language at the utterance, word and morpheme level.

In the following sections, I will describe the corpora included in the ACQDIV database and how they were selected. The language sample does not only show diversity with respect to geographical, cultural and linguistic factors, but the corpora are also very different from a technological and theoretical point of view.

2.3. Selection of Corpora

In order to ensure linguistic diversity, the corpora for the ACQDIV database were selected by applying a fuzzy clustering algorithm developed in Stoll and Bickel [2013] to language data from thousands of languages from the *World Atlas of Linguistic Structures* (WALS, (Dryer and Haspelmath [2013]))[4] and the *AUTOTYP* (Nichols et al. [2013])[5] database, as well as to various typological variables, which are encoded in different ways cross-linguistically. A set of languages with their typological feature values serves as input for the algorithm. The output is then a clustering of maximally diverse languages. Applied to the aforementioned data, a cluster of 5 maximally diverse language groups, ranging from more isolating to agglutinating

[4] `http://wals.info/`

[5] `http://www.autotyp.uzh.ch/`

to polysynthetic languages[6], was identified. To address the cross-linguistic nature of the ACQDIV project, two languages from each cluster were chosen. For nine out of those ten languages richly annotated longitudinal child language acquisition corpora already existed. A tenth corpus for the Athabaskan language *Denë* is currently being completed under the guidance of Dr. Dagmar Jung from the University of Zurich.[7] A general overview of the languages contained in the ACQDIV database is given in Table 2.1.[8]

	ISO 639-3	**Language**	**Speakers**	**Classification**
1	cre	Cree	87,220	Algic
1	chp	Denë	11,900	Na-Denë
2	ind	Indonesian	23,200,480	Austronesian
2	yua	Yucatec	766,000	Mayan
3	ctn	Chintang	3,710	Sino-Tibetan
3	ike	Inuktitut	34,510	Eskimo-Aleut
4	rus	Russian	166,167,860	Indo-European
4	sot	Sesotho	5,634,000	Niger-Congo
5	jpn	Japanese	128,056,940	Japanese
5	tur	Turkish	70,890,130	Altaic

Table 2.1.: Language Sample

Table 2.2 shows an overview with regards to the file format used in which the files of each corpus were coded, the number of target children involved, as well as the number of sessions and words (tokens) contained in each corpus in the ACQDIV database.

Each corpus is a detailed multi-year longitudinal study that consists of spoken language utterances by numerous participants in culturally distinct settings and contains target children and child directed speech mainly from mother-child interactions. Some corpora, like Chintang, contain a variety of participants, including parents, family members, playmates, etc. (Moran et al. [to appear]).
As is best practice in the development of child language acquisition corpora, recordings are made at regular intervals (e.g. every week or every two weeks) for one or more years and are centered around a number of target children. For example, the

[6]And also differing with respect to the number of contrastive sounds (Moran and Wright [2009]) and a number of certain typological parameters, such as the presence and nature of agreement and case marking, word order, degree of synthesis, polyexponence and inflecti onal compactness of categories, syncretism and inflectional classes.

[7]The Denë corpus contains audio-visual recordings of eight children ranging in age from 2-4 years and their families. It currently consists of 200+ sessions (190hrs+) with transcriptions and translations in another Toolbox-based format called ELAN, and glossing in Toolbox (Moran et al. [to appear]).

[8]Population figures are taken from the Ethnologue (Lewis et al. [2009]).

Language	Format	Children	Sessions	Words
Chintang	Toolbox	4	419	828272
Cree	CHAT	1	10	21525
Indonesian	Toolbox	8	997	2496828
Inuktitut	CHAT-like	5	77	73302
Japanese	XML	7	341	1235364
Russian	Toolbox	5	448	2022992
Sesotho	XML	4	129	237247
Turkish	CHAT-like	8	373	1139877
Yucatec	CHAT-like	3	234	120441

Table 2.2.: Corpora Contained in the ACQDIV database

Chintang corpus (Bickel et al. [2011]), which was compiled between 2004 and 2015, contains nearly 1 million words, hundreds of participants, morphological annotation, as well as part-of-speech tags, and is translated into English and Nepali (Moran et al. [to appear]). Figure 2.2 is an example of a conversational exchange encoded in the Toolbox format used in the Chintang corpus.
Although each corpus is encoded in a different format, they are all transcribed at the utterance, word and morpheme level and additional annotation tiers include utterance timestamps, morphological analysis, part-of-speech labels, etc.[9] In order to unify the data for the ACQDIV database, we developed our own ETL[10] (Inmon [2005],Kimball and Ross [2011]) pipeline. The following subsections will briefly explain the processes included in this pipeline.

2.4. ETL Pipeline

The ETL pipeline developed for this project is written in the programming language Python and transforms the original data of the various corpora into a single digital format encoded in relational tables using `SQLAlchemy` (Bayer [2016]) for its database ORM and `SQLite` for data storage (Moran et al. [to appear]).
In the **extraction** process the data we had to deal with included challenges such as: the corpus-compiler specific specifications encoded in Toolbox, various versions of the CHAT standard, as well as the nested (and often idiosyncratically coded) XML structure in the TalkBank XML format.

[9]The data in the ACQDIV project therefore provide not only a suitable platform for language acquisition research, but can also serve as a resource for developing data-rich NLP tasks *for* and *with* under-resourced languages data (Moran et al. [to appear]).

[10]Extract-Transform-Load.

The **transformation** process in the ETL pipeline included the following steps (Moran et al. [to appear]):

- Removing duplicate sessions from the corpora.
- Converting the files into Unicode (UTF-8 NO-BOM NFD) plain text.
- Correcting legacy character codes that were lost in translation and identified with unigram character code and grapheme models.
- Extracting annotator comments from the utterance level and removing punctuation from the utterances.
- Associating all annotations explicitly with the three main levels utterance - word - morpheme.[11]
- Unifying metadata standards and data types (e.g. unifying speaker role labels, age formats).
- Unifying linguistic terminology from the different annotations; creating terminological interoperability by mapping linguist expert opinion of grammatical categories to a unified set.
- Inferring additional information from annotations, e.g. determining the sentence type from punctuation in the translation.

After the transformation step, the sessions of the various corpora have been processed[12], the data is **loaded** into a simple relational SQLite database, which contains the following tables[13]: `sessions`, `speakers`, `utterances`, `words` and `morphemes`. Each session from the `sessions` table has multiple speakers and multiple utterances. Each utterance from the `utterances` table is in a one-to-many relationship with words and each word from the `words` table is in a one-to-many relationship with the morphemes from the `morphemes` table (Moran et al. [to appear]). Sessions contain additional metadata (e.g. date, location, speaker). Utterances contain information like timestamps and addressee. Words and morphemes contain linguistic analysis, e.g. part-of-speech tags, morphological glosses, and utterances may also have a phonetic transcription. Figure 2.4 shows an entity relationship diagram (ERD) of the ACQDIV database.[14] The content of the database can be mapped to various output

[11] Included in this step, conflicting structures in the original data that are used to align word and morpheme level annotations within an utterance often had to be reassembled as well.

[12] While the original subcorpora consist of several sessions, where each in turn may or may not be instantiated by several files, all subcorpora and all their session-related data are contained in a single file in the ACQDIV Corpus.

[13] The tables are linked to their initial corpus via session IDs and participant codes, respectively

[14] Beside text, the original subcorpora also contain media files (mostly digitized audio and/or video files). The ACQDIV Corpus does not include these files to protect the children's privacy-

formats, including R data frames (R Development Core Team [2008]) and simple CSV files.

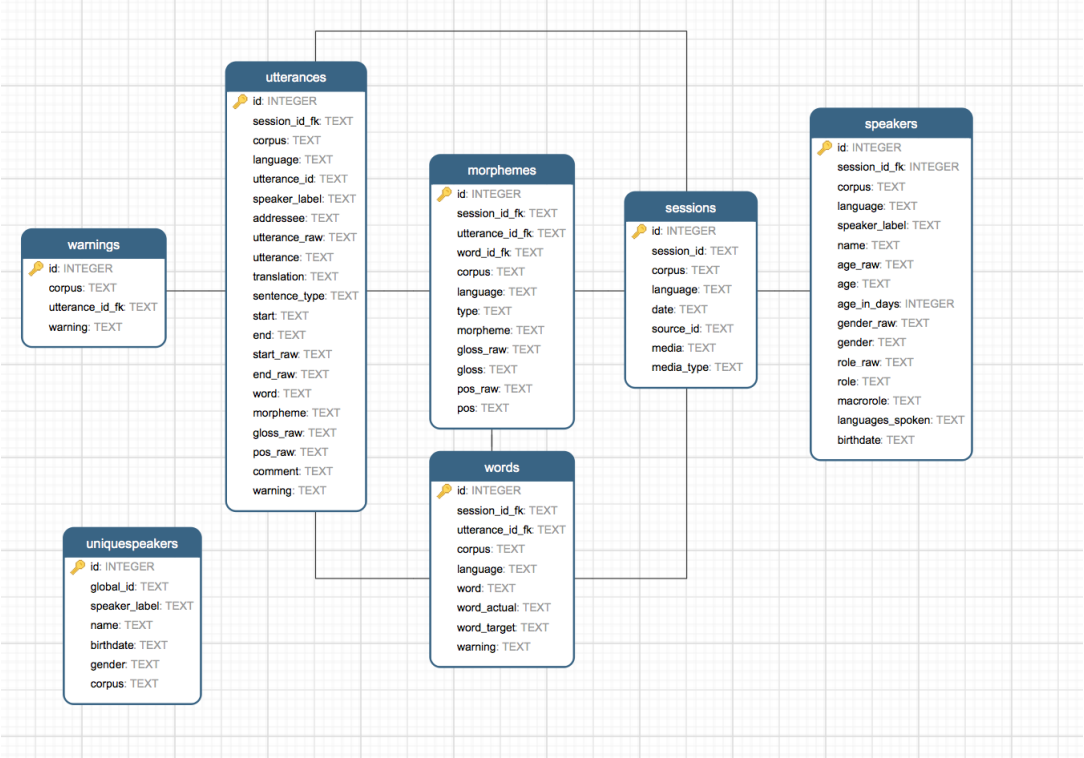

Figure 2.4.: Entity Relationship Diagram of the ACQDIV Database

After introducing the data and extraction of the underlying data for this thesis, I will focus on historical and theoretical aspects of (linguistic) data visualization in chapter 3, before discussing the creation of our own visualizations for specific research questions related to child language acquisition in chapter 3.4.

sensitive information. However, the names of the original media files can be found in the sessions metadata table.

3. Visualizing Data

Visualizations of textual data have strikingly grown in popularity in the last two decades in various thematic fields, because the amount of available data is increasing with such speed, that simple manual analyses cannot be conducted anymore in a reasonable amount of time. Also the development of more data visualization tools such as the JavaScript libraries `d3.js` (Bostock [2012]) and `p5.js` (McCarthy [2015]), as well as the graph visualization toolkit *Gephi* (Bastian et al. [2009]) has opened the field of interactive data visualization to a growing range of end-users. Nevertheless, even though the field of data visualization as it exists today has a rather young tradition, using abstract visual forms to communicate insights drawn from analyses of numerical and nominal data can be traced back to the 16th century, where the earliest forms of data visualization arose from geometric diagrams, tables of the positions of stars and other celestial bodies, as well as maps, which were mostly used as aids for navigation and exploration (Chen et al. [2007, 18]). In section 3.1, I will first present a historical overview of the development of using graphical representations to visualize statistical data, in order to show, on the one hand, the development of various visualization techniques, but on the other hand also how the handling of information changed due to various developments in the interplay of statistics and data visualization. Then, in section 3.3, I will narrow my focus to the usage of data visualization in the field of linguistics and present the main difficulties linguists face when trying to represent linguistic data in an abstract visual form. Section 3.4 will then be about using visualization forms specifically addressed at research questions from child language acquisition.

3.1. Historical Development in Data Visualization

When tracing back the origin of visual display of "real-world entities", one could principally also consider cave paintings, pictographic cuneiform inscriptions or rebus-texts as precursors of today's visualizations (cf. Collins [2005, 1]). However, as the focus in this thesis lies on visualizations which are created from a quantitative approach, I do not discuss iconic cave paintings nor textual writing in general in

this thesis.[1] Figure 3.1 taken from Chen et al. [2007, 18] shows a diagram of the historical "milestones" in the field of data visualization.

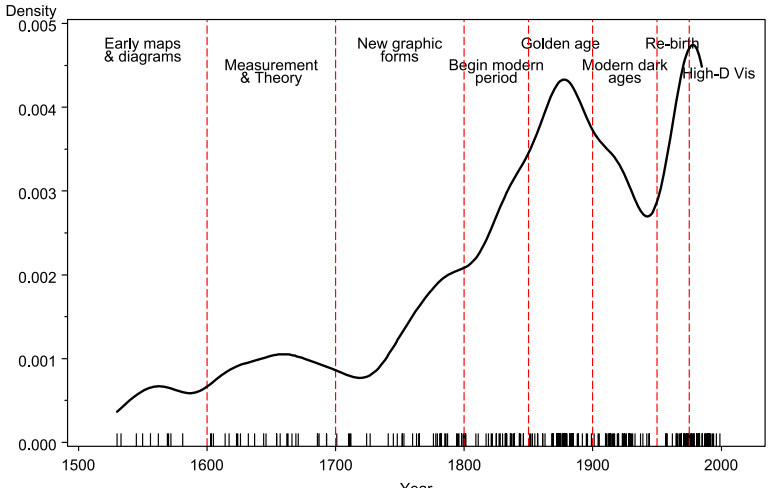

Figure 3.1.: Milestones in data visualization (source: Chen et al. [2007, 18])

In the above figure, the various epochs divided by the red dashed lines characterize the main accomplishments for each epoch, the black line (as well as the density plot at the bottom of the diagram) shows how the number of developed visualizations increased (but also decreased) over time. The main accomplishments within these epochs are characterized by Chen et al. [2007] as follows:

Early Maps and Diagrams (Pre-17th century)
Even though the beginnings of quantitative data visualization have started to appear mostly in the time when also statistical data became increasingly available, one of the earliest graphical representations of quantitative information (and probably also one of the best-known) dates already back to the **10th century** (see Figure 3.2). This visualization by an anonymous author shows the changing position (in space and time) of the seven most prominent planets. The horizontal axis of this diagram represents a time line and is divided into 30 (more or less equally sized) intervalls. The vertical axis represents the inclination of the planetary orbits. According to Chen et al. [2007], the grid-like division of the ground in the visualization is a remarkable point for such an early graphic, because the notion of a coordinate system was only fully developed by the 17th-18th century (Chen et al. [2007, 18]). Further developments in this epoch include the idea to plot relations between tabulating values by Bishop Nicole Oresme [1323-1382] and the idea of a theoretical graph showing distance versus speed by Nicolas of Cusa in the **14th century**. The **16th century** was then marked by new technological innovations, such as the cre-

[1]For a historical overview which also includes these kinds of visualizations, see Collins [2005].

ation of instruments and techniques to measure and observe geographic position, as well as physical quantities (Chen et al. [2007, 19]). Further innovations which influenced the early beginnings of real quantitative data visualization were the invention of the camera obscura by Reginer Gemma-Frisius, the usage of mathematical tables, as well as the first modern cartographic atlas by Abraham Ortelius (the *Theatrum Orbis Terrarum*), dating back to 1570 (ibid.).

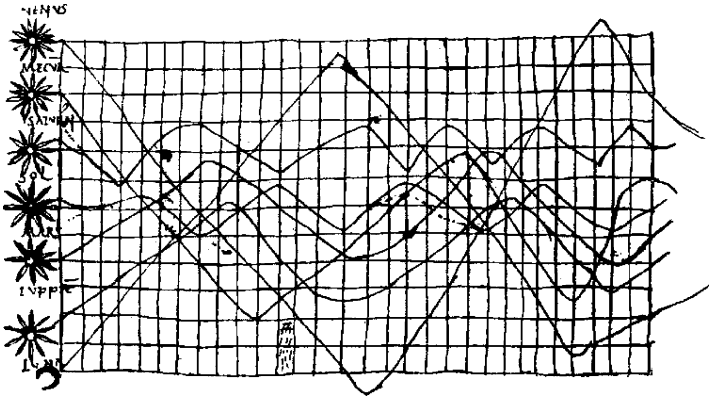

Figure 3.2.: Planetary Movement Over Time (source: Chen et al. [2007, 19])

Measurement and Theory (1600-1699)
The **17th century** was marked by the need of physical measurement: New techniques made nearly everything in the world quantitatively measurable: time, distance and space became important variables not only in map making and astronomy, but also in surveying. Probability theory and demographic statistics were more and more applied to study population, land marking and taxes (Chen et al. [2007, 20]).

With respect to data visualizations, the first example of a visualization which later was coined as *small multiples* by Edward Tufte (1983) appeared at the beginning of the 17th century. Figure 3.3 shows this visualization which illustrates changes in sunspots from October 23rd until December 19th 1611. The large circle in the upper left-hand corner shows the groups (labelled by the letters A-G) of the sunspots. The visualization with small multiples is then used as "a series of graphics, showing the same combination of variables, indexed by changes in another variable" (Tufte and Graves-Morris [1983, 168]).

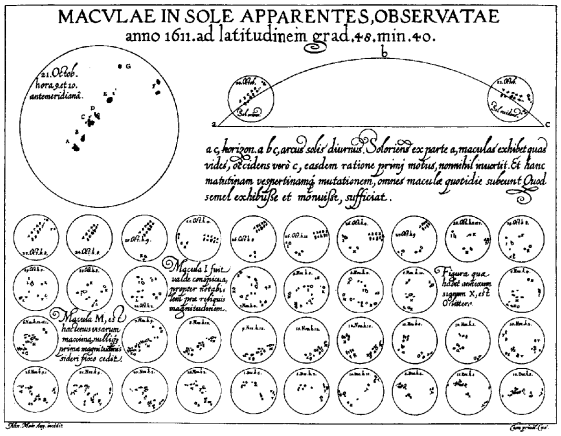

Figure 3.3.: Changes in Sunspots

Another visualization dating from this time, which represents another milestone in

data visualization history is a graphic by the Flemish astronomer to the Spanish court, Michael Floren van Langren [1600-1675]. His visualization (see Figure 3.4) shows in a one-dimensional line the distance estimated from Toledo to Rome by various astronomers. For Tufte, this visualization is "the first statistical graphic ever" drawn (Tufte and Robins [1997, 15]).
According to Chen, these two examples illustrate the beginnings of "visual thinking, a new field which emerged at a time where statistical data, some theory to make sense of it, and a few ideas for their visual representation became available" (Chen et al. [2007, 21]).

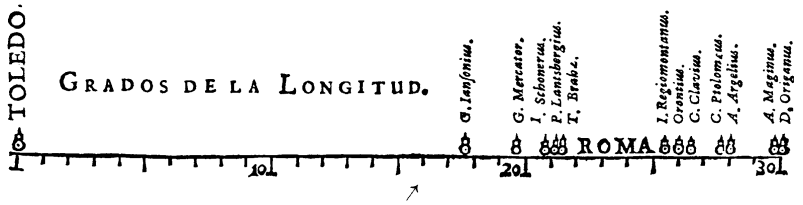

Figure 3.4.: Distances: Toledo to Rome (source: Tufte and Robins [1997, 15])

New Graphic Forms (1700-1799)
According to Chen, the **18th century** was essentially marked by the expansion of using visualizations in new domains and also by combining various graphical forms (Chen et al. [2007, 22]). Geographical maps were not anymore exclusively filled with geographic information, but with other statistical economic (and medical) data as well, which marked the birth of *thematic mapping*. Timelines included not only a temporal line, but also other geometric figures, whose sizes were drawn based on the underlying data. Furthermore, technological innovations such as the three color printing invented by Jacob le Blon in 1710 lead to further development in the field of data visualization (ibid.). Also in this time falls the beginning of the work of one of the most famous data visualizers ever: William Playfair [1759-1823], who is the inventor of many graphical forms which are still in use today (as for example the bar- and the piechart) (Chen et al. [2007, 23]). Figure 3.5 shows Playfairs pie-circle-line chart from 1801, which illustrates the comparison between taxes and population size in several nations. An interesting point to note here is, as mentioned by Chen, that Playfair did in this visualization something which would be considered today as a "sin in statistical graphics" (Chen et al. [2007, 24]): He used two vertical scales with different meanings (the left scale showing population and the right one taxes). Furthermore, as Chen notes, Playfair did not include the diameter of the circles (which represents the land area of the countries) to also pertain to the interpretation of the slopes, which leads to observer unknowing if the degree of the slope has a meaning as well or not (ibid.). Nevertheless, despite his "sin", Playfair's graphic still

succeeds in illustrating that the ratio of population and taxes is different for Britain and Ireland compared to the other states, where the slope shows in the opposite direction.

Even though states began to collect more and more statistical data, this data was very often fragmentary and, more importantly, not publicly available. As can be seen in Figure 3.1, there was a first big increase in data visualizations towards the middle of the 18th century. This increase is particularly shaped by the fact that states made their statistical data publicly available.

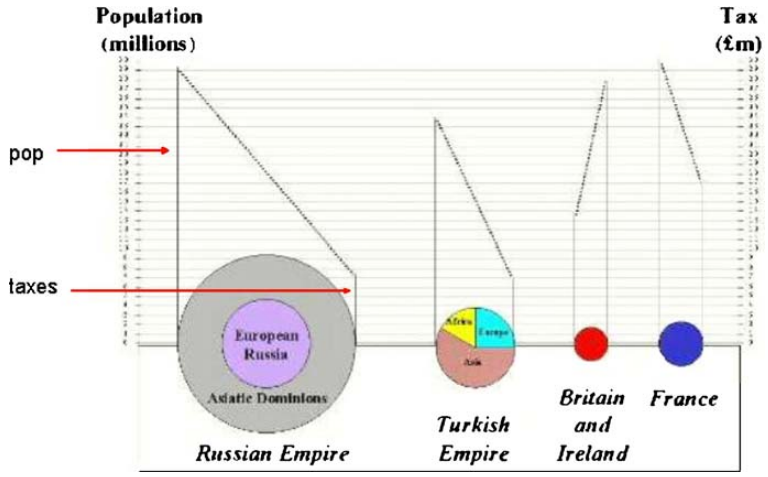

Figure 3.5.: Comparison of Population and Taxes (source: Chen et al. [2007, 24])

Beginnings of Modern Graphics (1800-1850)

As can furthermore be seen in Figure 3.1, the beginning until the middle of the **19th century** saw a massive rise in statistical data visualization and especially also again in thematic mapping. As already noted above, the increasing availability of publicly available data was a major catalyst for the production of statistical graphics. For example, the ministry of justice in France made crime reports publicly available in 1825, which were then used by a French lawyer called Andre-Michel Guerry in a work that is considered the "foundation of modern social science" (Chen et al. [2007, 26]). Furthermore, the severe cholera epidemic from 1813 to 1854 in Great Britain lead to the creation of one of the most famous thematic maps ever: *John Snow's cholera map of London*, which is considered the founding innovation for modern epidemiological mapping (Chen et al. [2007, 31]). With his map, Snow showed that many deaths caused by cholera clustered around the Broad Street pump in London – the common denominator of the people from this district. Using this information, people were able to detect the cause of the spread of the disease and to take measurements to stop the epidemic from spreading and eventually extinguishing it (Tufte and Graves-Morris [1983, 24]).

The Golden Age of Statistical Graphics (1850-1900)

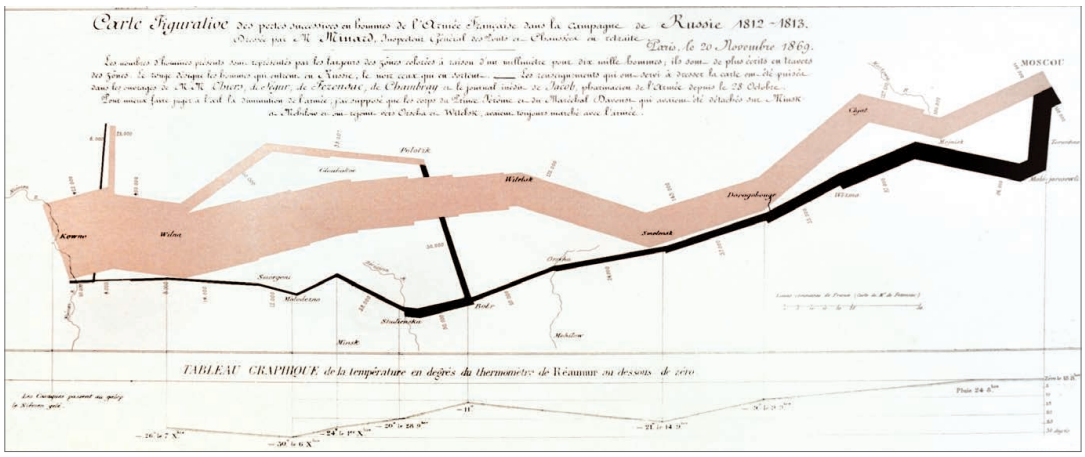

Figure 3.6.: Minard: Map of the fate of the French army during the Russian Campaign (1812-1813) (source: Kraak [2014, 17])

By the middle of the **19th century** using numerical information for transportation, industrialization and social planning and making it publicly available became a common task, which also lead to new insights gained in the field of statistics and mathematics. Again, with more variable data at hand, people further experimented with how to graphically represent this data. For example, the first experiments were conducted to display visualizations not only in a flat manner, but in a three-dimensional space. The first three-dimensional plots where then used to illustrate population data, and later three-dimensional visualization techniques were also applied in cartography, where the first 3-D contour maps where created (Chen et al. [2007, 30]). This period really was something like the "Golden Age of Statistical Graphics" (ibid.), as another two very famous visualizations were created at this time: *Minard's map of the fate of the armies of Napoleon and Hanibal from 1812-1813*[2], which Tufte calls "the best graphic ever produced" (Tufte and Graves-Morris [1983, 40]). In this graphic (shown in Figure 3.6), Minard shows how Napoleon's troops travel form Poland to Moscow (orange line) and back (black line). The main part of this graphic shows the movement of the troops and their size as they march along (with an army starting with around 420'000 men and ending with only 10'000 (Kraak [2014, 19]). The plot in the lower part of the graphic marking the temperature for every location of the troops, as well as the lines representing the rivers illustrate the threats the troops were facing. Providing the viewer with this information helps enormously in understanding what may have caused the death of thousands of men on their way back.

[2]Which was published in 1869 (Kraak [2014, 19]).

The second graphic is the polar area chart invented by the British nurse *Florence Nightingale* [1820-1910] to undertake a campaign for improving sanitary conditions for the treatment of the soldiers on the battlefield (Meirelles [2013, 94]). Nightingale's graphics showed that most soldiers did not die because of the attacks by the enemy, but from preventable disease and the consequences of infections. However, as every Golden Age in history comes to an end, the end of the Golden Age in data visualization was mainly caused by high production costs to produce new, innovative visualizations. Other factors, which lead to the "Modern Dark Ages" will be presented below.

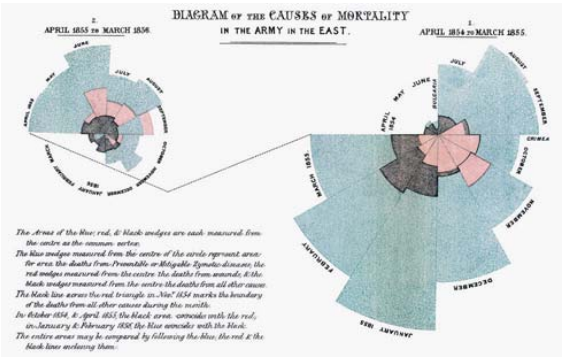

Figure 3.7.: F. Nightingale (1858): "Diagram of Causes of Mortality in the Army in the East" (source: Meirelles [2013, 95])

The Modern Dark Ages (1900-1950)

By the beginning of the **20th century** another important factor lowered the enthusiasm for novel, innovative statistical graphics: The rise of formal, statistical models. Numbers, parameter estimations (including standard errors) were considered precise calculations mirroring real-world data. Visualizations, on the other hand, started to be seen as just "pretty pictures, (...) incapable of stating a 'fact' to three or more decimals." (Chen et al. [2007, 37]). However, even though Chen named this period "The Modern Dark Ages", it actually was not *that* dark, as, at that time, the field of data visualization was characterized by "popularization rather than innovation" (ibid.). Statistical graphics became mainstream, they were printed in textbooks and commercials. The next missing factor, which would again lead to a rebirth of creating innovative data visualizations, was computational power.

Rebirth of Data Visualization (1950-1975)

According to Chen et al. [2007, 39-40], three main factors influenced the rebirth of data visualization: Firstly, John W. Tukey's work in establishing data analysis as a legitimate branch of statistics and his invention of various graphic displays (which were all still hand-drawn) in this book "exploratory data analysis (EDA)" published in 1977 (Tukey [1977]). With his work, Tukey succeeded in making data analysis (and hence also data visualization) again an interesting and respectable field. Secondly, the creation of the first high-level programming language FORTRAN in 1957,

as well as the increasing availability of computers offered the possibility to automatically construct old and new graphic forms by computer programs. The **1960s** where a time where the first interactive statistical applications including high-resolution graphics were developed. The third factor, which, according to Chen, influenced the rebirth of data visualization was an increasing collaboration between various fields: research in computer science would combine with developments in data analysis and display technologies, which resulted in creating and providing new paradigms, programming languages and software packages for implementing digitized graphics. According to Chen, by the end of this epoch, the first examples of modern geo-information systems (GIS) and interactive 2-D and 3-D statistical graphics would appear (Chen et al. [2007, 40]).

Interactive and Dynamic Data Visualization (1975-present)
The **last quarter of the 20th century** has then introduced a new level of interdisciplinary research, which resulted in a multitude of handbooks and software tools for a wide range of simple, but also highly sophisticated data visualizations. The **1970s and 1980s** were mostly characteristed by advances in statistical graphics which concentrated on generating static graphs for multidimensional quantitative data. Techniques for dimension reduction such as *principal component analysis, multidimensional scaling* and *discriminant analysis*, allowed the analysts to reduce the high-dimensional data set to a lower dimension by including only patterns which were of interest (ibid.). The main advantages in the development of data visualization within this period (especially since the last two decades) came surely from the development of dynamic and interactive visualizations, allowing the analyst to manipulate the data set and instantly see the influence on the (statistical and graphical) results (Chen et al. [2007, 41]).

The main developments in the field of data visualization during the **21st century** have produced a myriad of dynamic and interactive data graphics. As computational power is something that nowadays increases constantly, the new challenges in data visualization lay not anymore exclusively in acquiring enough data or developing more powerful computers and applications. Instead, as data visualizations seem to be omnipresent, new challenges also arise in critically discussing the steps involved in creating, but also applying these graphics to various data sets and also in various research fields. In order to collect parameters for a framework, in which data graphics can be analyzed and critically discussed, I will in the next chapter firstly present theoretical approaches to data visualization, which include insights from human cognitive perception, Gestalt theory and other guidelines and frame-

works for analysing data visualizations. These guidelines will then be used (and also discussed) in chapter 4 and 5, where the theoretical concepts will be applied to the conceptualization and creation of data visualizations for linguistic research questions in child language acquisition.

3.2. Theoretical Concepts

As we have seen in this historical outline, especially the 18th and 19th century were a period where the need to illustrate statistical data collected from social, medical and economic surveys served as a catalyst force to create graphics which would allow to examine correlations between various data points more closely and to present the data in more understandable way, but also to use such graphic representations for urban planning or political purposes (Friendly and Denis [2001]). With the increasing availability of more computational power, it has become more and more common to use visual representations as substitutes for alphabetic writing, especially when seeking information in large corpora of text (Collins [2005, 1]). Thus, when talking about data-driven visualizations today, two main categories can be distinguished (Chen et al. [2007, 1-3]):

- **Scientific Visualization**
  Scientific visualizations are used to illustrate scientific data and the relations in real-world data sets (for example in molecular structures or meteorological data). The goal in using scientific visualizations is to describe the underlying data as realistic as possible. Examples for scientific visualizations would therefore be 3-D models of e.g. molecules or genomes.

- **Information Visualization**
  Information visualizations are used when illustrating abstract data, be it from the field of economics, social statistics or linguistics. Because information visualization is primarily used to gain information, a commonly used definition from Card et al. [1999] is the following: Information visualization is " the use of computer-supported, interactive and visual representations of abstract data to amplify cognition" (Card et al. [1999] cit. in Collins [2005, 3])

The above definition for information visualizations is based on Scaife and Rogers'(1996) notion of ***external cognition*** (Collins [2005, 3]). External cognition is defined as the unification of internal and external cognitive processes, which produce thought (ibid). These thought processes can be supported by external aids, which enhance

the human cognitive capability of recognizing new patterns and relations. The following quote by Norman (1993) further explains this idea:

> The power of the unaided mind is highly overrated. Without external aids, memory, thought, and reasoning are all constrained. But human intelligence is highly flexible and adaptive at inventing procedures and objects that overcome its own limits. The real powers come from devising external aids that enhance cognitive abilities. How have we increased memory, thought, and reasoning? By the invention of external aids: **it is things that make us smart**. (Norman [1993], cit. in Collins [2005, 3], highlighted by D.P.)

A further dichotomy which is often used to characterize data visualizations is the one between ***explanatory*** and ***exploratory graphics***:

- **Explanatory Graphics**
  Explanatory graphics are also called *presentation graphics*, because the main purpose of such visualizations is to *explain*, i.e. to present and mainly focus on the findings of e.g. a study or certain aspects of the data which are already known by the creator of the visualization. Therefore, explanatory graphics are mostly used to prove a point or hypothesis, to support a decision and to communicate information (Ryan [2016, 184]).

- **Exploratory Graphics**
  Exploratory visualizations create a way to interact with a dataset or subject matter. The explorative approach to a data set may uncover stories, which can later be analyzed further and explained (ibid.). Exploratory graphics are intended to facilitate the discovery process that could potentially lead to the finding of different or new insights, or maybe none of them at all (which, of course, also depends on the data set and the research question).

Note that the distinction between explanatory and exploratory is not a clear-cut one, many visualizations can have characteristics of both of them.

The main goal of data-driven information visualization is therefore to increase the efficiency of (computer-controlled) information processing and display. However, an enhancement in efficiency will only be achieved, if the visualization establishes an assisting relation between the information within the data set and the cognitive processes involved in decoding the information (Collins [2005, 3])). Therefore, in choosing a suitable visualization form, one has to, on the one hand, consider the nature of the data set, the research question one wants to answer, but also the cognitive gain a user can obtain with a visual representation of the data. A further

point to decide upon is the primary goal of the visualization: Shall it be a graphic which is used to primarily show the results of a study, or is the visualization itself going to be a tool in the explorative analysis of the underlying data set? According to Collins (ibid), research in human-machine interaction and cognitive science has shown that well made information visualizations can lead to a faster and better understanding of the underlying data (ibid). Consequently, the question arises what it takes to produce "good" information visualizations: what parameters in the human cognitive processes of information processing play important roles in decoding visual information? How can we use graphical forms so that they really help in gaining new insights? In the following subsections, I will present findings from cognitive science on the one hand, but also from psychology and data visualization theory, in order to use these various insights and combine them in a framework based on which the visualizations created in chapter 4 and chapter 5 will be further analyzed.

3.2.1. Human Cognitive Perception

Information visualization has become a multidisciplinary filed consisting of various sub-fields such as human-machine interaction, computer science, mathematics, statistics, cognitive science and arts. According to Collins [2005, 6], insights from the above disciplines have shown that effective information visualizations can lead to the following benefits for human cognitive processes:

Benefit	Description
Comprehension	Enabling humans to understand huge amounts of data.
Perception	Revealing unexpected properties of a data set.
Quality Control	Discovering potential problems in the data (or in the data collection methods).
Focus and Context	Facilitating the understanding of small scale features in the whole data set.
Interpretation	Supporting hypothesis formation which leads to further investigation.

Table 3.1.: Cognitive Benefits of Good Information Visualizations

Based on these benefits, information visualizations can not only help to present or analyze the data, but they can also help as a tool for quality control. This goes in line with Keim et al. [2008] and their theory of *Visual Analytics*, which they define as an approach which includes more than just producing data visualizations. According to Keim et al., visual analytics is a process which combines "decision-making, visualization creation, human factors and data analysis" (Keim et al. [2008, 158]). Furthermore, according to the visual analytics approach, the challenge does not only lie in defining the right algorithms for an analysis task, but also in identifying its limits, which in turn can lead to alternative paths with alternative algorithms and

visualizations than initially planned (cf. ibid.).

In order to use visualizations as cognitive aids, one first has to look into the characteristics of human cognitive processes, in order to learn how to effectively use graphical representations of the underlying data. Spence (2001) describes the formation and interpretation of ***internal models***, which humans complement with ***external information***, in order to complete ongoing cognitive processes (Collins [2005, 8]). According to Spence, there are so-called "cognitive maps" (i.e. internal representation of relations within a data set) and "cognitive collages" (i.e. sets of weakly connected cognitive maps), which represent the basis of the internal models (ibid). If these internal models are incomplete (i.e. if information is missing), the need for additional information appears. If this new information is now visually encoded, there are some proposals as how to display them: too abrupt or too slow effects are to be avoided because internal models are not good at processing too fast or too slow changes. Therefore, it is better to display changes and navigation in visualizations gradually. This proposition is also supported by the gestalt law of *continuity* (the gestalt laws will be explained in more detail in the next section) and the ecological approach of Wise [1999]: the world we live in makes us predisposed to see continuity because events in the real world also happen continuously (Collins [2005, 8]).

3.2.2. Gestalt Theory

According to Spence [2001], our internal models are formed by our cognitive perception. This formation can be explained by a theory called ***Gestalt theory***. Gestalt theory initially started in the field of psychology and philosophy (especially in the work of Christian Ehrenfels) and was developed at the beginning of the 20th century by Max Wertheimer, Wolfgang Köhler and Kurt Koffka. Already at the end of the 19th century, Ehrenfels described that there are qualitative nuances within the human cognitive perception, which do not only result from the regulation of simple qualities of our senses (Rollinger and Ierna [2015]). With respect to information visualization, the important part of Gestalt Theory lies within the idea, that humans usually perceive more than what they are consciously aware of. This is the reason why we are able to perceive fragmentary information. This, as Collins calls it, "heightened perception" (Collins [2005, 9]) is possible because of laws (i.e. the *Gestalt laws*) which describe human tendencies to establish relations between unconnected objects. Even though this theory is not completely uncontroversial (cf. Collins [2005, 9]), its basic principles with respect to human perception can serve as guidelines for producing cognitively effective information visualizations. For exam-

ple, the Gestalt laws shown in Figure 3.8 describe principles which can be helpful when designing information visualizations. Table 3.2 includes an explanation (from Taylor [2014, 11]) for every of these Gestalt laws.

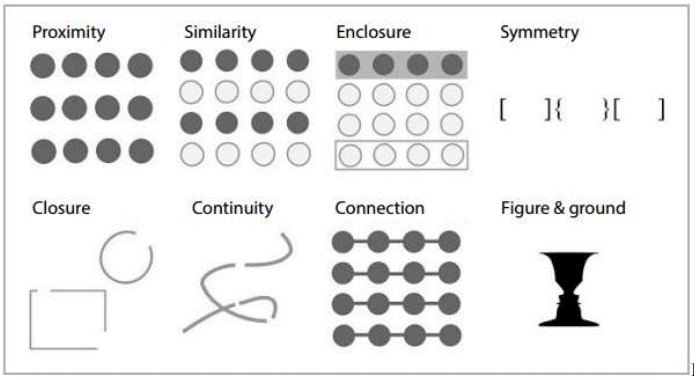

Figure 3.8.: Gestalt Principles (source:Taylor [2014, 11])

Gestalt Principle	Description
Proximity	Objects placed nearer together will be perceived as a unit.[3]
Similarity	Similar objects (in size, shape, colour, etc.) tend to be perceived as part of the same group.
Enclosure	We tend to group the first and and last elements of a graphic together.
Symmetry	Symmetrical pairs (if detectable) are perceived more strongly than parallel pairs.[4]
Closure	Contours with gaps will be perceptually closed.[5]
Continutity	Smooth and continuous lines will be perceived before discontinuous ones.[6]
Connection	Connected objects are perceived as belonging to the same group.
Figure & Ground	We either notice the vase or the two faces. Whichever form we notice becomes the figure, the other one the ground.

Table 3.2.: Description of Gestalt Principles (Taylor [2014])

The principles from Gestalt theory can not only be used to create visualization where patterns are more easily detectable, but also in order to create animations for interactive visualizations which should help the user to understand the changes within the visualization.

3.2.3. Visual Information-Seeking Mantra(s)

One of the most famous guidelines with respect to building interactive data visualizations comes from Ben Shneiderman (Shneiderman [1996, 337]), who introduced the ***Visual Information-Seeking Mantra***:

"Overview first, zoom and filter, then details-on-demand."

This trichotomy ensures that a graphic can be effective on a global as well as a local level. With a first step (***overview first***) the user is given the opportunity to detect

interesting patters. The parts which a user considers important can then be focused on by either zooming to the parts of interest or by filtering out other parts that are not of importance (***zoom and filter***). The scope chosen in step two can then be examined in again further detail (***details-on-demand***). Shneiderman's trichotomy introduces a good starting point for creating interactive visualizations. However, as it is presented, it could, potentially, be understood as a one-way process, but the information visualization mantra is actually a process in which the observer interacts with the data in a constant "information-seeking loop". This point has also been discussed by Keim et al. [2010, 10], who, therefore, extended Shneiderman's Visual Information-Seeking Mantra as follows:

"Analyse first, show the important, zoom and filter, analyze further, details-on-demand."

When comparing the two visualization mantras, one can see that the focus in the visualization mantra of Keim et al. [2010] seems to be equally balanced between data analysis[7] and data visualization. As already mentioned further above, in the field of visual analytics, a graphical representation of the data is not primarily used to show already computed results, but it is used to experiment with the data, to explore it, to (maybe) detect errors in the data set or to completely discard an already chosen path.[8] Therefore, the first two steps ***analyse first*** and ***show the important*** are concerned with a first visual overview of already filtered data. If the data shows interesting patterns, one can go on to ***zoom and filter, analyze further*** and (if necessary) get ***details-on-demand***.

Figure 3.9 shows the visual analytics working process, which is understood as a combination of using the strengths of computational *and* human data processing. Machines are used to mine the data, which can then be visualized in various forms. These visualizations then serve as "anchor points" which help in deciding if the approach taken in the data mining process, as well as in the visualization process brings new insights or the expected (or maybe also unexpected) results. In visual analytics, we are in a "semi-automated analytical process" (Keim et al. [2008, 156]) of human-machine interaction.

[7] Hence the name of their research field "visual analytics".

[8] Which, according to visual linguistics is nothing bad, because insights from failure can still be used to turn one's sight to alternative approaches, which might bring the desired effects, or lead the researchers to completely change either the data set or the research question (Keim et al. [2008, 163]).

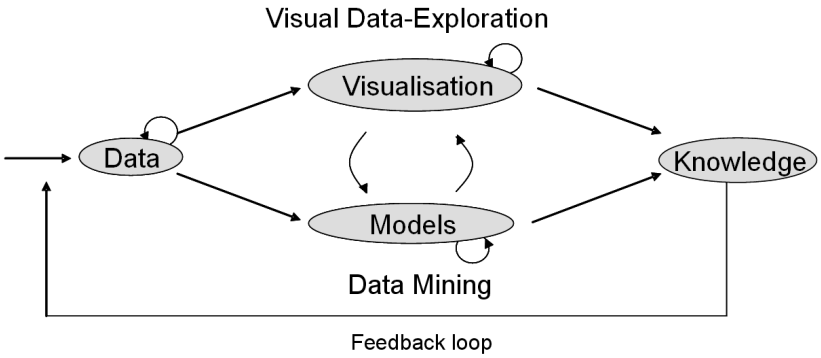

Figure 3.9.: Visualization Pipeline (Source:Keim et al. [2008, 156])

After having discussed and presented parameters from human cognitive perception, as well as new approaches to data visualization, I will in the next section introduce a further level for analyzing data visualizations which is more concerned about the graphical parameters, as well as the characteristics in general of diagrams: Sibylle Krämer's notion of *Diagrammatik*.

3.2.4. Diagrammatik

Sibylle Krämer introduces with her concept of "Diagrammatik" a theoretical background to define and describe characteristics of diagrams. These attributes belong to, as Krämer [2013] calls it, the "Grammatik der Diagrammatik", the 'grammar of the Diagrammatik'. Parameters, which belong to this grammar are the following:[9]

- **Planarity**
  Planarity is a parameter which enables the beholder to gain an 'overview' of the data by abstracting it down onto a 2-D plane. By looking at the data from a "bird's eye" perspective, planarity reduces the complexity of the data to a low-dimensional, abstract representation.
- **Graphism**
  According to Krämer, the interaction of points and lines, placed on a plane as information-bearing entities, unfolds the force of the graphism. Points and lines are the root of the drawing and the script, because by using them as means of connection or separation, they are assigned a meaning. Krämer describes graphism like another form of a "language" (Krämer [2009, 165]), with the difference that meaning is not encoded temporally and sequentially. Instead,

[9]The parameters, as well as their description is taken from Krämer [2013, 164-167].

graphism unfolds its meaning by creating simultaneity, which is organised in a spatially radial manner (ibid.).

- **Relation**
  The important aspect of diagrams is not the fact *that* they show relations, but *how* they do it. They serve as medium of "topographical arrangements" (ibid.), where spatial and non-spatial, arbitrary and regular correlations are made visible and thus also manageable.
- **Syntheses of Picture and Text**
  Diagrams achieve a symbiosis between text and picture: schematic drawings and written words are combined in order to shape a new abstract whole.
- **Usefulness**
  In contrast to art, the main characteristic of diagrams is not self-reference, but reference to an other: diagrams do not "show themselves, they show *something*" (Krämer [2009, 166]).
- **Generating Knowledge**
  By looking at a diagram, the beholder is enabled to gain new insights (and can thus also acquire new knowledge).

The grammar of the Diagrammatik seems at times more like a philosophical essay behind describing what actually carries meaning in a diagram and how it does that. However, the above listed parameters can indeed be used to characterize data visualizations, as the "points and lines" in data graphics can be assigned different meanings. When collecting various examples of data visualizations, Krämer's parameters can be used as part of a general framework to describe and compare various data visualizations. As part of the project *Visual Linguistics*, we have developed a further framework which includes Krämer's diagrammatical parameters, as well as other aspects from semiotics, characteristics of the underlying data and metadata, as well as technical attributes in order to describe and compare already existing visualizations.

3.3. Linguistic Data Visualization

"A picture is worth more than a thousand words."

This common saying, which is said to have emerged in the USA in the early 20th century, has been widely used for many years to propagate visual representations of data, especially in the commercial field (Hepting [2008]). As we have already seen in

chapter 3.1, the developments in the field of data visualization have recently focused on improving the quality of the graphics, as well as on increasing the computational power in order to be able to process more and more data. With the wide distribution of static and interactive graphics over the world wide web, applications of data visualization have also expanded to new domains. One domain, which has also seen an extreme growth in the number of graphics created is linguistics. Even though the visualization of language has its roots far deeper than in the late 1990s (Collins [2005]), it was at about this time, when the first sporadic visualizations from linguistic data were created (Lyding [2012]). The decade from 1990 until 2000 has seen new tools being developed for visual linguistic data analysis. Such tools include e.g. *DocumentGalaxies* (Wise et al. [1995]) and *Leximancer* (Smith and Humphreys [2006]), as well as forms of keyword visualizations.[10] The Mid 2000s were then characterized by an increased use of so-called *word clouds* (e.g. Figure 3.10) and by the late 2000s, the number of visualization websites and blogs, such as `eagereyes.org`[11] or *ManyEyes*[12] further pushed the popularity of linguistic data visualizations. Furhthermore, especially the field of computational linguistics started to organize workshops and conferences which were explicitly aimed at the subject of visualizing linguistic data (Lyding [2012]). However, a critical reflection about applying theories and techniques from data visualization to linguistics has only started in recent years (Bubenhofer et al. [2016]).

Figure 3.10.: Wordcloud Visualization

The quote at the beginning of this paragraph might be true in many cases, but a meaningful conversion from "a word" into "a picture" is often far more complex than it might seem at first sight. Linguistic data often does not consist of a small group of words, but rather of linguistic collections with millions of words, all being connected not only in a hierarchical way (for example in compositional (e.g. syntactic) structures of linguistic units), but also in a relational way (in e.g. semantic relations and/or associations between words), which both lead to a multifaceted and

[10] Which include also the Key-Word-In-Context (KWIC) visualizations in linguistic corpora.

[11] `https://eagereyes.org/`

[12] Note that ManyEyes has been shut down last year and is since incorporated into IBM's Watson Analytics. Nevertheless, it's still noted here, as ManyEyes was one of the first approaches in making interactive web-based visualizations publicly available for data analysis (especially also for textual data). Also, its founders Martin Wattenberg and Fernanda Viégas are two of the most known people in data visualization research.

multidimensional relationship of various linguistic units within collections of textual data (Lyding [2012]). Such multidimensional relations within a data set allow for a multitude of different research approaches to one and the same data set. However, this characteristic makes it also extremely difficult to break down and abstract a data set which is inherently already an abstract representation of an even more abstract concept created by mankind, namely the human language. Textual representation of language is not something that is given by nature, obeying natural laws and having a certain range of development that could easily be . Instead, language is like an ever evolving organism, created and constantly shaped by social interaction. The dynamic and innovative nature of language is one fact which makes not only visualizing linguistic data difficult, but also applying various statistical algorithms to calculate various phenomena (as, for example semantic similarity). Furthermore, when we compare the types in which data is commonly classified, we can use the following labels (Lyding [2012], Hearst [2009]):

Quantitative data: Are numerical values, which can be processed arithmetically (e.g. integers and real numbers).
Qualitative data: Is everything else, which can be further divided into:

- **Interval data**: Is actually essentially quantitative data as well, which has been discretized and made ordered with measurable distances (e.g. time converted to days, months and years).
- **Ordinal data**: Is data, which can be placed in an order, but without definition-depending measurable distances (e.g. first-second-third or hot-warm- cold).
- **Nominal data**: Is data without inherent organization (e.g. weather types, colors, gender, textual data).
- **Hierarchical data**: Is also data without an inherent order, but which can be arranged into subgroups (e.g. [[mammals, [cat ...], [dog [pug...],...],...],...] , etc.).

In the above list we can already spot another challenging factor when visualizing textual data: its lack of inherent ordering (Hearst [2009]).[13] Of course, one could order textual data alphabetically, but again, this is not an inherent feature of all textual data, because an alphabetical ordering might be logical in our central European culture, but for other cultures this might not be the case. What is more, often an alphabetical ordering does not mirror the importance of words based on some other statistical results.

[13] An exception here are visualizations which illustrate for example syntactic dependency. Here, we of course have a structured ordering of the textual units which obey language-specific grammatical rules, but these rules are created from human beings by defining rules and conventions.

A further challenge in visualizing textual data lies in the often huge size of linguistic corpora. There are several kinds of visualization forms that work well with limited amounts of nominal data, as, for example tree diagrams or network graphs, which can be used to convey various kinds of relationships among nominal variables. However, such visualizations (as we will also see in our two use-cases in chapter 4), start to look messy very quickly, which very often requires strong filtering of the data set. This is why effective filtering techniques are especially important in the field of linguistic data visualization. But again, in order to be able to filter the data, specific algorithms have to be applied, which might be suitable for a specific data set (or a specific research question), but not for another one. Also the fact that linguistic variables can be interpreted as specific "units" on various levels such as on the level of phonemes, morphemes, POS-tags, lemmas, tokens, phrases, sentences or even whole text documents makes a clear definition of the level of the underlying textual data set a necessary preprocessing step. A necessity, which we do not face in other fields that also work with "big data". Furthermore, when seen from a semiotic point of view, a textual representation of a concept is in fact already an abstract visual representation of it. Textual characters are highly symbolic in their nature, which makes visualizations of textual data not a first level, but a second level of abstraction. Therefore, visualizing language and/or specific aspects of language requires many pre-processing steps and theoretical decisions which influence the type of algorithms that can be applied to the data, but also the visualization forms in order to graphically represent either the data set itself or the computed outcomes. Nevertheless, because human language *is* a highly structured system, the detection of patterns and clusters is of course possible. One only has to correctly (and adequately) apply research questions and techniques.
However, even though adult language production is (normally) highly structured, for child language this often does not seem to be the case. Toddlers start from producing single sounds, to words to phrases and whole complex sentences in an incredibly fast pace, but how that exactly happens is still a great debate in child language acquisition research. Because the subject of this thesis is to apply and experiment with data visualization techniques in order to find a way to visualize language development in child language acquisition, I will in the next section further describe the challenges one has to face with processing and visualizing child language acquisition data.

3.4. Visualization of Linguistic Development in Child Language Acquisition

The main difference when analyzing (and consequently also visualizing) child language is its big amount of unstructuredness when compared to adult language, and also the speed at which a child goes from making single sounds to uttering its first words to speaking in fully grammatical sentences.
To what extent either *nature* or *nurture* influences the development of first language acquisition in children has been a much disputed debate for many decades not only in the field of first language acquisition research, but also in psychology: Is language an innate skill of every human being, or is it the result of various influencing factors in the child's environment (Hoff [2013, 17])? In Chomskian *generativist* research tradition, the ability to learn a language is innate for every human being. Specific areas in the brain only have to be activated in order for a child to learn how to apply language-specific rules to produce grammatically correct utterances (Hoff [2013, 13]). Whereas the *generativist* approach has been very well received in syntax research (cf. Tomasello [2000, 2017]), there are other linguistic areas where, nowadays, a more *functionalist* approach is being followed by assigning more weight to the influence of various factors in a child's environment on it's language development. The following quote of Tomasello (2000) illustrates very well the seemingly paradox ability of language speakers to act, on the one hand, conventionally, to use already existing patterns and rules, but also to creatively invent new concepts in order to ensure communicative acts.

> To become a competent speaker of a natural language it is necessary to be conventional: to use language the way that other people use it. To become a competent speaker of a natural language it is also necessary to be creative: to formulate novel utterances tailored to the exigencies of particular communicative circumstances. (Tomasello [2000, 2010])

According to Tomasello (ibid), researchers have tried to explain this paradox ability since the beginning of cognitive science. Because of the fact that already very young children are able to produce multiword expressions, which they themselves have never heard before, many researchers from the generative research tradition have postulated that children operate with an adult language competence from very early on (ibid). This so-called *continuity assumption* (ibid) is, according to Tomasello, in many respects the fundamental theoretic postulate of the generativist approach to child language acquisition. This has been the case for a long time because generativist researchers thought that only this theory was able to sufficiently describe

the language of infants with an adult formal grammar (ibid). However, new studies which included more and more data to their analyses have shown that children do produce new (multiword) utterances, but that this creativity is also more limited than initially thought. This insight has, amongst others, lead to the fact that in today's language acquisition research, one does not anymore exclusively act on the assumption that language is innate, but that linguistic, social, cultural and cognitive *nurture* is given much higher importance as it was the case at the beginnings of language acquisition research (Hoff [2013, 18]). The creation and consequently also the availability of an increasing number of language corpora containing longitudinal language acquisition data of small children and their adult care takers, allow us now to cross-linguistically investigate the development of various linguistic features for an increasing number of children and languages. Transcripts from recordings in which children were mostly recorded while playing with their adult peers allow furthermore a comparison of the data from a socio-contextual perspective as the recording situations are in most cases very similar.

In the following sections, I will present three experimental studies that have been conducted together with Jekaterina Mažara and Prof. Dr. Sabine Stoll for this thesis in order to reflect on the theoretical background from data visualization theory and also to experiment with visual representations of the development of verbal morphology in five children and their adult care takers from the Russian subcorpus of the ACQDIV project. In a first part, I will introduce the theoretical background in the field of child language acquisition and Russian verbal morphology, with focus on the acquisition of the verbal aspect. Then I will present the three visual studies in more detail and analyze and discuss them from a *visual linguistics* perspective.

4. ACQDIVZ I – Visualizing Development In Russian Verbal Morphology

4.1. Theoretical Background

The fact that the correct usage of the Russian verbal aspect shows high complexity in morphology, semantics and pragmatics, makes the verbal aspect a linguistic category that is not innately available at the beginning of a language acquisition process.[1] Rather, the acquisition of the verbal aspect is influenced by various factors, such as *Aktionsart*, verbal morphology, complexity of the discourse, as well as narrative competence of the speaker (Stoll [2001]). This is for example why, according to Stoll, the developmental pattern for verbs with a telic aktionsart is different from a verb with an ingressive aktionsart. Furthermore, the frequent occurrence of certain aspectual forms in certain contexts is a factor which influences the acquisition of the verbal aspect.

4.1.1. Grammatical and Lexical Aspect

According to Stoll, there is no generally accepted definition of the Russian verbal aspect. The only universally accepted characteristic is that there is a perfective and an imperfective aspect, where aspect – in Slavic aspectology – is usually considered a binary category, where each verb form belongs to either the ***perfective*** or ***imperfective aspect***.[2] In Russian, the verbal aspect is generally considered a temporal category. However, in comparison to tense, aspect has no temporal deictic com-

[1] According to Stoll [2001], the acquisition of the verbal aspect is often not terminated even at the age of 6 years.

[2] Note that there exist also a few biaspectual verbs, where one verb form can in some cases bear a perfective, in other cases an imperfective meaning (Stoll [2001, 27])

ponent, but is used to emphasize the temporal structure of the action it describes (Stoll [2001, 27]) or even carries pragmatic meaning (Karavanov [2008]). Furthermore, there is another dichotomy between ***grammatical*** versus ***lexical aspect***. In Russian, the grammatical aspect is labelled by markers for perfectiveness and imperfectiveness (such as suffixes, pre- and postfixes). The lexical aspect, however, is inherently present in the meaning of a verb in a given context. One of the best known categorizations of lexical aspectual forms comes from Vendler [1957], where he differentiates between *achievement, accomplishment, activity* and *state* verbs[3], which differ in their telicity, punctuality and dynamics[4] (Shirai and Andersen [1995]).

	state	**activity**	**accomplishment**	**achievement**
punctual	-	-	-	+
telic	-	-	+	+
dynamic	-	+	+	+

Table 4.1.: Lexical Aspects According to Vendler (1957)

The Sentences 4.1 to 4.4 include examples for verbs denoting *state, activity, accomplishment* and *achievement* in Russian:

(4.1) *Lera* ***ljubit*** *Mishu.* *(state)*
F.SG.NOM.AN love.IPFV.NPST.3SG.IRREFL M.SG.NOM.AN
'Lera loves Misha.'

(4.2) *Olja* ***chitala*** *knigu.* *(activity)*
F.SG.NOM.AN read.IPFV.PST.F.SG.IRREFL F.SG.ACC.INAN
'Olja was reading the book.'

(4.3) *Masha* ***vyuchila*** *tancevatq.* *(accompl.)*
F.SG.NOM.AN lern.PFV.PST.F.SG.IRREFL dance.IPFV.INF.
'Masha has learned to dance.'

(4.4) *Leonid* ***postroil*** *dom.* *(achiev.)*
M.SG.NOM.AN build.PFV.PST.M.SG.IRREFL M.SG.ACC.INAN
'Leonid has built a house.'

Normally, when investigating the characteristics of the Russian grammatical verbal aspect (i.e. perfective vs. imperfective), it is often studied together with the lex-

[3] *Achievement* verbs do not last in time and are therefore instanteous (e.g. *knock*), *accomplishment* verbs consist of a process as well as of an outcome (e.g. *build a house*), *activity* verbs are lasting in time (even if only shortly) and *state* verbs are stative and durative and the information if there is an endpoint is irrelevant (Kibort [2008]).

[4] *Telicity* refers situations and processes which are leading up to a terminal point, *punctuality* refers to situations which are not conceived as lasting in time and *dynamic* refers to situations which change over time (Kibort [2008]).

ical aspect of the verbs (cf. Filiouchkina [2005], Gagarina [2000], Gvozdev [1949]). However, as the verbs in our corpus are not coded for lexical aspect information (and also because the focus of this thesis lies in discussing various visualization possibilities for illustrating the development of the grammatical aspect), I will only consider the distribution of the grammatical aspect with respect to its combination with finite[5] and non-finite (infinitive) verb forms. The theoretical background for this visualization study comes from the *distributional bias hypothesis* (Shirai and Andersen [1995]) and will be explained in more detail in the next section.

4.1.2. Aspect- and Distributional Bias Hypothesis

Verbal aspect has become an important part in language acquisition research since the 1980s (Stoll [2001]). One of the most important findings with respect to child language acquisition and verbal aspect was the correlation between aspect, tense and aktionsart. For example, according to Stoll [2001], there is a strong correlation between the usage of the telic aktionsart with the perfective aspect (in the past tense), and the usage of the durative aktionsart (*states* and *activities*) and the imperfective aspect (in the present tense) in many languages such as English, French, Italian and Greek. Gagarina [2000] has investigated the usage of the perfective and imperfective aspect in early child language acquisition data. In her study, Gagarina found out that children start to use both aspectual pairs relatively at the same time, but also that the biggest part of the utterances were simple, imperfective verb forms in their present tense (Gagarina [2000]). All verbs that were uttered in the past tense exhibited a telic aktionsart. These results confirmed the findings of Gvozdev (Gvozdev [1949]), namely that both aspectual pairs are used from the beginning, but that they are biased in their distribution: imperfective verbs tend to be uttered in their present tense, whereas telic perfective verbs tend to be uttered in their past tense (cf. Stoll [2001, 11-13]). With regards to the biased distribution of the perfective and imperfective verbal aspect, Shirai and Andersen [1995] introduced already in the 1990s the so-called *Aspect Hypothesis* in first language acquisition:

> Children first use mainly verbs in their past tense with *achievement* and *accomplishment* verbs. In languages which differentiate a progressive aspect, children use the progressive marker mainly with *activity* verbs. This marking is then also extended to *accomplishment* and *achievement* verbs in a later stage. A (wrong) overgeneralization by marking also *state* verbs with the progressive lexical aktionsart does not happen.

[5] Tense (past vs. non-past) and mode (imperative).

According to Bickerton (cf. Shirai and Andersen [1995]) children are able to distinguish between state and process from very early on. This was corroborated by results from multiple studies where children only very rarely made the mistake to mark *state* verbs with progressive markers. However, Li and Shirai (1995) note that these usage tendencies could also ground in the fact that children simply take over these tendencies from their adult peers, as tense and aspect is similarly distributed in adult speech, where the perfective aspect is rather used with *accomplishment* verbs and the imperfective aspect with *activity* verbs (Shirai and Andersen [1995]). In their study from 1995, Shirai and Andersen [1995] recorded three English-speaking children of an age range from 1;6[5] to 4;10 years for longer periods of time while playing at home with their mothers. The focus of this study was the question of how the linguistic utterances of the mothers (which serve as input to the children) and the utterances of the children (the output) are going to change in the course of time. After the recordings were finished, the data was annotated with regards to the grammatical and the lexical aspect, in order to see if the *Aspect- and Distributional-Bias-Hypothesis* can be confirmed with the data of the target children of this study. Overall, Li and Shirai analyzed 3370 verbs according to pre-defined criteria and assigned them to one respective lexical aspect.
The results of their analysis showed that the mothers used verbs in the past tense predominantly with *achievement* verbs and verbs denoting duration with *activity* verbs. The children in this study generally also used the progressive form first with *activity* verbs (then also with *iterative achievement* verbs). The children used verbs in their past tense in a first phase (first to second year) almost exclusively with *achievement* verbs, whereas the mothers used the past tense also with verbs of other lexical aspects. In their study, Shirai and Andersen showed that in English, the progressive form is learned first with verbs in their present tense, and verbs in their past tense are first mostly used with verbs having telic meaning. These findings support the *Distributional-Bias-Hypothesis*, as both grammatical aspects (perfective vs. imperfective/progressive) were used with the tense form which is most commonly used with either of the aspect forms (Shirai and Andersen [1995, 757-760]).

A further study where the correlation between verbal aspect, tense and aktionsart was investigated is Filiouchkina [2005]. In her study, Filiouchkina looked at which of the two factors – tense or aktionsart – plays the most important role in acquiring verbal aspect. The data in her study was also annotated for tense, grammatical aspect, finiteness, usage of auxiliary verbs and then classified as belonging to a certain lexical aspect. Filiouchkina showed in her study, that language-specific mor-

[5]Read one year, six months.

phological, syntactic and semantic characteristics do have a great influence on the acquisition of the verbal aspect in English and Russian.

Concerning the *Distributional-Bias-Hypothesis*, one part of this thesis consisted of experimenting with various visualization methods in order to make potential distributional patterns with respect to the combination *aspect-tense* visible. Other than simply plotting the distribution of aspect and tense by only using abstract numbers, the challenge lay in creating visual representations which on the one hand use frequency distributions in order to highlight potential patterns, but which also enable the user to concentrate on a specific verb form if desired, thereby keeping the "big distributional picture" of all the other verb forms in the background. In the following sections, I will present three visualization forms which have been created from morphological information of verbs from the Russian corpus of the ACQDIV project. I will first briefly present the underlying data set and then analyze the three visualization forms based on the following factors which were discussed in chapter 3: the type of the data, how the visualization uses visual clues to facilitate human cognitive perception, parameters from Gestalt theory and Diagrammatik. Furthermore, the whole process of generating the visualizations will also be discussed from a "visual analytics" perspective. All of the below presented visualizations are available as an interactive HTML website. Because the focus in this thesis lay on conceptualizing and creating the visualizations, their full potential for further research is still to be exploited. Figure 4.1 and Figure 4.2 show two example screenshots from the interactive website.

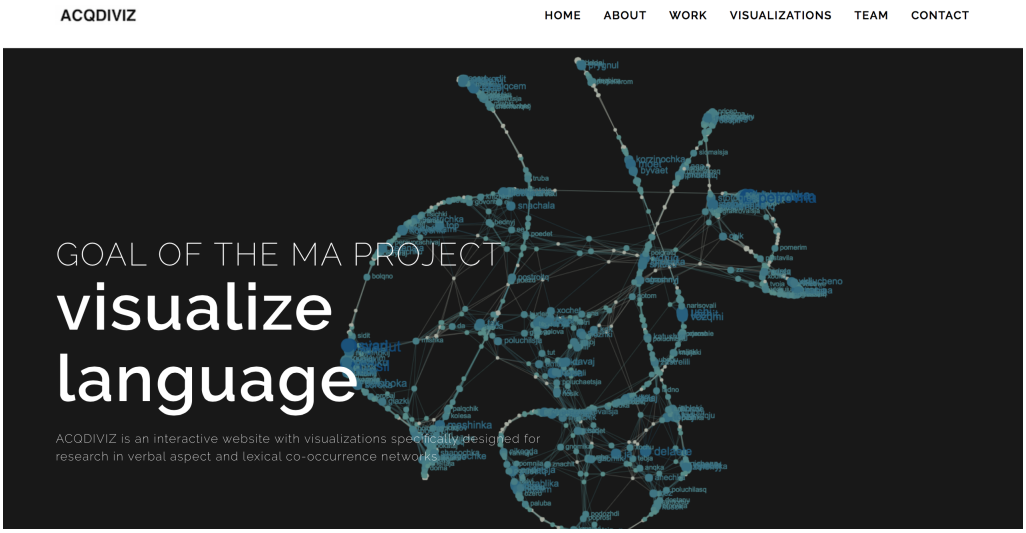

Figure 4.1.: Screenshot ACQDIVIZ: Main page

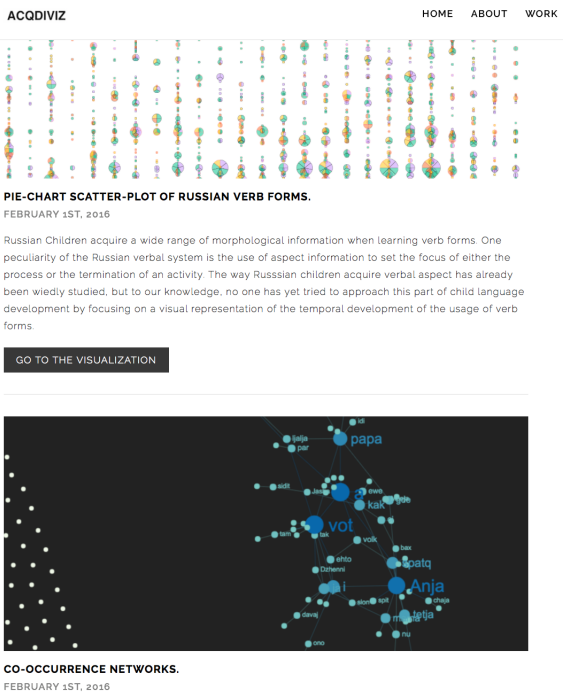

Figure 4.2.: Screenshot ACQDIVIZ: Visualizations

4.2. Data Set

The data for the fist ACQDIVIZ use-case come from the **Russian language acquisition corpus** which was compiled for the work of Stoll [2001] and is also known as the *Saint Petersburg Language Acquisition Corpus* (SPLAC). This corpus contains recordings of five children, ranging in age from 1;3.26[5] to 6;8.12 years. The sessions for this corpus were mostly recorded in weekly intervals, where the children were videotaped while playing at home with their mothers and/or other caregivers. Table 4.2 gives a global summary of the data used with respect to the number of utterances, the number of tokens as well as the number of verbs used for each child and all the adult speakers which were present in the sessions for each child.[6]

[5]Read one year, three months and 26 days.

[6]Note that for this study, all the adult speakers were pooled as one group of *child-directed speech* (CDS) because we also wanted to get as many data as possible from the adult speakers surrounding a child in order to see how all the adult caregivers from a child utilize the verbal aspect. In the second use-case, where we investigated the growth of lexical co-occurrence networks, we followed another approach by only taking the main caregiver of a child as a reference (which were the mothers in Russian but again all the adult speakers in Chintang because in the Chintang culture the upbringing of a child is far more commune- than mother-centered). These points will be discussed in more detail in chapter 5.

Speaker	Age	Nr. of Utterances	Nr. of Tokens	Nr. of Verbs
RUChild1	1;6.10 - 5;4.18	124269	543244	39609
RUChild2	1;3.26 - 4;11.0	45420	662743	10886
RUChild3	3;1.8 - 6;8.12	64711	642479	29378
RUChild4	1;11.28 - 4;3.14	43973	329522	16880
RUChild5	1;4.22 - 5;6.26	53508	506678	5529
RUChild1 adults	-	117364	300898	62333
RUChild2 adults	-	146738	420989	86038
RUChild3 adults	-	140465	404782	68038
RUChild4 adults	-	76508	219481	43963
RUChild5 adults	-	151283	418735	66763

Table 4.2.: Basic Statistics: RU data

As already stated above, the first use-case of this thesis is concerned in creating a visual representation of the usage of Russian verbal morphological patterns with special respect to grammatical aspect and tense. The first step in creating all of the below visualizations was to choose a way how one can visualize morphological information for Russian. An example for a grammatical tag of the Russian verbs *datq* 'give', *bezhatq* 'run (away)' and *upastq* 'fall down' is given below.[7]

(4.5) *datq*
`IMP.2SG.IRREFL.PFV`
'give'

(4.6) *bezhatq*
`NPST.3SG.IRREFL.IPFV`
'run (away)'

(4.7) *spatq*
`INF.IRREFL.IPFV`
'sleep'

(4.8) *upastq*
`PST.SG.F.IRREFL.PFV`
'fall down'

As can be seen in the examples 4.5 - 4.8, all the verbs are coded for aspect information and reflexivity, however, the morphological information before these two tags highly depends on whether a verb occurs a) in a finite or non-finite form, b) in a certain mode (e.g. imperative) and c) in a certain tense (if the verb form is finite). Therefore, a first step consisted in reordering the morphological tags so that the main dichotomy would be between the two verbal aspects (perfective vs. imperfective), then between finiteness (finite vs. infinite), and then between mode and tense, as well. Furthermore, if a verb occurs in a finite form and either in the past or present tense, further distinction has to be made, because in Russian, past-tense verb forms do not distinguish between person, however they include information for number and gender. What further complicates dividing up past tense verb forms is that also the distinction for gender is only given in the singular form. Plural forms, on the other hand, do not contain the information if a subject or an object the verb refers to is masculine or feminine. Because of the hierarchical nature of the data

[7]Note that the verb we used in our visualizations always contained the lemma, but the grammatical information actually belongs to the actual word form that occurred in the data.

set, our first idea for a visualization was to represent the tags as a tree diagram. In the next section, I will explain in further details the functionality and conceptual background of the morphological tree diagram visualization.

4.3. Morphological Tree Diagram

Because the process of writing the code to re-order the morphological information felt like following a tree-like (i.e. hierarchical) structure, where one would start at the root, then decide to take either the perfective or imperfective branch and then depending on this decision, make its way further until one reaches the actual verb form. A process which functions very much like syntactic dependency graphs, which are in their nature hierarchical structures.[8]. In a next step, following Shneiderman (1996) and Keim's (2008) "visual information-seeking mantra(s)", we clearly wanted to create a visualization which enables to user to see the big picture of all the displayed verb lemmas and their morphological information, thereby enabling the visualization to make full use of its planarity (Krämer [2013]) feature, where the beholder is put in a bird's eye position from where he or she can overlook the structure of the whole data set. Nevertheless, if a user only wants to focus on a specific verbal aspect (or only on a specific tense), we also wanted the visualization to be flexible enough to allow for that. It was also clear from the beginning that we will use the programming language JavaScript to create our visualizations, which can then be used (and manipulated) in a web browser. All of the three visualizations for visualizing verbal morphology have been created by mainly using the `d3.js` (Bostock [2012]) library. Figure 4.3 shows a zoomed version for imperfective, non-past tense irreflexive, as well as perfective infinitive verbs uttered by target child RUChild1 for an age range between 2;.06.21 and 2;.07.29 years.

[8]Dependency graphs will be discussed in further detail in the light of graph structures in general in chapter 5.

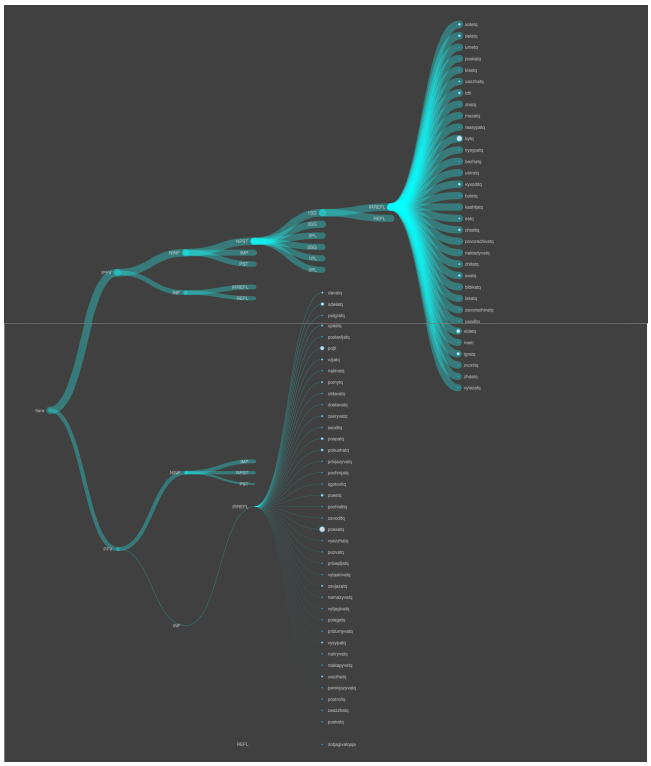

Figure 4.3.: RUChild1: Tree-Diagram of Verbal Morphology

The visualization is created with the d3.js library[9] and constructed in such a way that it starts at the root of the tree (the left-most node) and then divides the morphological tags according to the distinctions described above. The width of the branches proportionally indicates the number of sub-branches. That is, the thicker a branch is, the more sub-branches it contains. Hovering over a branch will highlight it from the other branches and hovering over a node will increase the name of the node and color it differently, as shown in Figure 4.4. As can furthermore be seen in Figure 4.4, the size of the node just before the verb lemma (i.e. the right-most node) indicates how frequently a verb appeared with this particular morphological information. This

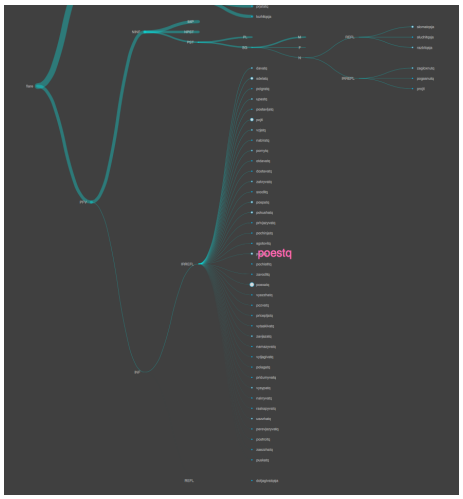

Figure 4.4.: RUChild1: Tree-Diagram Zoom

[9]The source code comes from the d3 example website `http://bl.ocks.org/d3noob/8375092`.

kind of visualization might seem promising at first sight, as it allows for a bird's eye perspective, includes a zooming function and follows the gestalt laws of proximity (morphologically similar items are automatically placed near each other in a tree diagram), similarity, as well as connection. However, as the structure of a tree will always look very similar, and placing more verbs increases the size of the visualization dramatically, we decided to look for an alternative visualization, which would allow us to firstly better detect patterns, and which would secondly not increase in size so that the user constantly has to scroll up and down when looking at the whole data set.

4.4. Morphological Sunburst Visualization

When thinking about another form to visualize the verbal morphology data set, it was clear that we had to move from a visualization placing it's items based on a y-/x-coordinate system to another visualization form which uses a radial system. Using a radial system would, on the one hand, help us to reduce the space needed on the screen, but it would, on the other hand, still allow us to represent the hierarchical structure of our data. Therefore, the next visualization we experimented with was a so-called sunburst visualization.[10] Figure 4.5 illustrates the main components of our visualization. Conserning the user interaction, we allow for choosing various parameters in order to select the data set for one particular target child. We included three drop down menus where the user can choose the corpus[11], select "Age Range" if he or she wants to use a more aggregated version of the data set (i.e. monthly grouped sessions instead of weekly grouped ones) and also choose the target child. A slider can then be used to create the visualization for different time spans (i.e. the monthly grouped data).[12] With respect to the coloring, it can be seen that we colored the visualization in a way that only the information we are mainly interested in stands out and the other information is kept in a lighter tone. We used blue and red for the main distinction between the two verbal aspects in the inner-most circle: red denotes imperfective aspect and blue denotes perfective aspect. Following the Gestalt laws of similarity and continuity (on a color level), we colored the information for past tense also in a blue tone and the information for non-past tense also in a red tone, in order to create a visual relationship to the tense

[10]This visualization was build on the source code by K. Rodden: `http://bl.ocks.org/kerryrodden/7090426`, we adapted the visualization to our data set and also for its zooming functionality.

[11]Note that this visualization currently only works for Russian.

[12]The exact age ranges for every age group in this visualization are given in Table A.1 in the appendix.

and mode information we are interested in seeing if it follows the aspect distribution hypothesis. Because the imperative mode is not really discussed in the literature about verbal aspect acquisition (at least not with respect to the distributional bias hypothesis), we marked it in yet another color. We did not use turquoise as for the other arcs because we thought that it would be interesting to see how the imperative form distributes with respect to the grammatical aspect of the verbs.

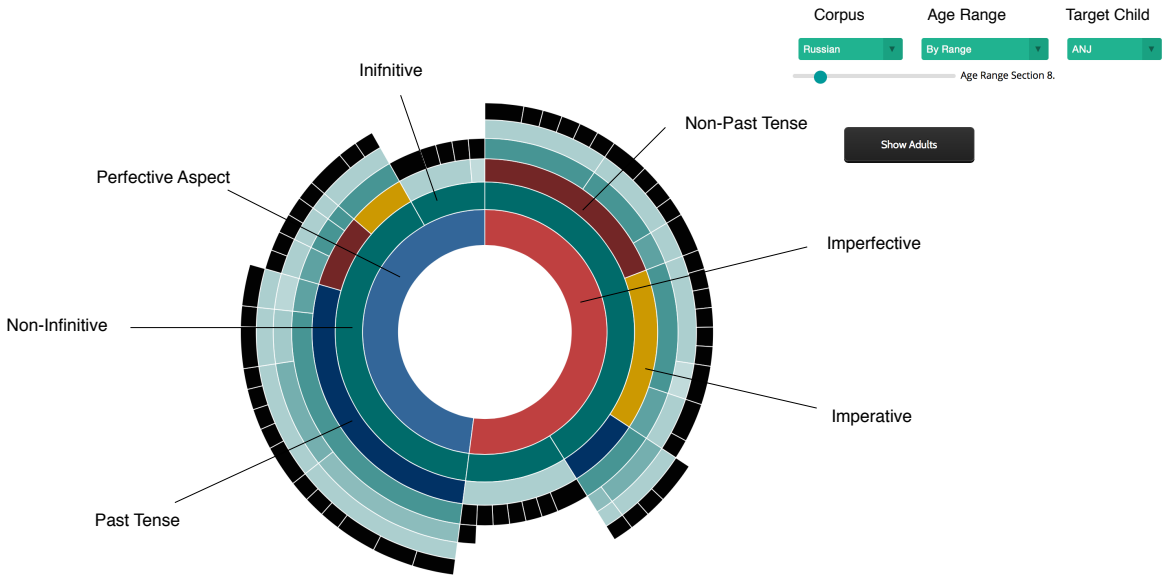

Figure 4.5.: Sunburst Visualization: Legend

The arcs colored in turquoise contain the morphological information that is not important for our distinction. We use darker turquoise to mirror the nesting structure of our data (i.e. darker colored arcs contain more sub-branches than lighter ones based on our partition of the morphological information). We also used a gradient shading for the turquoise arcs in order to lead the observer's eye from the inner arcs to the outer arcs. In order to highlight our arcs of interest even more, we used the opposite direction of shading for our target arcs, so that they would lead the observer's eye from the outer arcs to the inner ones (cf. Figure 4.6).

Figure 4.6.: Sunburst Visualization: Visual Clues

Additionally to the hover functionality, we added a zooming function to our sunburst visualization in order to enable the user to follow the information visualization mantra by choosing "details-on-demand". Figure 4.7 shows the hover, as well as the zooming functionality. Hovering over an arc will display the morphological information up to this arc at the top of the browser window (it will also show in absolute and relative numbers how many verb forms are included). The zooming function was explicitly coded as a transformation process, where the viewer can understand the reordering of the visualization, following the gestalt law of continuity.

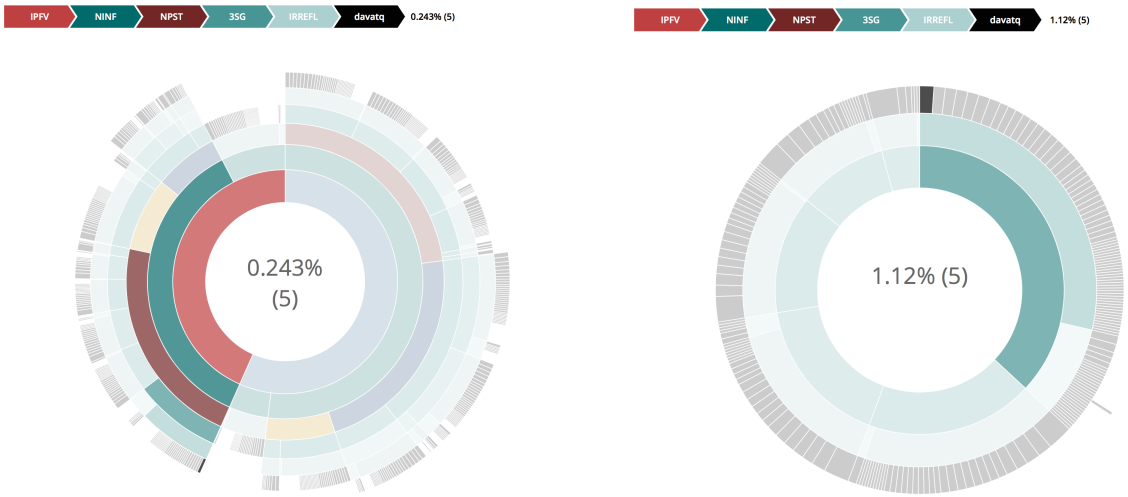

Figure 4.7.: Sunburst Visualization: Hover and Zoom

As indicated with the selection options described further above, one of the main motivations for this visualization was to have a possibility to reduce the morphological information to a very abstract level, but also to allow for comparison between the children and their adult caregivers. By using the slider function, we hoped to be able to detect interesting distributional patterns over time. We wanted to test a reduced version of the distributional bias hypothesis, by only concentrating on the

distribution of the combination *verbal aspect – tense/mode* because this is the information we have in our Russian corpus. Therefore, our hypothesis was that the usage of the verbal aspect in the adult data would favor the combination *perfective aspect – past tense* and *imperfective aspect – non-past tense*, which we expected to find also in the data of the children. We therefore created a so-called "small-multiples" (Tufte and Robins [1997, 105]). We decided to turn our own visualization into yet another visualization form (the small-multiples), because this visualization form is ideal in order to show temporal changes when still keeping the option of seeing how the structure of the whole picture changes over time (that is, it is something like a "static" animation, which allows both, seeing changes over time, but also the data at a specific point in time).[13] According to Tufte, multiple images are the elements which define the whole idea of visualizing information because they can reveal "repetition and change, pattern and surprise" (ibid.) Furthermore, as Tufte states as well, small multiples perfectly use the nature of paper and computer screens, as, even though the pictures are used "only" in a two-dimensional space, small multiples create their depth by "arraying panels and slices of information" (ibid.), which help the beholder to "analyze, compare, differentiate and decide" by "amplifying, intensifying, and reinforcing the meaning of images." (ibid.).

When analyzing the small multiples more closely, one can clearly see in Figure 4.8 that the distributional pattern for RUChild2 with respect to the perfective aspect shows mostly blue-blue patterns, when only highlighting the arcs which contain the most verb forms. However, we can also see that in the first five age stages, child RUChild2 mostly uses the perfective aspect with imperative forms.[14] For the distribution of the perfective aspect also in Figure 4.8, the distribution is also heavily biased towards the red-red pattern, i.e. the combination perfective aspect – non-past verb forms (there are only three age stages, where child RUChild2 uses the imperfective aspect more often with past tense or imperative verb forms).[15]

[13]The "specific point in time" would be a specific age range in our case, to be precise.

[14]Mostly with the verb *datq*.

[15]An interesting fact here might be that these age stages are directly following each other.

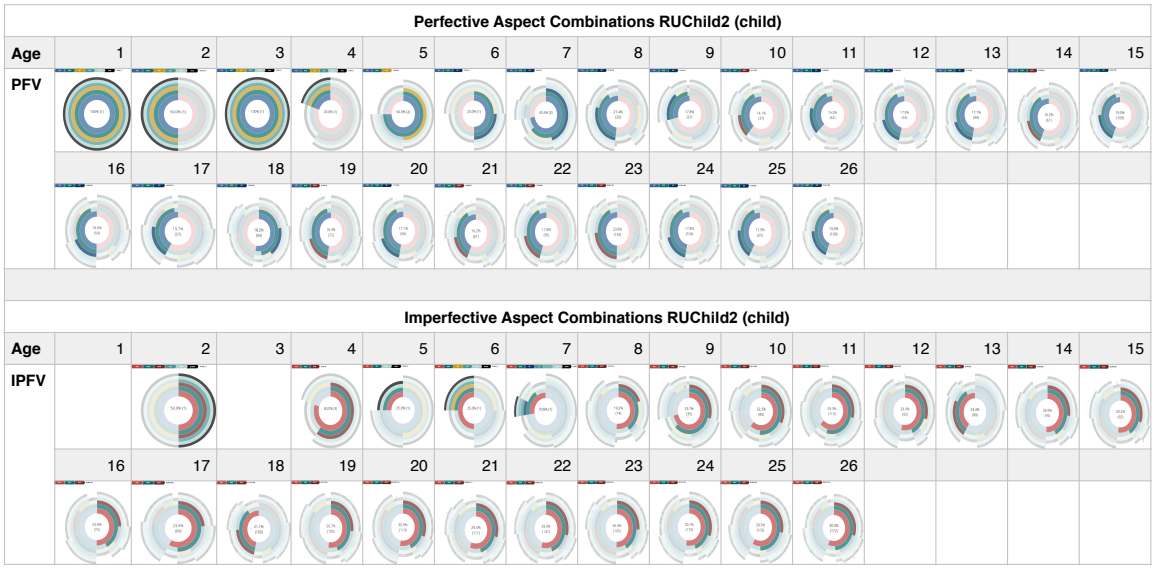

Figure 4.8.: Sunburst Small-Multiples: RUChild2

When comparing the aspect-tense-mode distribution with the adult caregivers of child RUChild2, the patterns for the perfective aspect look quite differently. At first sight, the combinations perfective aspect – past-tense and perfective aspect – non-past tense seem quite evenly distributed. In fact, in 9 of the 20 age stages the perfective aspect is used with the past tense, and in 17 stages with the non-past tense. This certainly differs with regards to the distribution of the child data. The imperfective aspect, however, is exclusively used with non-past tense verb forms, which follows the distributional bias hypothesis.

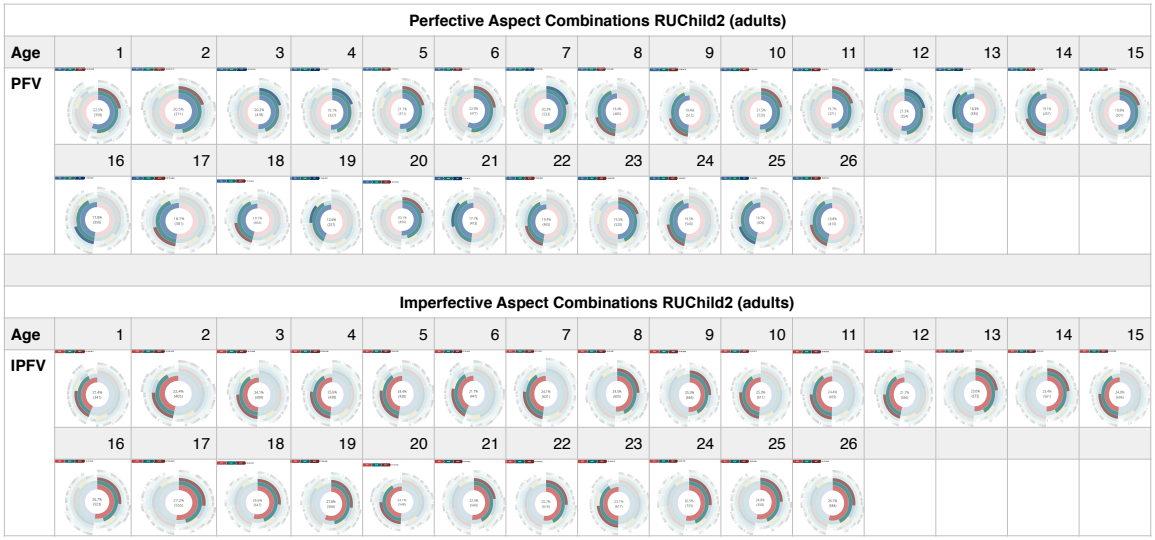

Figure 4.9.: Sunburst Small-Multiples: RUChild2 adults

When comparing to another child (Figure 4.10), the distributional patterns for

the perfective aspect look quite different at first sight, because the data for child RUChild5 does not show a clear pattern change as child RUChild2. But again, when we compare the numbers, we can see that in 11 out of 20 stages, child RUChild5 uses the perfective aspect with the past-tense form, in 5 stages with the non-past form and in 4 stages with the imperative form. This examples illustrates very well the additional benefit we have when using a visual representation of the data: even though the numbers of the distributions are similar, the distributional pattern over time is quite different, a fact which becomes visible only within the data graphic. Considering the imperfective aspect, there is also with child RUChild5 a clear bias towards the usage with the non-past tense form.

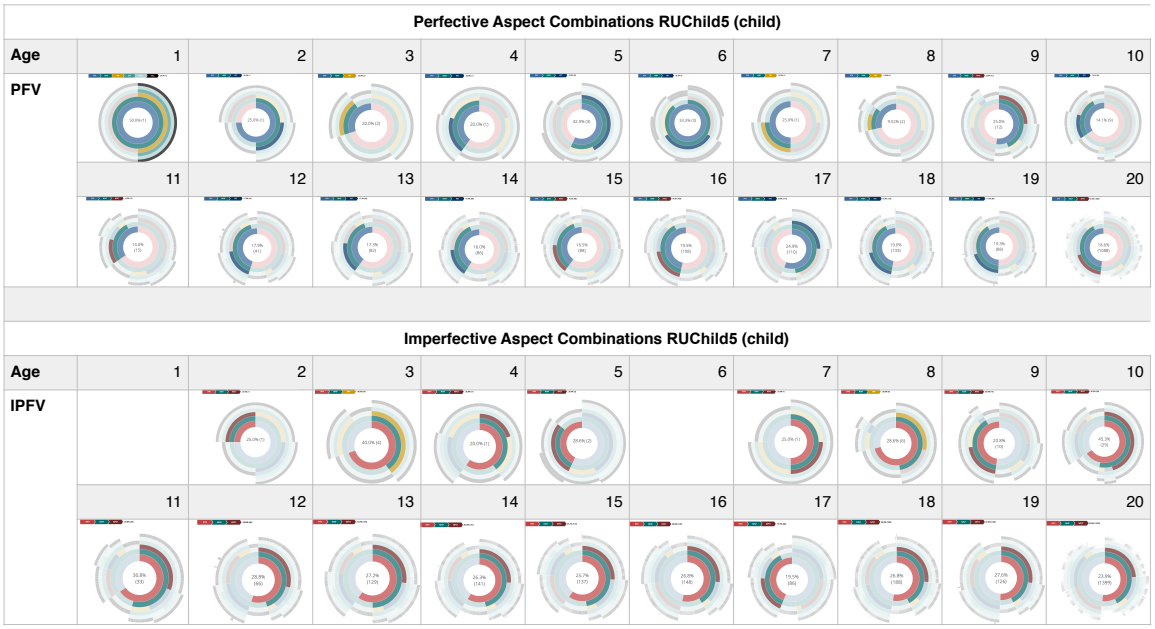

Figure 4.10.: Sunburst Small-Multiples: RUChild5

When again comparing with the adult caregivers of child RUChild5 (Figure 4.11), there is a clear tendency detectable that the adult speakers use the perfective aspect predominantly with past tense verb forms (in 15 out of 20 age stages), but that there are also stages, when the adult speakers used the perfective aspect more with non-past tense verb forms. Considering the imperfective aspect, we can again see a clear bias towards using it with non-past verb forms (in 20 out of 20 age stages).

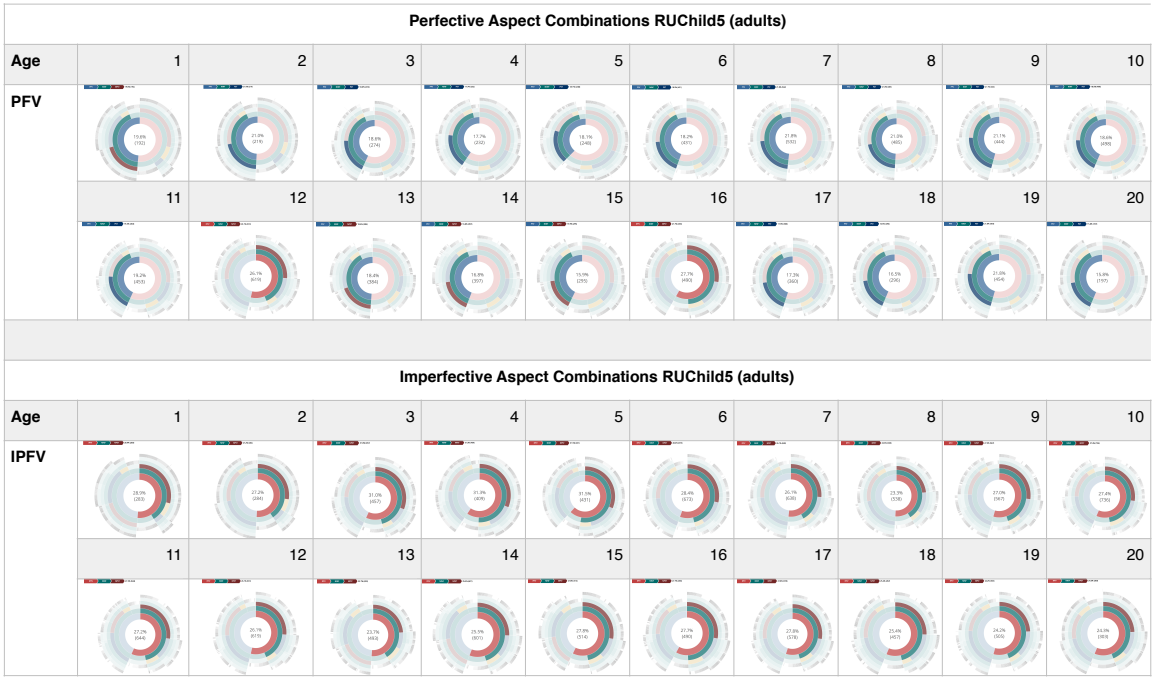

Figure 4.11.: Sunburst Small-Multiples: RUChild5 adults

Of course, principally, one could display this whole data also in a table (as in Table 4.3), where one could as well detect the above described tendencies. However, how these combinations are distributed over the age stages only becomes apparent with the visualization.

Aspect-Tense-Mode Distribution								
Speaker	**PFV-PST**	**PFV-NPST**	**PFV-IMP**	**PFV-INF**	**IPFV-PST**	**IPFV-NPST**	**IPFV-IMP**	**IPFV-INF**
RUChild2	57% (15)	23% (6)	20% 5	0% (0)	4% (1)	88% (21)	4% (1)	4% (1)
RUChild2 adults	34% (9)	66% (17)	0% (0)	0% (0)	0% (0)	100% (26)	0% (0)	0% (0)
RUChild5	55% (11)	25% (5)	20% (4)	0% (0)	0% (0)	89% (16)	11% (2)	0% (0)
RUChild5 adults	75% (15)	25% (5)	0% (0)	0% (0)	0% (0)	100% (20)	0% (0)	0% (0)

Table 4.3.: Distribution Aspect-Tense-Mode

With regards to the distributional bias hypothesis, our data clearly confirmed the pattern that the children are using the imperfective aspect predominantly with non-past tense verb forms (as it is also the case with the data of the adult speakers). Concerning the perfective aspect, the distributional bias hypothesis seems to hold true for child RUChild5 and his caregivers, as the most frequently used combination is perfective aspect – past tense verb forms. However, this does not hold for child RUChild2 and her caregivers, as they predominantly use the perfective aspect with the non-past tense forms. Child RUChild2, on the other hand, mostly uses it with past tense verb forms.

4.5. Pie-Chart Bubble-Plots

Even though the sunburst visualizations proved to be useful to detect patterns with respect to the aspect-tense-mode distribution, it still did not fully satisfy us, because it did not make obvious one major point in child language acquisition: the temporal aspect of when a verb enters the vocabulary of a child for the first time. Therefore, we conceptualized yet another diagram where we wanted to visualize a) the distribution of all verbs in a child's (and his caregivers') vocabulary over time, as well as b) the growth in different verb lemma forms per verb over time. Following the visual analytics pipeline (cf. Figure 3.9), experimenting with the previous visualizations increased our knowledge about the data, and only this experimentation process led us to realize how multi-leveled visualizations of linguistic data in practice have to be. Consequently, by combining the insights gained from our previous visualizations, we knew we wanted to create a graphic which shows us "everything we want". JavaScript proved to be an excellent choice, because it allows for data manipulation on many levels, and using the browser window as a "canvas", which can be "painted at" as much as we wanted, allowed us to create the last form of our visualizations as product of all our creative experiments.

For our last data graphic, we decided to combine two diagrams into one, which would allow us to realize our vision of showing development, as well as pattern detection over time. Figure 4.12 shows the logic behind our final visualization.[16]

Figure 4.12.: VerbsPie-Scatter-Plot: Schema

[16]This visualization was also created using the `d3.js` as well as the `dimple.js` JavaScript libraries.

By combining a scatter-plot with a piechart visualization, we used the x-axis to visualize the verbal development per verb lemma over time, with respect to the number of different verb forms used, coded with different colors for aspect, tense, mode and finiteness information. The ticks in the x-axis correspond to the age of a child during the various recording sessions.

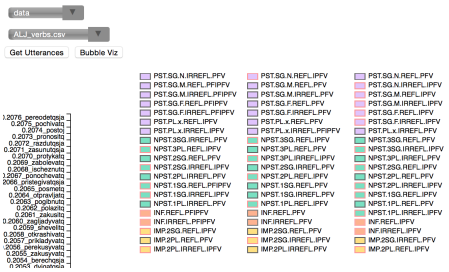

Figure 4.13.: VerbsPie: Legend

The size of the piechart circles is determined by the number of different verb forms in which a verb lemma occurs per session. Following the gestalt laws of connection and proximity, the viewer is enabled to concentrate on two levels simultaneously: if one wants to see how many different verb forms a child uttered in a session, the y-axis will be the main focus, where the verb lemmas are ordered according to their first appearance in the vocabulary of a child (i.e. when a child uttered a new verb for the first time).[17] If, on the other hand, one is more interested in the development of a certain verb lemma, then the focus can be shifted to the x-axis. Different coloring is used to again summarize the data into the two groups *perfective* vs. *imperfective* aspect. Figure 4.13 shows the color coding used, where tense, mode and finiteness is used as the inside color, and red (again for the imperfective aspect) and blue (perfective aspect) is used as the edge color to mark the aspect of a verb form.

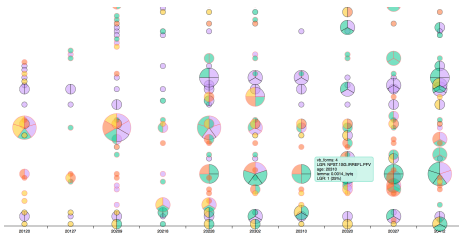

Figure 4.14.: VerbsPie: Hover

In order to facilitate the orientation in the visualization[18], a tooltip showing the number of different verb forms, the age of the child and the verb lemma is included when hovering over a segment in a piechart (cf. Figure 4.14). The button "Get Utterances" can then be used to go to an interface where the ACQDIV database can directly be queried for example utterances for a specific child, his or her age range as well as a specific verb lemma. What is more, in order to facilitate

[17] Note the numeration of the verbs in Figure 4.13, this was unfortunately a necessary step in order to arrange the verb lemmas correctly.

[18] Unfortunately, one big drawback is that the user must constantly scroll from left to right or up and down in order to follow the pattern of interest. However, this can be helped when using the zoom-out/minimize window function of the browser window. Additionaly, a PDF file with custom size can be exported from the visualization, where then again zoom or crop functions can be used to focus on specific parts.

the detection of verbs that were used in many different forms (as opposed to verbs which only appeared in one or two verb forms), we added a last column to the visualization which serves as a summary of the various verb forms a lemma appeared in within the data set.

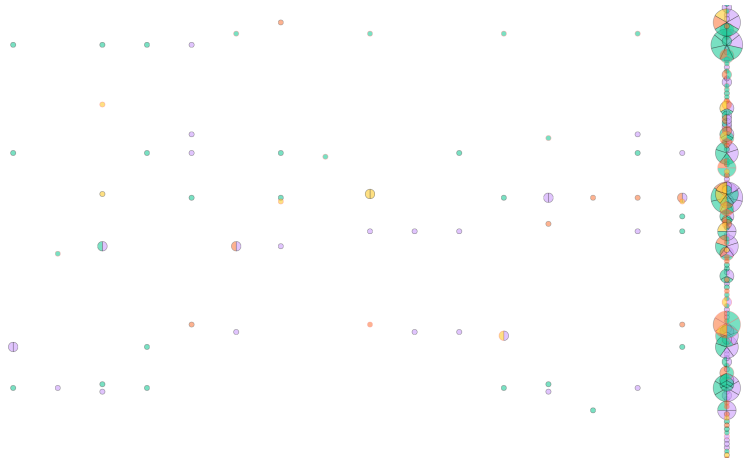

Figure 4.15.: VerbsPie: Summary Column

5. ACQDIVZ II – Language Acquisition & Network Theory

Graph theory, often also called *network theory*, is an area of study in mathematical science used to model relations between objects. It is said to have been invented in 1739 by Leonhard Euler when he developed a technique to prove that it was impossible to cross the seven bridges in Königsberg without crossing any bridge twice (Euler [1741]).[1] Euler's groundbreaking innovation was his idea of turning a mathematical problem into a graph, where the land between the bridges denote the nodes and the bridges themselves serve as links between those nodes (Meirelles [2013, 49]). Even though graph theory was invented at that time, the visual representation of mathematical graphs only followed more than a century later.[2] According to Kruja et al. [2001], the earliest graphical representations resembling graphs date from Ancient Egypt times (1400-1366 B.C.) and have the form of Mill game boards which explicitly depict nodes and edges. However, their usage was not motivated by any underlying mathematical theory. Other examples that Kruja et al. mention as early graph drawings include trees of religious figures and their virtues, as well as genealogical trees such as family trees.[3] Graph drawings, as opposed to tree drawings, were, according to Murdoch (1984), used during the Middle Ages to represent and visualize abstract information and were mostly used as "squares of opposition" that were

> "pedagogical tools used in the teaching of logic, particularly the relations between propositions or syllogisms. They were designed to facilitate the recall of knowledge that students already had, and hence did not contain

[1] Note that the term *graph* was not coined by Euler himself, but only 1878 by the English mathematician James Joseph Silvester (Kruja et al. [2001, 277]). Today, the term *graph* is mostly used in mathematics, whereas in computer and social science the term *network* is more frequent. In this thesis, the two terms are used interchangeably.

[2] An interesting fact to note here is that Euler himself did not use any *abstract* visual graph representation to illustrate his mathematical problem (Kruja et al. [2001, 277]). In this article published in 1736, we see an iconic drawing of the seven bridges (Figure 5.1), but no abstract graph representation of it.

[3] Which were commonly used to decorate the atria in Patrician Roman villas, or also served as evidence at court trials to prove one's ancestry (Kruja et al. [2001, 273]).

complete information." (Murdoch [1984], cit. in Kruja et al. [2001, 276])

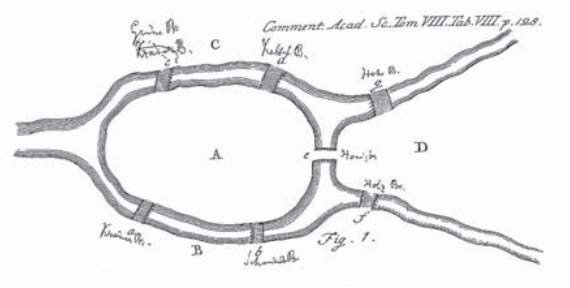

Figure 5.1.: Euler 1936: 7 Bridges of Königsberg (source: Meirelles [2013, 49])

According to Kruja et al., examples of modern graph drawings started in the late 18th century, and only from the beginning of the 19th century, they appeared frequently in many research areas such as mathematics and chemistry (ibid.). Thus, they were not anymore exclusively used in "pedagogy, exposition and record keeping", but served as "visual aids to solving problems" (Kruja et al. [2001, 283]). However, network theory remained a typical research area in mathematics and physics for many decades. The development of more powerful computers on the one hand, as well as the ground breaking findings from Watts and Strogatz (1998) with regards to the common properties of networks generated from real-world data, and the findings of Barási and Albert (1999), who showed that there are common properties in the structure of these networks which may be responsible for their efficient processing and robustness against failure and damage (De Deyne et al. [2016, 48]), led to an increasing popularity of seeing many other real-world phenomena represented as networks as well.

In this chapter, I will present the theoretical background and research which applies techniques from network theory to linguistic research questions. In section 5.1, I will introduce network theoretical approaches to linguistic and psycholinguistic research questions by outlining their historical development. In section 5.2, I will introduce the statistical parameters that are frequently used when analysing networks from linguistic data, and I will give an overview of various approaches where these parameters have already been successfully applied to language acquisition research. In section 5.4, I will then present our own research, which was conducted together with Dr. Steven Moran, where we analyzed lexical co-occurrence networks of five Russian and four Chintang children and their caregivers by calculating the statistical parameters introduced in section 5.2, as well as by taking a visual approach with the graph analysis toolkit *Gephi* (Bastian et al. [2009]) to study the development of so-called *hubs* from a more dynamic perspective, and also with a tool which would again allow us to follow a visual analytics approach by again applying the information-seeking visualization mantra.

5.1. Historical Development

The idea to use networks as mental representations for linguistic concepts can be traced back to Aristotle, who introduced the notion of semantic networks (Ke [2007, 2]). Apart from philosophy, the notion of networks has also already been present for a long time in the fields of psychology and psycholinguistics. As Mihalcea and Radev (2011) point out, Quillian's (1968) theory of semantic memory and the theory of connectionism by Fodor and Pylyshyn (1988) were initially proposed in cognitive psychology as models for representing human language and reasoning, but they were soon also used in a variety of applications in computer science (Mihalcea and Radev [2011, 3]). But the first paper to have a wide impact on the application of network theoretical approaches for linguistic research questions was the above mentioned work of Quillian. In his work, Quillian introduced the idea that linguistic concepts are organized in a network fashion. He was the first to demonstrate, how semantic networks constructed from words and their relationships [4] can be used to successfully identify semantic relations and (multiple) meanings of words. What is more, he also showed that such semantic netowrks can even be used to generate whole text passages (Ke [2007, 3], Mehler et al. [2015, 11]). Nowadays, applications to NLP that still profit from Quillian's work are word-sense desambiguation, automated reasoning, automatic text generation and summarization, as well as information retrieval (Mihalcea and Radev [2011, 1]).

Even though graph theory, NLP and information retrieval have been well-studied disciplines for a long time, they have not been used in combination in computational linguistics until the 1970s (Mihalcea and Radev [2011, 2]). Only from that time on, people have approached linguistic research questions by combining techniques from mathematics, physics, computational science and linguistics, combining them all in the field of complex network theory. When applying the concept of complex networks to linguistics, the nodes of these networks are constructed from linguistic units which can be everything from single phonemes, morphemes, words and sentences to whole document collections. Network creation and analysis in linguistics has recently seen an increasing popularity and research has been conducted in order to capture dynamics and global properties of complex networks on various linguistic levels (ranging from the phonological to the semantic and also pragmatic level (Mehler et al. [2015])). Similar to linguistics and psycholinguistics, one of the first areas in computational NLP where a network representation has been used to represent linguistic information, comes again from semantics. An example taken from Mihalcea and Radev [2011, 2] that shows an automatically derived semantic

[4] Which were derived from dictionary definitions.

network from definitional links in a dictionary can be seen in Figure 5.2.

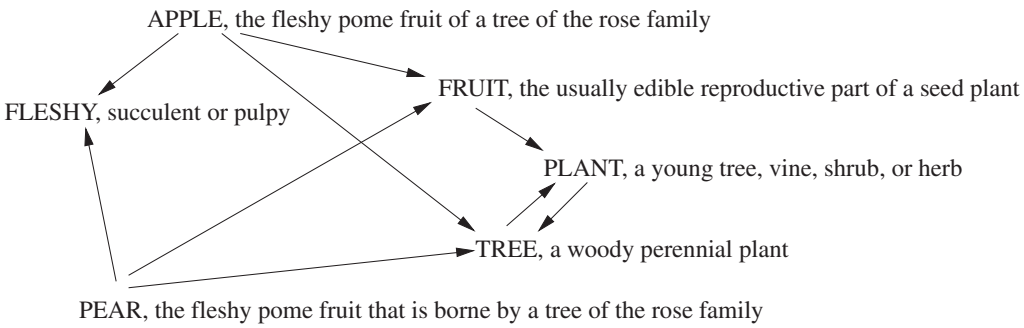

Figure 5.2.: Semantic network derived from dictionary entries (source: Mihalcea and Radev [2011, 2])

This network is constructed by using specific words (concepts) as nodes and creating links between them whenever one of the concepts is used within the definition of another concept. A more modern and interactive version of a semantic network based on lexical semantic relationship stored in the WordNet (Miller [1995]) database is shown in Figure 5.3 for the concept *dog*. In this visualization made for the project *Visuworkds* (Breckon [2015]), nouns, adjectives and verbs are grouped into so-called *synsets* (sets of cognitive synonyms) that each express a distinct concept. Furthermore, the visualization includes additional information by coloring the concepts differently, based on their POS information, as well as by representing different semantic relations using different graphical forms and colors:

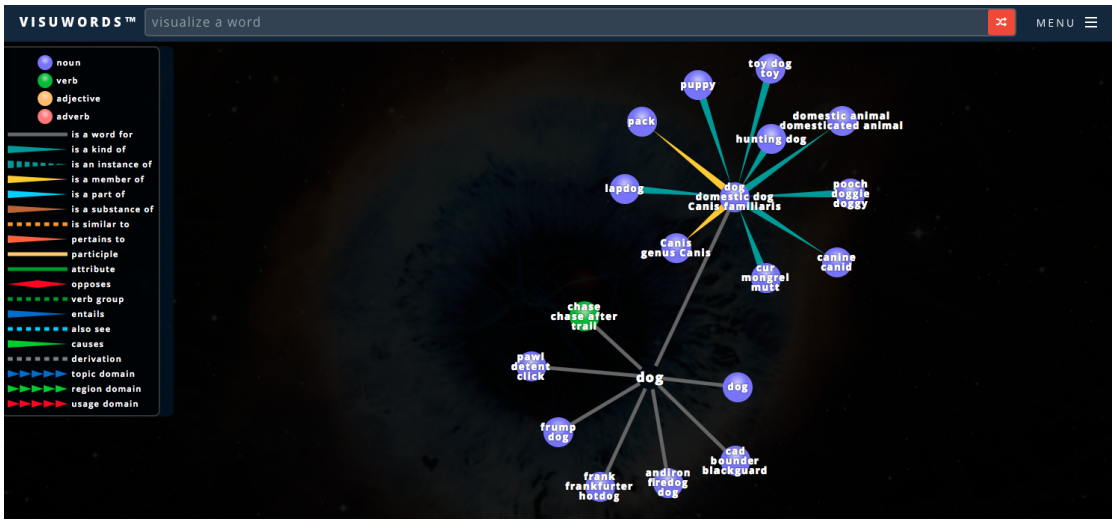

Figure 5.3.: Visuwords Semantic Network

Another early example of a hierarchical representation for linguistic concepts would be graphs constructed from syntactic information. Even though many syntactic

analyses are not based on approaches originating in network theory, tree-like structures to represent syntactic information were principally already introduced in the middle of the 20th century by the French linguist Lucien Tesnière in his famous work *Éléments de syntaxe structurale* (1959). His approach to analyzing the syntax of natural language has served as the starting point for a syntactic theory called *dependency grammar*. Within this theory, graphical representations of dependency graphs are still widely used to visualize the syntactic structure of sentences. Creating a dependency graph includes, next to an initial phase of part-of-speech tagging, syntactic parsing where another annotation layer is added with syntactic information of a word (for example, if a word in a sentence serves as a subject or an object, etc.). Figure 5.4 (taken from Mihalcea and Radev [2011, 142]) shows a syntactic dependency graph for the sentence

(5.1) *The monthly sales have been setting records every month since March.*

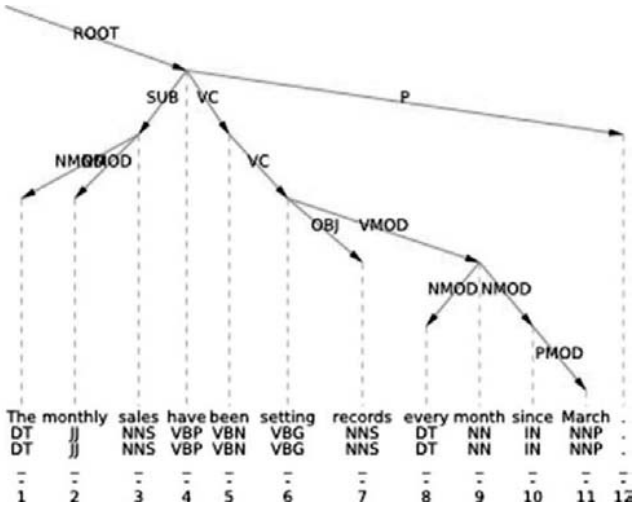

Figure 5.4.: Syntactic dependency graph (source: Mihalcea and Radev [2011, 142])

In this tree-like representation of the above sentence, the method which is used in dependency parsing is also represented visually, i.e the predicate *set* (in its past continues form) is used as the starting point of the analysis and is hence represented at the top level as the root of the dependency tree. *Sales* serve as the subject of this sentence and is further modified by the determiner *the* and the attribute *monthly*. Also dependent on the root, the word *records* serves as the object of the sentence. What can further be seen is that the phrase *every month since March* is used as a modifier of the root.

As we can see, visualizations of linguistic information that use a graph-like (or tree-like) representation have been used successfully and widely in certain linguistic domains. However, there are other domains, where the application of network

theoretical analyses has only been made possible decades later. A major impediment early researchers were facing when applying network theoretic analyses to their data was the huge complexity of the resulting structures, which computers at that time were not able to process. Consequently, research was often conducted on "toy-size problems, and the scalability of the models was rarely evaluated, if ever" (Mihalcea and Radev [2011, 4]).[5] This situation has completely changed in recent years, and the area of graph-based representations and algorithms applied to natural language processing and information retrieval has experienced tremendous growth. Graph-ranking algorithms such as the HITS (Kleinberg [1999]) and the PageRank algorithm (Brain and Page [1998]) have been used successfully on a large scale, for example in the analysis of the network structure of the World Wide Web (Mihalcea and Radev [2011, 4]).

The development in network research has consequently also inspired interests in networks in child language acquisition. Because of the fact that network theory is still not very frequently used within the area of child language acquisition research, I will, in the next subsection, first introduce the most important concepts within network theory, in order to provide the theoretical background necessary to understand the research questions and approaches that have been carried out so far in child language acquisition.

5.2. Theoretical Background

Before starting with the theoretical background, there is one important fact I would like to stress with regards to the graphs and networks analyzed in the following sections. As already pointed out by Vitevitch [2008, 2], the networks discussed in the following sections are not artificial neural networks that are used to model cognitive processes.[6] The graphs that will be presented here do neither have activation states or rules to change them (as this would be the case in artificial neural networks), nor do the links follow certain learning rules to e.g. change the connection weights between nodes. Instead, all of the following networks have been constructed by showing linguistic relations on various levels that have been extracted (and calculated) from the underlying textual data.

[5] For example, in Quilian's work, he evaluated his proposed algorithm for word-sense disambiguation on nineteen ambiguous words (Mihalcea and Radev [2011, 4])

[6] Neither are they "social networks" from linguistic interlocutors of the children.

5.2.1. Graph Properties

A network (or sometimes also called a *graph*) is a data structure that consists of a set of nodes, which are connected by a set of edges that can be used to model relationships among various objects in that network (Mihalcea and Radev [2011, 11]). A graph is therefore defined as set $G = (V, E)$, where $\boldsymbol{V}$ is a collection of ***nodes***[7] $V = \{V_i, i = 1, n\}$ and $\boldsymbol{E}$ is a collection of ***edges***[8] over $V, E_{ij} = \{(V_i, V_j), V_i \in V, V_j \in V\}$ (ibid.). Graphs have certain properties and occur in certain types which allow them to model various relations. In the following subsections, I will introduce the most commonly studied graph properties, as well as the most studied graph types in child language acquisition research. Of course, graphs have many more properties than those listed below, but because the focus of this thesis lies on linguistic research, I will limit the description of graph properties to those which have been used in linguistic analyses.[9]

5.2.1.1. Graph Directedness

Graphs can be either ***directed*** or ***undirected***, depending on whether a diretion in the modeled relationship is defined or not. In a *directed* graph (which is also called a ***digraph***), an adge E_{ij} can be traversed from V_i to V_j but not the other way around. In an *undirected* graph, edges can be traversed in both directions. Moreover, nodes and edges of a graph can also be assigned *attributes* such as a weight (for example to indicate the strength of the relation between two nodes). A graph that has weighted verteces or edges is called a ***weighted graph*** (Mihalcea and Radev [2011, 13]). Figure 5.5 taken from Mihalcea and Radev [2011, 12] shows an example of an undirected **(a)**, a directed **(b)**, as well as a weighted undirected graph **(c)**.

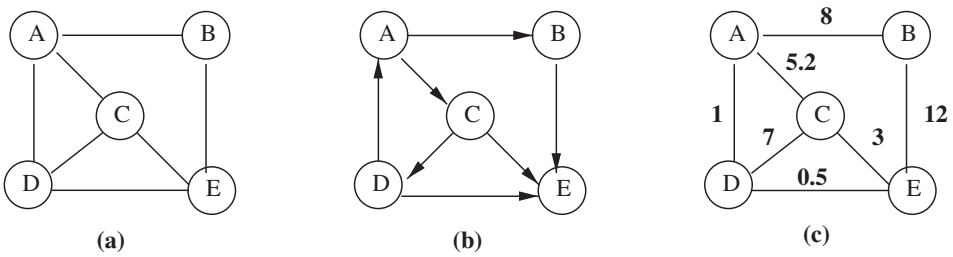

Figure 5.5.: Examples of Graph Structures (source: Mihalcea and Radev [2011, 12])

[7] Also called *vertices.*

[8] Also called *links.*

[9] For an in-depth introduction to graph types and their properties, see Mihalcea and Radev [2011].

5.2.1.2. Node Degree (<k>)

One of the most important properties of a node in a network is its ***degree***, which is defined as the total number of edges coming in or going out from that node. In a directed graph, the degree of a node V_i is split into the ***in-degree*** and the ***out-degree***. The *in-degree* denotes the number of edges coming in to V_i, the *out-degree* is the number of edges going out from V_i (Mihalcea and Radev [2011, 12]). For example, in graph (b) from Figure 5.5, node C would have an in-degree of 1 and an out-degree of 2; node E would have an in-degree of 3 and an out-degree of 0. The ***average degree***, which denotes the average number of edges a node can have in a graph, is calculated by dividing the sum of the number of all edges by the number of all nodes, as shown in the following formula (Where N is the number of nodes in the graph.):

$$<k> = \frac{1}{N} \sum_{i=1...N} Degree(V_i)$$

5.2.1.3. Degree Distribution (P(k))

The ***degree distribution*** of a graph is the probability distribution calculated over the degrees of the nodes in the entire graph, i.e. the proportion of nodes [P(k)] that have a given number of edges (Mihalcea and Radev [2011, 54]). In a degree distribution that resembles a normal bell-shaped distribution (also known as a Poisson or Gaussian distribution), a small number of nodes will have less than the average number of edges per node, and a small number of nodes will have more than the average number of edges per node, but most of the nodes will have an average degree, i.e. the average number of connections per node (ibid.). This type of degree distribution is typically found in *random networks*.[10] As we will see with the degree distribution of the data from our study, many graphs created from real-world data have a degree distribution that resembles a power-law, which is mathematically defined as $P(k) = k^{-\gamma}$ where γ is a constant typically ranging from $2 < \gamma > 3$ in scale-free networks (Mehler et al. [2015]). A network with a degree distribution that follows a power-law means that only very few nodes have a high degree, whereas most of the nodes in such a network have small degrees (Vitevitch [2008, 5]). This kind of degree distribution is typically found in networks created from real-world data (Barabási and Albert [1999]). A typical (bell-shaped) degree distribution of a random graph is shown in Figure 5.6, a degree distribution following a power-law is shown in Figure 5.7 from Mihalcea and Radev [2011, 66-67], plotted on a log-log

[10]Graph types, including random networks will be explained in more detail in chapter 5.2.2.

scale as the function P(k) (y-axis) with respect the the value of degree k (x-axis).

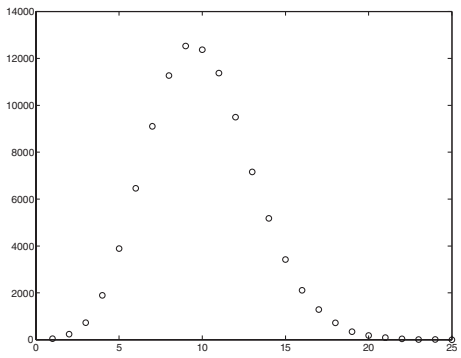

Figure 5.6.: Normal Degree Distrib.

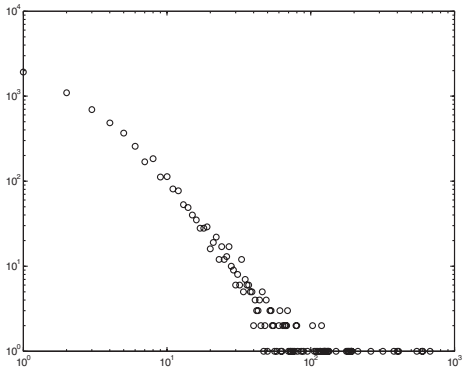

Figure 5.7.: Power-Law Degree Distrib.

Because nodes in networks with a power-law degree distribution do not have a "typical degree", such networks are also called ***scale-free*** (Mihalcea and Radev [2011, 67]).

5.2.1.4. Connected Components (NCC)

A ***subgraph*** of graph $G = (V, E)$ is defined as a graph $S = (V_S, E_S)$, in which V_S and E_S are subsets of V and E and E_S contains edges connecting nodes in V_S. Conversely, when graph S is the subgraph of graph G, G is also called the ***supergraph*** of S (Mihalcea and Radev [2011, 13]). A maximally connected subgraph derived from a supergraph is then called a ***giant (or largest) connected component*** (ibid.). A giant connected component is a subgraph where every node can reach every other node in the network. This structure is so tight, that adding any additional node will violate this property (ibid.). Because many graphs created from real-world data are weakly connected in their global structre[11], there are certain calculations which are done on the giant connected component of a network, because they cannot be computed on weakly connected networks.

5.2.1.5. Clustering Coefficient (CC)

Given a node v_i in a graph and the set of its neighboring nodes that are connected by an edge to v_i, the ***clustering coefficient of a node*** v_i is defined as the number of all existing edges between all of the nodes in this neighborhood (i.e. between the

[11] That is, they have many nodes that are either not connected at all in the network, or only with very few nodes

node and its neighbors, as well as between the neighbors themselves) divided by the total number of possible edges in the neighborhood of a node (Mihalcea and Radev [2011, 13]). In other words, the clustering coefficient of a node measures how well it's neighbors are connected to each other (Ke [2007, 29]). The values for the clustering coefficient of a node range between 0 and 1. A clustering coefficient of 0 means that none of the neighbors of a node is connected to the other neighbors. A clustering coefficient of 1, on the other hand, means that all of the neighbors of a node are connected to each other. Values for the clustering coefficient ranging between 0 and 1 imply that a number of neighbors of a node are also neighbors of each other (Vitevitch [2008, 4]). The ***clustering coefficient of a graph*** is defined as the average of the clustering coefficients of all nodes in the graph.[12] Mathematically formulated, the average clustering coefficient of a directed graph is the average of the local clustering coefficients of all the nodes in a graph. Therefore, we first compute the local clustering coefficient of every node (v_i) with the following formula:[13]

$$C_i = \frac{|e_{jk} : v_j, v_k \ni N_i, e_{jk} \ni E|}{k_i(k_i - 1)}$$

Where the clustering coefficient (C_i) for a node v_i is given by the proportion of links between the nodes in iths neighborhood, divided by the number of edges that could exist between them (i.e. in a neighborhood N_i of node v_i, there could be $k_i(k_i - 1)$ edges between this node and the nodes in its neighborhood).[14] The overall clustering coefficient of a network is then the average of all the clustering coefficients of all the nodes in that network:

$$\bar{C} = \frac{1}{n}\sum_{i=1}^{n} C_i$$

5.2.1.6. Average Path Length (L)

The average path length of a network refers to the average distance between one node and every other node in the network (Vitevitch [2008, 4]). If we have $d(v_1, v_2)$, which denotes the shortest distance between the two nodes v_1 and v_2[15], then we

[12] **Note** that in this thesis the term *clustering coefficient* is used to refer to the clustering coefficient of a graph.

[13] Source: `https://en.wikipedia.org/wiki/Clustering_coefficient`

[14] k_i is the number of neighbor nodes of a node v_i.

[15] Assuming that $d(v_1, v_2) = 0$ if node v_2 cannot be reached from node v_i.

calculate the average path length for the whole networks as:[16]

$$L = \frac{1}{n \cdot (n-1)} \cdot \sum_{i \neg j} d(v_i, v_j)$$

A short average path length is the most characteristic feature of so-called small-world networks because it illustrates why we are talking about "small" worlds. The word "small" refers here to the fact that any two nodes in such a network can be reached through a few intermediate nodes (Watts and Strogatz [1998]).

The above presented network properties are very characteristic for networks created from real-world data and can therefore be found in networks created from email communication, gene interaction, neural systems in the brain (which have been applied in computer sciences as artificial neural networks (Ke [2007])) and networks created from linguistic units (Solé et al. [2010], De Deyne and Storms [2008], Ke [2007], Vitevitch [2008]). Following Barabási and Albert [1999], a very interesting point discussed by Beckage et al. [2011], Ke [2007] and Vitevitch [2008] is that these properties may account for the rapidness, accuracy and robustness of linguistic networks, because, in certain graph types (small-world and scale-free), they are said to be responsible for "optimal navigation in the mental lexicon for speech production" (Ke [2007, 7]).

5.2.2. Graph Types

Certain graph types have been of particular interest since the findings of Watts and Strogatz [1998], who showed that graphs with a so-called "small-world" structure have a type of organization which is responsible for keeping the networks between overly structured and totally unstructured organizations (Mehler et al. [2015, 7]). Furthermore, so-called "scale-free" networks have been shown to account for particular types of growth processes, which also have been used to explain language acquisition (Barabási and Albert [1999], Steyvers and Tenenbaum [2005]).

Depending on their structural properties, graphs can be divided into several types (Mihalcea and Radev [2011] and Mehler et al. [2015]):

Bipartite graphs
Are graphs in which their nodes can be divided into two disjoint subsets (Mihalcea

[16] source: `https://en.wikipedia.org/wiki/Average_path_length`

and Radev [2011, 32]).

Random graphs

Are graphs whose edges are created randomly. That is, when starting with a set of N nodes, a random graph is created by randomly placing edges between those nodes. Because random graphs have been studied extensively in mathematics, and are fairly well understood, they are often used as a baseline with regards to the calculation of the above presented graph properties, in order to compare the results with graphs which were not created randomly.[17] One of the mathematically best understood models for random graphs comes from Paul Erdős and Alfred Rényi(Vitevitch [2008, 3]), where random graphs are created with the *G(n,p)* model, where n is the number of nodes in the data graph and p is the probability of edge creation (which is calculated as $2m/(n(n-1))$, where m is the number of edges in the data graph).

Regular graphs

Are graphs in which all the nodes have the same degree. There can be zero-degree regular graphs, with no edges at all, one-degree regular graphs where every node is connected to exactly one other node, etc. (Mihalcea and Radev [2011, 14])

Scale-free graphs

Are graphs where degree distribution of their nodes can be well approximated by a power-law (Mehler et al. [2015, 10]). As already mentioned in section 5.2.1, many graphs created from real-world data are scale-free graphs. The main characteristic of such a structure is defined by the existence of a few central nodes (nodes with many out-degree connections, so-called hubs), and a far greater amount of nodes with only very few connections. According to Barabási and Albert [1999], this structure may also be an evidence for a process called "preferential attachment", where new nodes joining a network typically do so by connecting to a present node in the network which already has many connections. This "rich gets richer" effect (Mehler et al. [2010, 10]) is in turn responsible for the fact that there are only few nodes with a high degree, and very many nodes with a low degree in such networks.

Small-world graphs

Small-world graphs are probably the most widely studied graphs in social sciences (Vitevitch [2008]). The concept of "small-world" originally dates back to the work of Milgram [1967], where he conducted an experiment to measure the "social *distance*" (Vitevitch [2008, 7]) between the people in the United States.[18] Compared to a

[17] In the study where we have examined the properties of lexical co-occurrence graphs from Russian and Chintang children and their caregivers, the differences between random and small-world and scale-free graphs and their implications will be discussed in further detail in section 5.4.

[18] In his study, Milgram randomly chose a person who was to receive the name of a target person, to

random network, small-world networks tend to have a short average path length (i.e. no matter how large a network is, any node can be reached from any other node by traversing only a small number of intermediate nodes (Ke [2007, 2]). Additionally, small-world networks tend to have a high clustering coefficient, because the nodes in such networks tend to share common neighbors, thus forming clusters (ibid.). What is more, findings from analyzing small-world and scale-free networks have lead to important insights which could not have been achieved with traditional approaches. An example from the health sector is mentioned in Barabási and Bonabeau (2003), where it was shown that random immunization in a social network could easily fail because of the scale-free property of social networks. This means that networks with a scale-free property have a small number of very important nodes (hubs) and a large number of nodes that have only a few neighbors. Therefore, in such networks, an effective way to stop epidemics is not by randomly immunizing people, but by firstly identifying the most important nodes (people) of a network by means of applying analysis techniques from network theory. Consequently, by immunizing the most influenciable people in a social network, the spread of an epidemic can be stopped much faster (Hills et al. [2009]). With regards to networks created from linguistic data, and the fact that they very often show this kind of network structure in many different languages, indicates, according to Mehler (2015), that also a "cognitive system may take advantage of these structural properties" (Mehler et al. [2015, 8]).

5.3. Research Findings: Overview

The interesting findings of Watts and Strogatz [1998] and Barabási and Albert [1999] with regards to small-world and scale-free networks have lead researchers to apply approaches from network theory also to linguistic research questions. According to Ke [2007], in an initial phase, there have been mainly two types of networks created from linguistic data: networks which contain semantic relationship or networks which are constructed from grammatical relationships between words. In recent years, network theory has been expanded to other linguistic subfields, such as networks created from phonological and also syntactic relations. In the next sections, I will summarize the main findings of studies which have applied network theory

which a package had to be delivered. The person chosen by Milgram was instructed to forward the package either directly to the target person (if the person was known) or to forward the one person in his or her social circle, who would most likely know the target person. This process had to be repeated, until the target person received the package. After analyzing the results of the study, Milgram found that, on average, six intermediate people were required to deliver the package to the target person. Milgram's work contributed to the idea that we all live in a "small world" and that there are only "six degrees of freedom between any two people on the planet." (Vitevitch [2008, 7]).

to examine semantic, phonological as well as lexical relationships between linguistic items. Interestingly, despite the different nature of these networks, they all exhibit small-world properties, which again favors the view that there exist some universal mechanisms, which also account for linguistic networks. Considering the degree distribution of these networks, a power-law distribution has not been found in all of them.

Network theory has, according to Gegov et al. [2011]) gained wide popularity as an analysis tool because it is extremely powerful for modeling and analyzing various relationships on the micro-level (i.e. the relationship between individual words in the lexicon) as well as on the macro-level (which would be the overall structure of the lexicon) of a network (Goldstein and Vitevitch [2014]). Furthermore, network theory has successfully been applied to various field of child language acquisition, such as in semantic networks (Steyvers & Tenenbaum 2005), phonological networks (e.g. Vitevitch [2008]), collocation networks (Ke and Yao [2008], Adamo and Boylan [2008]) and syntactic networks (Corominas-Murtra et al. i Cancho and Solé [2001], Adamo and Boylan [2008]). Modeling language acquisition usually requires a lot of data (to e.g. train machine learning models on maternal utterances and then compare the resulting utterances generated by the machine to those uttered by the children (cf. the MOSAIC model of Freudenthal et al. [2007])). However, linguistic data sets are typically very noisy and very difficult to compare directly (Gegov et al. [2011, 1]). By modeling language data sets using a network representation (and also using analysis techniques from network theory) one can embed all the information from the data set into a single network, which can be analyzed and compared with other networks created from other data sets.

5.3.1. Semantic Networks

As we have already seen in the example figures (Figure 5.2 and Figure 5.3) of semantic networks, they are created by placing words (or concepts) as nodes and connecting them with edges if they are semantically related to each other. The weight of the edges in such networks can be defined by assigning them values of certain association or similarity measures (such as point-wise mutual information or the cosine similarity, respectively). Semantic networks have usually been build from either thesauri (Motter et al. [2002]) or databases (Steyvers and Tenenbaum [2005]) or from human similarity tasks (Steyvers and Tenenbaum [2005]). As to the development in applying network theory to semantic networks, Ke [2007] notes that early studies mostly focused on refining networks of specific semantic domains, thereby often neglecting global properties which can be observed by applying analysis tech-

niques from network theory (Ke [2007, 7]). Later, when these analysis techniques became more widely applied, researchers started to also examine global properties of semantic networks. For example Motter et al. [2002] have constructed a network from synonyms from the Moby thesaurus dictionary. In their study, they found that also this semantic network showed small-world properties. The following parameters are indications for this: The studied network was highly clustered (CC = 0.52) in comparison to a random network, it had a small average path length (L=3.16) considering the huge size of the network (over 30'000 nodes). Another interesting point they were able to show in their study is that the degree distribution of their semantic network showed to regimes: an exponential distribution for the nodes with a low degree, and a power-law distribution for the nodes with higher degrees. Another study was conducted by Steyvers and Tenenbaum [2005], where one of they build three synonym networks using a free-association database, WordNet and the Roget's Thesaurus. In this study, similar findings were found, as the networks also had short average path lengths, a high clustering coefficient and a degree distribution following a power-law.

With respect to semantic networks created from child language data, work from De Deyne and Storms [2008], Hills et al. [2009] and Steyvers and Tenenbaum [2005] has shown that networks created from word association tasks also share small-world and scale-free characteristics. Zortea et al. [2014] investigated in their study the network properties of semantic networks from children, adults and elder people. The results of their study showed that all the three networks had similar measures with respect to the number of nodes, edges, the clustering coefficient, as well as the average path length (of course, the children's networks had a lower number of nodes with lower links between them, which is also a reason why the networks of the adults and the elderly are more connected, denser and have fewer isolated nodes (Zortea et al. [2014, 95]).

5.3.2. Syntactic Networks

Even though lexical networks have shown to also account for syntactic characteristics within a language (Ke [2007], Ke and Yao [2008], Vitevitch [2008]), work on syntactically annotated corpora, where also long-distance syntactic relations can be captured between words, has only recently been conducted. For example, i Cancho et al. [2004] created networks from syntactically annotated corpora and analyzed them for three different languages German, Romanian and Czech. In their study, i Cancho et al. [2004] showed that there are similar characteristics detectable also in syntactic networks, even though syntactic networks would be expected to differ more

cross-linguistically than e.g. semantic networks. In the work of i Cancho et al. [2004], all the three networks showed small average path lengths and a high clustering coefficient (indicators for a small-world structure). Furthermore, the three networks uniformly showed a power-law distribution and dissortative mixing by degree (i.e. where highly connected nodes tend to be connected with other highly connected nodes). Because syntactic networks have been the least studied linguistic networks by applying graph theory, syntax is a very interesting area for further research, especially when comparing syntactic networks cross-linguistically on maximally diverse languages.

5.3.3. Phonological Networks

The most exhaustive study with regards to phonological networks has been conducted by Vitevitch [2008]. In his study, Vitevitch examined the structure of phonological word-forms in the lexicon of adult speakers in order to gain possible insights on the constraints that may influence lexical acquisition and word learning. His main motivation was to find out if the same network structure found in a network of semantic representations will also be found in a network of phonological word-forms. In his study, Vitevitch constructed graphs by defining the phonological representation of entries in the *Merriam-Webster Pocket Dictionary* (Woolf [1974]) as nodes and placing edges between those nodes, if their phonological form differed by adding, deleting or substituting only one single phoneme. For example the word *cat* would have direct connections to the following words: *hat, cut, cap, at* (Vitevitch [2008, 2]). According to Vitevitch, phonological similarity has been used in various psycholinguistic studies and has shown to influence language processes in children such as "the acquisition of sounds, the acquisition of words, spoken word recognition (also with young adults with hearing impairment) and spoken word production" (Vitevitch [2008, 3]).
In order to have a comparison for the structure of the obtained graphs, Vitevitch also constructed random graphs with the same number of nodes and the same number of average connections per node as the graphs created from the *Merriam-Webster Pocket Dictionary*. The parameters used in Vitevitch's study were ***average path length, clustering coefficient, degree distribution*** and **"the extent of assortative mixing by degree** in the network" (ibid).[19] The comparison of the results for the average path length as well as the clustering coefficient in Vitevitch's study suggest that also phonological networks show small-world characteristics be-

[19] *Assortative mixing by degree* refers to the probability of a highly connected node being connected to other nodes that are also highly connected.

cause they have an average path length that is comparable to the average path length of a random network, and they also have a clustering coefficient that is much greater than the clustering coefficient of a random network having the same number of nodes and the same average degree (Vitevitch [2008, 7]).
Concerning the ***degree distribution*** of the network, it seemed to better fit an exponential than a power-law distribution in Vitevitch's study, a feature that distinguishes phonological networks from semantic and lexical networks, where the degree distribution has been shown to follow a power-law distribution (Steyvers and Tenenbaum [2005])). This is a surprising finding, particularly with regards to other studies which have shown that new words are acquired faster if they resemble phonologically already known words. Such a characteristic, which is also referred to as ***preferential attachment***, is actually one of the main features of scale-free networks (Barabási and Albert [1999]). As possible explanation for the lack of a power-law distribution within the phonological network, Vitevitch mentions the work of Amaral et al. [2000], who have shown that, if there is a cost associated with the attachment of a new node, it might prevent the degree distribution to follow a power-law distribution (Vitevitch [2008, 12]). Factors which might be responsible for limiting the number of nodes that can be attached to already existing nodes within a phonological network are, according to Vitevitch, word length, the phonemic inventory of a language, as well as language specific phonotactic constraints (i.e. sound sequences in words). However, the exact reason for the lack of a power-law in the degree distribution is a topic that needs further investigation (ibid). By taking a network theoretical approach to analyzing the characteristics of an (English) phonological network, Vitevitch has revealed very interesting insights with regards to the structure of the phonological network in his study. Despite the fact that the phonological network showed small-world characteristics by showing an average path length that is similar to the path length of a random network, as well as by having a clustering coefficient that is much higher than the clustering coefficient of the random network, the striking finding of his study was that the degree distribution of the phonological network did not follow a power-law. Instead, the degree distribution seemed to better fit an exponential function. A possible explanation for this might lie in language-specific phonological features and rules. Another feature Vitevich observed in his study was that the phonological network exhibited assortative mixing by degree, which might be an important factor for assuring rapidness, accuracy and robustness of linguistic networks.

5.3.4. Lexical Networks

Apart from semantic and phonological networks, there has been increasing interest in the past years in the study of networks created from lexical co-occurrences. With respect to the construction of lexical co-occurrence networks, there have been only nodes directly adjacent to each other taken as nodes, or in other studies only words with a certain POS-tag were added as nodes to the network. According to i Cancho and Solé [2001], networks created from word co-occurrences also show small-world characteristics by also having a small average path length, and a much higher clustering coefficient than in a comparable random graph. As to the degree distribution, word co-occurrence networks also tend to show a scale-free distribution of the node degrees. Ferrer-i-Cancho and Solé created lexical co-occurrence networks from running text extracted from the British National Corpus.

With regards to lexical co-occurrence networks constructed from child language acquisition data, researchers were able to gain new insights on the development in word learning, which "traditional" measures such as MLU (mean length of utterance) and frequency measures were not able to detect. For example, by comparing the size of networks created from word co-occurrences of child language acquisition data, as well as their connectivity, Ke and Yao (2008) were able to show that children with a small vocabulary size, but a comparably high average degree exploit the words they know with more flexibility.[20] Furthermore, an increasing network size as well as an increasing average degree account for a growing complexity of the produced utterances Ke and Yao [2008, 76]. Apart from these global structures which have nowadays been applied to many other linguistic networks as well, Ke and Yao were furthermore interested in studying local properties of lexical co-occurrence networks in more detail, which, also according to Ke [2007]) may not only account for lexical growth in the language network of a child, but also for the emergence of syntactically more complex utterances that start to include more function words.[21] In their study, Ke and Yao have extracted so-called "hub" and "authority" nodes from staged networks in order to account for a more detailed network structure over time by identifying the most influential nodes in networks created from grouped data

[20] Because a higher degree of a node means that it is combined with many other nodes.

[21] Even though the analysis of lexical networks can account for syntactic development to some degree, cf. Ninio [2014, 632], who criticizes this approach by noting that real syntactic complexity also depends on the length of the utterances, and not only on the combinatorical characteristics of the words used in the utterances. For example, the structure of a lexical network might look the same for a child that produces many two-word utterances as for a child that produces fewer utterances, but longer ones, even though the syntactic structure of longer utterances (e.g. sentences consisting of ten words) are clearly more complex than shorter utterances (e.g. sentences consisting of only two words).

sets based on different MLU values of the children's data.[22] ***Hubs*** are those nodes in a network that have high out-degree values and point to many nodes that are important ***authorities*** in a network (i.e. nodes which are pointed to from many hubs). As opposed to other nodes with also high in- or out-degree values, hubs and authorities are those nodes which tend to connect with other *important* nodes (i.e. with other nodes that are either important hubs or authorities themselves). The detection of hubs and authorities is an important task in network theory because a change in nodes denoting hubs and authorities accounts for a change in the entire structure of a network. With respect to linguistic networks created from lexical co-occurrences, this means that hubs and authorities can account for the transition of the linguistic system of a speaker [Ke, 2007, 85], in particular for the emergence of syntactically more complex utterances. In their study, Ke and Yao have analyzed the development of hubs and authorities in lexical co-occurrence networks of 12 English-speaking children (ranging in age from 1;8.22[23] to 2;0.25) and their mothers from the Manchester corpus Theakston et al. [2001] of the CHILDES MacWhinney [2000b] database. In their data set, the language of the mothers remained relatively stable across the time span of the recorded data. This is also mirrored by the fact that the five most important hubs and authorities rarely changed. This stays in contrast to the children's networks, which show a much greater variety in their development of hubs and authorities [Ke and Yao, 2008, 86]. Whereas words denoting hubs and authorities in the early age of the children consist of content words or names such as *mama, juice* and *whale*, they later seem to assimilate to the mothers' hubs and authorities which usually denote function words such as conjunctions, prepositions, determiners and pronouns (ibid). Ke and Yao have also shown that there are individual differences among children with regards to the development of hub and authority nodes.

While Ke and Yao analyzed the development of hubs and nodes by looking at the ten most highly rated hubs and authorities in each network, Adamo and Boylan 2008 measured in their study also the ***in-*** and ***out-degree centralization*** of various staged networks created from the data of one English-speaking child (raging in age from 1;01.02 to 4;00.02) from the Providence corpus of the CHILDES database. By including in-degree and out-degree centralization measures we can investigate how these degrees are distributed within a network. Higher values for centralization mean that in-degree/out-degree centralization is focused on a few nodes, and lower centralization values indicate that in- and out-degree centralization is spread more evenly among multiple nodes in a network Adamo and Boylan [2008]. As Adamo

[22] Note that with "staged" networks they mean networks generated from data that was grouped based on different MLU values of the underlying data.

[23] 1 year, 8 months, 22 days.

and Boylan have shown, out-degree centralization can account for interesting developmental "milestones" in child language development by pointing to stages where firstly, the grammatical nature of the hubs changes (from more content words to more function words) and secondly, the development of hubs and authorities levels out and approximates the measures found in the data of the mothers. By examining the networks of these "changing" stages more closely, Adamo and Boylan showed that a first increase in out-degree centralization appears when the child starts to use the first function words, a following decrease in out-degree centralization then happens when the number of function words in the child's vocabulary increases, thereby mirroring a more evenly distributed out-degree centralization across the network.
Another study with regards to language acquisition and lexical co-occurrence networks was conducted by Gegov et al. [2011], where lexical networks were generated from the data of the Manchester corpus of the CHILDES database (MacWhinney [2000b], Theakston et al. [2001]). The characteristics (as well as the correlations between them) of the following eight statistical parameters were investigated: *mean length of utterance, number of nodes, number of edges, giant connected component, average degree, average geodesic length, average clustering coefficient, node degree distribution*, as well as the *ranked link frequency distribution.*[24]. In a network constructed of word co-occurrences, a power-law distribution means that language productivity is very biased towards some word co-occurrences, which are produced much more often than most other co-occurrences (Gegov et al. [2011]). In their study, Gegov et al. constructed the networks of lexical co-occurrences of all the utterances uttered by the children of the Manchester corpus, as well as of the utterances uttered by the mothers. The data set used in their study was divided into three non-overlapping developmental stages for the children and into one stage for the mothers.[25] The networks where then constructed by defining the various words as nodes (thereby also deleting duplicates because every node only occurs once in the network) and placing a link between two nodes if they co-occur within an utterance.

[24]The ranked frequency distribution, shows how the degree of link frequencies decreases when the frequencies are sorted in descending order, thus also following a power-law distribution

[25]The argumentation for treating the mothers' data as only one stage is that Gegov et al. expected the language of the mothers to "remain fairly stable". Making such generalizing assumptions seemed quite daring to us. This is the reason why, in the study conducted for this thesis, we decided to partition the data of the mothers (adults) and the children into sets of temporally more or less stable ranges by using the same partitioning for the adults' data as in the children's data.

5.4. ACQDIV & Lexical Co-Occurrence Networks

As part of this thesis, we conducted two studies where we applied analysis techniques from network theory to lexical co-occurrence networks generated from data of the ACQDIV database. In a first study, we were interested in seeing if small-world and scale-free properties can be detected in all languages of the ACQDIV database, despite their typological differences and, consequently, their diverging notions of what denotes a "word" in the linguistic sense. Therefore, for this first study, we created lexical co-occurrence networks of all the child directed speech data from every corpus in the ACQDIV database. We compared the resulting networks to random graphs with regards to the following graph properties: **Network Size (N)**, **Connected Components (CC)**, **Average Path Length (L)**, **Average Clustering Coefficient (C)** and **Degree Distribution (*P(k)*)**. Our hypothesis was that all of the networks would exhibit small-world and scale-free characteristics, despite their typological differences. In a second study, we used the same parameters to compare if lexical co-occurrence networks created from the data of the children in our database will also exhibit small-world and scale-free properties. In order to account for the developmental change in the networks of the children's data, we followed the approach taken by Ke and Yao [2008] and Adamo and Boylan [2008] and also constructed staged networks (by grouping all recording sessions into groups consisting data of a time-span of approximately one month) which were then further examined on a micro-level by calculating the ten most important hubs of each staged network and by using the graph analysis software *Gephi* (Bastian et al. [2009]) in order to visualize the temporal development of these hubs. Based on former research summarized in 5.3.4, our hypothesis was that words denoting hubs in early child language acquisition are words with low semantic content. This expectation is in line with previous findings that show in English that hubs in early child language are mostly semantically vacuous lexical categories, i.e. indefinite and definite articles. However, these lexical categories do not appear across languages, so we test our hypothesis on Russian and Chintang, two languages that lack a clear-cut article distinction and for which there are richly annotated large longitudinal child language acquisition corpora. Additionally, we are interested in seeing if the caregiver's hubs change with time as well, reflecting a language adaptation to meet the children's linguistic abilities.

5.4.1. Data

For the two above mentioned studies, we used two different data sets because in the global study, we were interested in seeing if networks created from lexical co-occurrences exhibit small-world and scale-free properties in general, we only took the data of the adult speakers into consideration for this first study. On the other hand, we used only the data of two corpora of the ACQDIV database for our second study, where we were interested in comparing the generated networks language wise on a child-by-child and a child vs. caregiver(s) basis, as well as cross-linguistically, by comparing the results of our analysis across the two languages.

5.4.1.1. Global Networks

In our "global" study, we expected to find the same small-world properties that have been found for networks of other languages like English (Motter et al., 2002, Steyvers and Tenenbaum 2005, Sigman and Cecchi **?**, Cancho and Solé 2001), Czech, German and Romanian (Cancho et al., 2004), despite the typological diversity of the languages in our study. To illustrate the typological diversity, consider Russian, which is one of the Indo-European languages with a fairly complex morphological system, for example it has inflectional classes both in the nominal and verbal domains and often expresses a large number of categories by a single morpheme. The examples 5.2 and 5.3 show the same bundle of grammatical functions (PL.GEN) expressed by very different morphs due to nominal inflection classes. By contrast, Chintang does not feature any inflectional classes, has less compact grammatical morphemes, and may even express a single function several times within a single word, as shown by the complex verb form in 5.4:

(5.2) *Skol'ko* *produkt-ov* *papa* *nam* *privez?*
How.many product-PL.GEN dad.NOM 1PL.DAT bring.PFV.PST.M.SG.S/A
'How many products has dad brought us?'

(5.3) *Im* *mnogo* *konfet-Ø* *togda* *ne* *da-esh'.*
3SG.DAT much sweet-PL.GEN then NEG give.IPFV-NPST.2SG.S/A
'Don't give him too many sweets then.'

(5.4) *Athom* *u-patt-a-ng-s-a-ng-ni-ng=kha.*
before 3A-call-PST-1sP-PRF-PST-1sP-3p=NMLZ
'They had called me before'

The data from the ACQDIV database with regards to number of sessions and utterances of each corpus is summarized in 5.1. Because of the very heterogenous number of utterances and sessions in our database, we initially hesitated in using all of our

corpora for the global network study, because we feared if there would be enough data for reliable results. Luckily (as the results from our second study confirm), there was no problem with regards to the sample size.

Corpus	**Utterances**	**Sessions**
Chintang	393030	477
Cree	20648	25
Indonesian	915759	997
Inuktitut	46683	77
Japanese	437348	362
Russian	827589	450
Sesotho	69575	115
Turkish	401262	373
Yucatec	93185	125

Table 5.1.: Corpus Size

5.4.1.2. Local Networks

The data used for the second part of our study comes form the Russian Stoll and Roland [2008] and the Chintang Bickel et al. [2011] corpus of our database. The Russian corpus is the same as used in chapter 4.

The data of the **Chintang subcorpus** was compiled between 2004 and 2015 in the course of several research projects now summarized as the *Chintang Language Research Program* (CLRP). This corpus contains recordings of seven children ranging in age from 0;7.23 to 4;4.14 years.[25] The sessions for this corpus were recorded in monthly intervals where the children were recorded for approximately 4h (taken during several sessions within a single week) while mostly playing outside close to their home with their mothers and/or other relatives. In order to keep the age range for the two languages as comparable as possible, we will only use the four older children from the Chintang corpus.

The basic statistics of the number of sessions, utterances and unique words for every speaker used in our study are shown in Table 5.2 and 5.3.[26]

[25]The recordings for one child were cancelled early.

[26]Note that in the second study, we used only the mothers data for Russian in order to construct reference networks, but for Chintang we took the data of all the adult speakers that were present in recording sessions for every target child. This methodological decision was made due to cultural factors (because in Chintang there is no "main" caregiver role for a child as this is for example (mostly) the mother in Western-European cultures).

Russian				
Speaker	**Age Range**	**Sessions**	**Utterances**	**Tokens**
RUChild1	1;6.10 - 5;4.18	129	124269	543244
RUChild2	1;3.26 - 4;11.0	124	45420	662743
RUChild3	3;1.8 - 6;8.12	119	64711	642479
RUChild4	1;11.28 - 4;3.14	67	43973	329522
RUChild5	1;4.22 - 5;6.26	130	53508	506678
RUChild1 aunt	x	91	47555	382037
RUChild2 mother	x	88	95608	282483
RUChild3 mother	x	85	92504	496120
RUChild4 mother	x	54	43586	264982
RUChild5 mother	x	123	141313	477916

Table 5.2.: Russian Data

Chintang				
Speaker	**Age Range**	**Sessions**	**Utterances**	**Tokens**
LDCh1	2;1.9 - 3;5.25	90	17677	182534
LDCh2	2;0.29 - 3;5.13	115	21046	228244
LDCh3	3;0.14 - 4;4.25	147	28459	267833
LDCh4	2;11.2 - 4;3.14	115	20168	272344
LDCh1 adults	x	89	30868	87190
LDCh2 adults	x	114	39593	111220
LDCh3 adults	x	147	37774	105964
LDCh4 adults	x	113	37937	111071

Table 5.3.: Chintang Data

5.4.2. Methodology

5.4.2.1. Global Networks

The relational SQLite database we have build (cf. chapter 2.4) allows us to make language specific queries on various linguistic levels (phonological, morphological, lexical). For both studies, we queried the database for language- and speaker-specific utterances and further created lexical co-occurrence graphs by splitting the utterances on white space characters delimiting unique word forms.[27] The types of unique word forms are then used to represent the nodes in our graphs. A link is placed between two nodes if they directly co-occur (i.e. if they are directly adjacent) within an utterance. A mini graph in Figure 5.8 created from two example utterances 5.5 and 5.6 shall illustrate our methodology.[28]

(5.5) *ALJ* *idi* *mjachik* *narisuju.*
M.SG.NOM.AN IPFV.IMP.2SG M.SG.ACC.INAN PFV.NPST.1SG
'ALJ come I'll draw you a ball.'

(5.6) *Davaj* *mjachik* *narisuju.*
IPFV.IMP.2SG M.SG.ACC.INAN PFV.NPST.1SG
'Give I'll draw you a ball.'

The mini example also illustrates that nodes appearing multiple times in the data

[27] Here we followed the various approaches that have already been taken in creating graphs from lexical co-occurrences by using the unique word form instead of the lemma of a word. Ke and Yao justify this decision with the fact that by taking the actual word form, we avoid the problem in determining weather a word is learned as individual lexical item or derived from morphological rules [Ke and Yao, 2008, 75]. However, given the typological variety in our database, treating characters separated by whitespace as single words is not always trivial because, when analyzed in detail, it is rather difficult to define what a word is cross-linguistically (including whether it is phonological, morphological or orthographic). In some languages words represent full phrases, in others the word and morpheme are nearly synonymous.

[28] The size of the nodes in this example is determined by the degree of the node, i.e. the more links a node has, the bigger it is drawn.

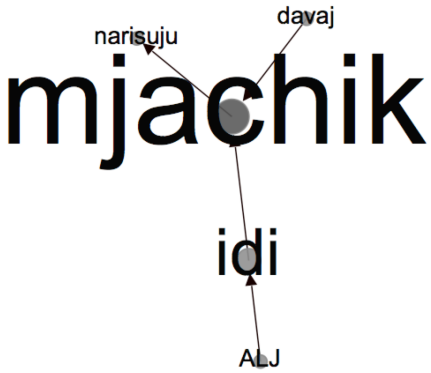

Figure 5.8.: Example graph for two sentences

set will only appear as a single node in the graph. Furthermore, if an utterance only consists of a single word (and if this word never appears in any multi-word utterance), this word will be placed as a single node in the graph (not shown in Figure 5.8). Note also that this method actually creates a *multidigraph* (because it allows multiple links from a source to a target node). As already mentioned by Ke 2007, this is to be avoided because such graphs have very complex structures and are hence computationally very expensive to process. Therefore, after querying the database with a Python script and extracting all the necessary information with the `SQLAlchemy` (Bayer [2016]) and `Pandas` (McKinney [2012]) library, we use the graph library `Networkx` (Hagberg et al. [2005]) to convert this multidigraph into a *weighted digraph* (i.e. a directed graph), where the weight of the edges corresponds to the frequency with which the edge from target to source node occurs in the data set. This digraph is then exported to a special Graph Exchange XML format (`.gexf`). This graph files are then loaded into R (Ihaka and Gentleman [1996]), where the following statistical parameters are computed using the `igraph` (Csardi and Nepusz [2006]) library: *number of nodes (N), number of edges (E), average degree (<k>), average path length (L)* as well as the *average clustering coefficient (C)*. We also construct random graphs for comparison with the *igraph* library. Here we use the Erdős-Rényi *G(n,p)* model[28], where n is the number of nodes in our data graph we want to compare, and p is the probability of edge creation (also calculated on the data graph we want to compare).[29].
For our first analysis concerning the global, language-specific comparison, we created graphs for all the utterances spoken by adult Russian and Chintang speakers, and

[28]Note that our random graphs are also directed graphs.

[29]We calculate p as $2m/(n(n-1))$ where m is the number of edges in our data graph.

a graph for all the utterances produced by each child.[30] The results of this analysis are shown in Table 5.5 and will be discussed in more detail in chapter 5.4.3.1.

5.4.2.2. Local Networks

For our second study, we also used the Python libraries `SQLAlchemy` and `Pandas` to query the database to extract the utterances from the data. But this time we followed the approach applied by Ke and Yao [2008] and built graphs for different age stages of the target children, where we also used a time-stamp for every new node and edge entering the graph in order to be able to replicate the growth of the lexical network for multiple age stages. We used the same age stages for partitioning the data of the mothers because we are interested in seeing if the language input for a child changes in a similar way in which the produced output changes. As the Russian data consists of sessions that were recorded on a weekly basis, we partitioned this data into bigger groups by always combining sessions that were recorded within a time span of approximately one month to one big session (which allows us to create the graphs from bigger data groups).[31] For Chintang we also grouped all the sessions that have been recorded within the time span of a month, thereby keeping the stages across corpora as consistent as possible. An example structure of a *.gexf* file is shown in Figure 5.9, where we can see that each node and each edge between two nodes is assigned a time-stamp (the date is extracted from the session date), depending on when it appeared for the first and last time during our grouped sessions:

[30] We are not grouping the data of the children because children show such variability thats pooling their utterances may obscure detectable patterns.

[31] Note that this approach turned out to be problematic, as it does not account for "real" developmental patterns of the child within the data, it actually only accounts for a temporal development. As the goal for this thesis was to test if former research questions would yield similar results, we decided to follow this "temporal" path of partitioning the data. An approach suggested by Gries and Stoll [2009], where the underlying data is first partitioned according to developmental patterns, would be desirable for further research.

gexf_example.gexf

```
<?xml version="1.0" encoding="UTF-8"?>
<gexf version="1.1" xmlns="http://www.gexf.net/1.1draft" xmlns:viz="http://www.gexf.net/1.1draft/viz" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://www.w3.org/2001/XMLSchema-instance">
  <graph defaultedgetype="directed" mode="dynamic" timeformat="datetime">
    <attributes class="node" mode="dynamic">
      <attribute id="0" title="duration" type="string"/>
      <attribute id="1" title="end" type="string"/>
    </attributes>
    <nodes>
      <node id="ama" label="ama" start="2004-04-18 00:00:15" end="2004-04-18 00:13:14">
        <attvalues>
          <attvalue for="0" value="2004-04-18 00:00:15"/>
          <attvalue for="1" value="2004-04-18 00:13:14"/>
        </attvalues>
      </node>
      <node id="khoi" label="khoi" start="2004-04-18 00:00:16" end="2004-04-18 00:12:41">
        <attvalues>
          <attvalue for="0" value="2004-04-18 00:00:16"/>
          <attvalue for="1" value="2004-04-18 00:12:41"/>
        </attvalues>
      </node>
      <node id="Debuci" label="Debuci" start="2004-04-18 00:00:35" end="2004-04-18 00:00:35">
        <attvalues>
          <attvalue for="0" value="2004-04-18 00:00:35"/>
          <attvalue for="1" value="2004-04-18 00:00:35"/>
        </attvalues>
      </node>
      <node id="Debi" label="Debi" start="2004-04-18 00:00:36" end="2004-04-18 00:00:37">
        <attvalues>
          <attvalue for="0" value="2004-04-18 00:00:36"/>
          <attvalue for="1" value="2004-04-18 00:00:37"/>
        </attvalues>
      </node>
    </nodes>
    <edges>
      <edge end="2004-04-18 00:00:15" id="2" source="ama" start="2004-04-18 00:00:13" target="khoi" weight="0.2"/>
      <edge end="2004-04-18 00:00:14" id="3" source="khoi" start="2004-04-18 00:00:14" target="ama" weight="0.1"/>
      <edge end="2004-04-18 00:00:29" id="11" source="Debuci" start="2004-04-18 00:00:29" target="Debi" weight="0.3"/>
      <edge end="2004-04-18 00:00:30" id="12" source="Debi" start="2004-04-18 00:00:30" target="Debi" weight="0.3"/>
    </edges>
  </graph>
</gexf>
```

Figure 5.9.: Example for dynamic .gexf file

Assigning time-stamps to the nodes and edges allowed us then to load this .gexf file into Gephi and use the timeline option in order to simulate the growth (and the changes over time) in the network. We created a graph for each age stage for each child and compared these graphs over time with regards to the following global network parameters: **N, E, <k>**. We also compared these parameters to the MLU of our target children and their mothers. Further building on the work of Adamo and Boylan 2008, we calculated the in- and out-degree centralization for the different age stages. Abrupt changes in the out-degree centralization values were then be used to examine the age ranges before and after these changes more closely by simulating the growth of the graph with the timeline option as well as by dynamically applying the HITS Kleinberg [1999] algorithm (used to detect hubs and authorities within a network) in Gephi. In doing so, we were interested in seeing how nodes denoting typical hubs in a network firstly change over time, and secondly, connect to other nodes in the network. The results of this study will be discussed in further detail in section 5.4.3.2.

5.4.3. Analysis & Results

In the following sections, we will present the analysis and results of our global and our local study.

5.4.3.1. Global Networks

CDS	N	E	E_{random}	<k>	$<k>_{random}$	L	L_{random}	C	C_{random}	γ	R^2
Chintang	60535	173350	347894	5.727265	11.49398	3.861909	6.497079	0.01949274	0.0001666573	1.44	0.93
Cree	4577	9945	19943	4.345641	8.714442	3.925917	5.89046	0.0354484	0.002096497	1.523	0.947
Indonesian	26264	257285	516430	19.59222	39.32607	0.87	3.174789	3.73446	0.07556786	1.72	0.854
Inuktitut	11705	7851	15703	1.341478	2.683127	6.714106	23.04485	0.008594746	0.000283802	1.658	0.974
Japanese	24051	151235	303140	12.57619	25.2081	3.156601	4.261239	0.03455292	0.001044477	1.429	0.893
Russian	47895	290832	581179	12.14457	24.26888	3.382483	4.599158	0.03299296	0.0004945935	1.577	0.93
Sesotho	5519	22409	44735	8.120674	16.21127	3.265634	4.363033	0.05016898	0.002802469	1.286	0.932
Turkish	61537	324164	648276	10.53558	21.06947	3.621385	4.933678	0.03675761	0.0003510219	1.525	0.934
Yucatec	15967	38638	77134	4.839732	9.661677	4.020815	6.317499	0.02289779	0.0006281727	1.466	0.945

Table 5.4.: Network parameters comparison

Considering our global study, where we wanted to see if small-world and scale-free characteristics can be detected in all of our languages, despite their typological difference, the results shown in Table 5.4 were obtained: Inuktitut has the smallest number of edges and also the smallest node degree, but the highest average path length. This is in line with our expectations given Inuktitut's regular agglutinative morphology; there are few combinations of bigrams delimited by white space. On the other hand compare Indonesian, which has a higher key parameter. This finding is also in line with our expectations because Indonesian's morphology is isolating and words are combined more much frequently than in the morphologically more complex languages in our sample.

The characteristics of small-world graphs hold for each language-specific network in our sample: the number of edges is a lot greater in the randomly-generated graphs; degree is higher in the random graphs; the principally short average path lengths are similar as in the random graphs; and the clustering coefficient is much higher in the child-directed speech networks than in the random graphs. In order to see if the networks from our global study also exhibit scale-free properties, we plotted their degree distribution on a log-log scale. Figure 5.10 shows the degree distribution from our Russian data, it clearly looks similar to the "theoretical" power law distribution by Mihalcea and Radev [2011] (shown in Figure 5.7). Furthermore, the degree distribution of the corresponding random graph also shows as expected a (more or less) normal distribution.

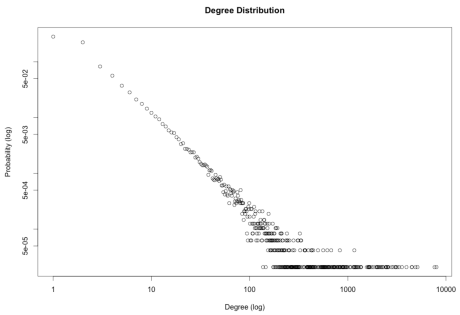

Figure 5.10.: RU degree distribution

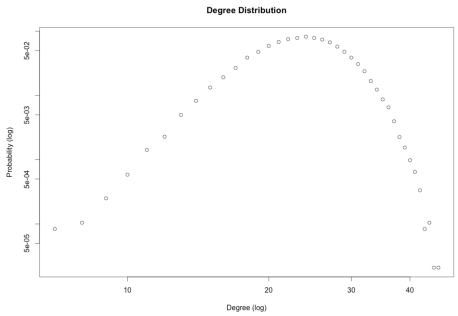

Figure 5.11.: RU random degr. distrib.

In order to further test how well the normal distribution models our actual degree distribution, we followed an approach mentioned in Vitevitch [2008], Zortea et al. [2014] and Liu and Cong [2013] and fitted an expected power-law distribution by plotting a cumulative degree-distribution on a log-log scale (the red line in Figure 5.12) and also measures the value of the constant γ as well as R^2, which tells us how well a power-law distribution models the distribution of our data .

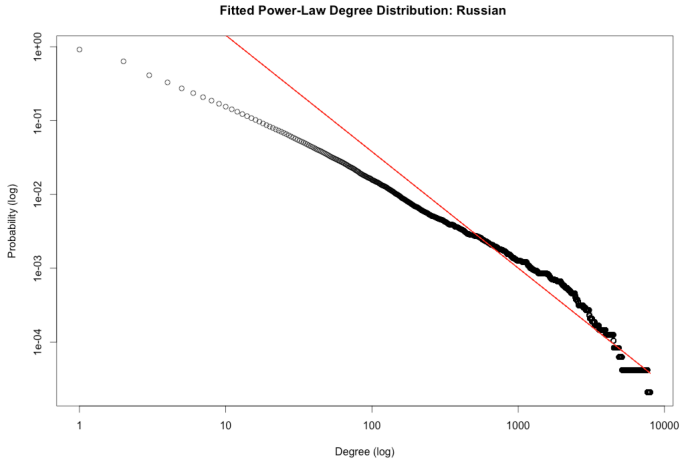

Figure 5.12.: RU fitted power-law distribution

The value for γ and R^2 for our Russian data were 1.577[32] and 0.93, respectively, which indicates that a power-law distribution fits our distribution to 93%. As shown in 5.4, a power-law distribution fits all of the languages of our study >90%, except Japanese and Indonesian, which is most probably due to their typological difference, but exact reasons for this need further investigation. The degree-distribution plots for all the other languages can be found in the appendix of this thesis.

Consequently, we were able to show that networks created from child-directed speech

[32]Which is around 1.5 in a cummulative degree-distribution (Vitevitch [2008, 7], Masucci and Rodgers [2006] in their lexical co-occurrence network from George Orwell's novel 1984, the value for the exponent γ was 1.1. (Mihalcea and Radev [2011, 79]).

from a cross-linguistic sample of nine longitudinal corpora all exhibit small-world and most of them exhibit scale-free structural properties even though the networks from these morphologically very different languages reflect the frequency effects of what linguists consider a word. Our finding is in line with network construction in child language acquisition models that have defined links in terms of semantic or grammatical relationships, both of which exhibit convergent features in their global structures (Ke [2007]). As we have noted, it has been suggested that small-world and free-scale structural characteristics reflect self-organization in the lexicon – a feature that may account for universal properties like fast retrieval from the mental lexicon, but which may also help to account for the fact that children can learn any language's patterns, despite their remarkable diversity in morphological structures.

The resutls of our global study also applied to networks generated from the linguistic data of four Chintang and five Russian children are summarized in Table 5.5. Interestingly, small-world and scale-free properties can also already be detected in the graphs created from all the utterances per child. Of course, when compared to the adults' graphs, the children's graphs show much lower average degrees, but this is to be expected as the vocabulary size of the children is much smaller. When comparing the graphs language wise, we can see that the Chintang graphs have less edges (and consequently also a smaller average <k>) in their network. This is due to the typological characteristics of this language, where information is often coded on a morphological level and not necessarily on a lexical level as this is more the case with Russian.

Speaker(s)	**N**	**E**	$\mathbf{E_{random}}$	**<k>**	$\mathbf{<k>_{radnom}}$	**L**	$\mathbf{L_{random}}$	**C**	$\mathbf{C_{random}}$	γ	$\mathbf{R^2}$
RU adults	47895	290832	581179	12.14457	24.26888	3.382483	4.599158	0.03299296	0.0004945935	1.577	0.94
CTN adults	60535	173350	347894	5.727265	11.49398	3.861909	6.497079	0.01949274	0.0001666573	1.44	0.93
RUChild1	14705	50948	101509	6.929344	13.80605	3.482718	5.187775	0.04734716	0.0009721003	1.474	0.93
RUChild2	6898	14796	29466	4.289939	8.543346	3.826677	6.243457	0.0476125	0.001251505	1.402	0.924
RUChild3	12952	45767	92054	7.067171	14.21464	3.605029	5.070269	0.04565106	0.001017294	1.629	0.912
RUChild4	9976	26696	53330	5.352045	10.69166	3.740627	5.680254	0.03934938	0.001157816	1.416	0.927
RUChild5	8061	15763	31617	3.910929	7.844436	3.83316	6.698409	0.05082797	0.0007765891	1.331	0.968
LDCh1	5246	7962	15754	3.035456	6.0061	4.159497	7.687513	0.05263286	0.001389796	1.43	0.935
LDCh2	5287	8146	16285	3.081521	6.160393	4.004297	7.607707	0.0494303	0.001411327	1.462	0.955
LDCh3	10514	20282	40433	3.858094	7.691269	4.124715	6.990922	0.03149554	0.0007716719	1.447	0.946
LDCh4	8095	14474	28707	3.576035	7.092526	4.115316	7.170245	0.03776784	0.000794363	1.395	0.91

Table 5.5.: Networks Parameters Comparison (incl. Children)

5.4.3.2. Local Networks

For our second "local" network analysis, we first compared the the MLU values of each child in Figure 5.14. Here we can see that according to this measure, child RUChild5 would be considered as a *late speaker* because his MLU values stay rel-

atively low for a long time compared to the other children. The MLU values for child RUChild4 and RUChild3 do not show such a steep rising curve as for the other children. For child RUChild3 this can be explained by the fact that his recordings start when we was already 1162 days old. Child RUChild4, on the other hand also starts off with a very high MLU already at a young age (781 days). In contrast to child RUChild5, he would be considered as a *early talker*. Comparing development of MLU for the Russian mothers, we can see that as child RUChild4, the mother of RUChild4 also shows high MLU values. An interesting trend can be observed with the mother of RUChild3: her MLU rises considerably stronger when compared to the child RUChild3. As in the graph for the MLU values of the children, the mothers of RUChild3 and RUChild4 show the highest values. Interesting to note is also the difference between RUChild5 and the mother of RUChild5. Her MLU seems to be average when compared to the other mothers (it starts even higher than with the mother of RUChild1). However, child RUChild5 shows very low MLU values for quite a long time.

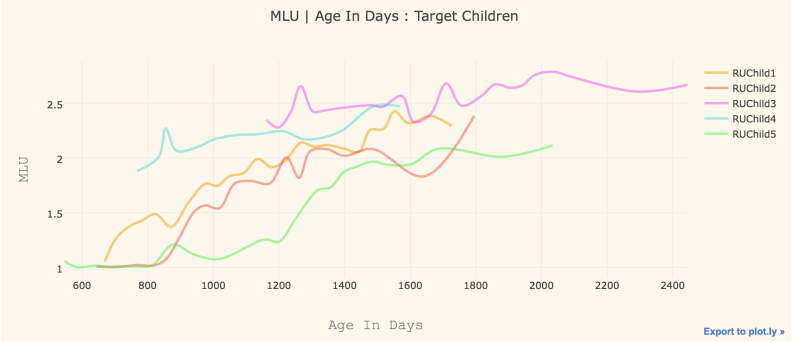

Figure 5.13.: MLU for Russian children

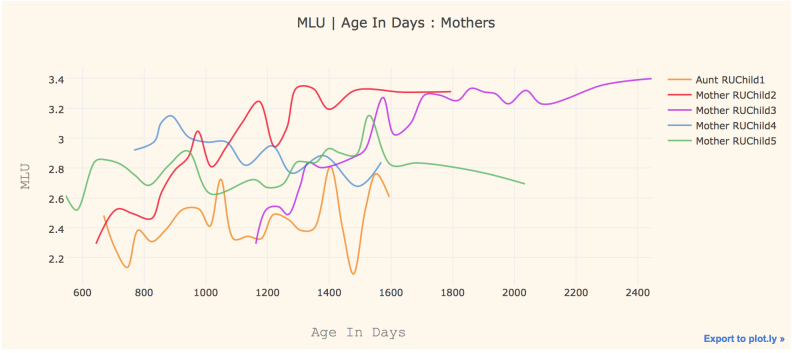

Figure 5.14.: MLU for Russian Mothers

When comparing the changes in the size of the networks and the average degree over time (shown in Figure 5.15 and 5.16), we can see that the children's networks grow in size and degree, indicating a growth in their vocabulary (cf. Ke and Yao 2008). An interesting development can be seen for child RUChild3: the size of his

network is smaller than for the networks of RUChild1 and RUChild4 in the age range after roughly 1300 days, but the average degree levels off at more or less the same range as for the other two children. This indicates that child RUChild3 is using his vocabulary more flexibly than RUChild1 and RUChild4 do.

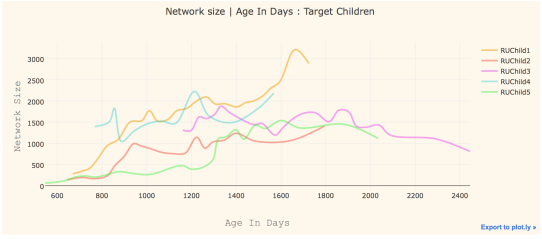

Figure 5.15.: N RU children

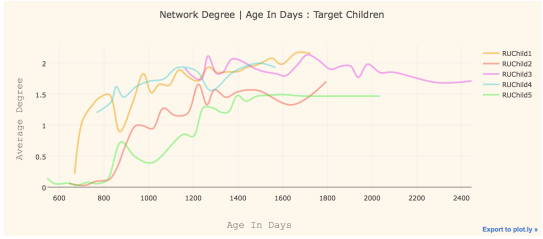

Figure 5.16.: <k>RU children

Compared to the graphs for the development of the network size and the average degree for the Russian mothers, we see that there is no clear evolution taking place as with the children. The network size fluctuates a lot more, whereas the average degree stays much more constant (except for the mother of RUChild1, who shows a very different pattern from all the other mothers. An explanation for this requires further investigation).

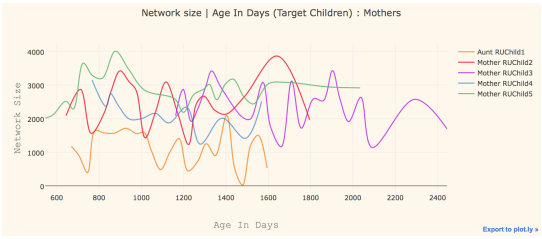

Figure 5.17.: N RU mothers

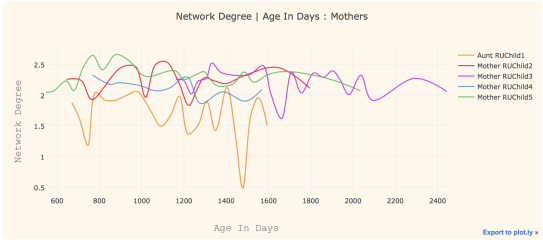

Figure 5.18.: <k>RU mothers

With regards to analyzing network properties, the above measures may only describe the global structure of the networks. But if we want to study the lexical development from a more syntactic perspective, we need to apply measures which account for changes in the internal structure of the networks. In-degree and out-degree centralization are such measures. Following the work of Adamo and Boylan 2008, we also measured the out-degree centralization for the different staged networks for every child. As can be seen in Figure 5.19 the first "peak" in the out-degree centralization as observed by Adamo and Boylan 2008 for the English child, also seems to occur in the data for the Russian children, where for most of them (except child RUChild1, whose peak happens a lot earlier)[32] show the first significant rise in out-degree centralization roughly at around the age of 850 days. A second "global"

[32]The initial very high out-degree value for child RUChild5 also needs further investigation.

rising seems to occur at around age 1200 days and 1500 days, before leveling off at an out-degree centralization between 0.6 and 0.8. The data of the mothers in Figure 5.20 again shows no rising trend, which also for our data supports findings from Ke and Yao [2008] and Adamo and Boylan [2008] that the mother's networks change less over time.

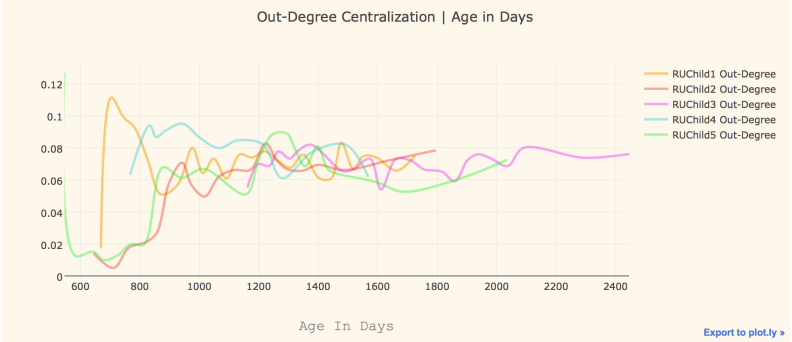

Figure 5.19.: Out-degree centralization for Russian children

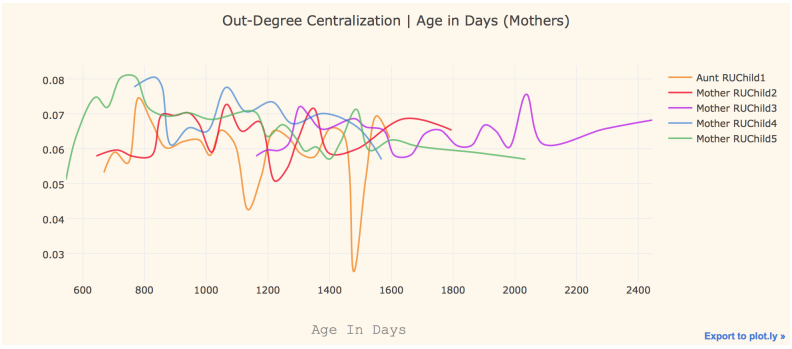

Figure 5.20.: Out-degree centralization for Russian mothers

Adamo & Boylan showed in their study that this initial peak for the out-degree centralization in a child's network happens when the first hubs appear which can combine with many other nodes. In the case of the English-speaking children in Adamo and Boylan's as well as Ke and Yao's 2008 study, these hubs are words which denote conjunctions, prepositions and pronouns (replacing hubs from earlier age stages which are mostly nouns or proper names). In our approach, we used the ages where the above peaks occur to partition our data into four further stages:

- **Stage 1**
  Before the initial peak, where we expect the words denoting hubs to be content words.
- **Stage 2**
  The phase after the first peak, from where we expect the hubs to change from being content words to more function words.

- **Stage 3**
  A further phase where we expect more function words to come in.
- **Stage 4**
  A last phase of levelling off, where we expect the hubs to roughly stay the same, but co-occur with many other words.

For a more global analysis of the development of hubs in the children's as well as the mothers' data, we used the ten highest ranking hubs of each stage and plotted them with a punchcard visualization, where the size of the circles denotes the frequency in how many sessions of our first grouping these hubs occurred. In Figure 5.21 on page 89, we can see that most of the words that only appear in stage 1 are also here mostly content words, names or "baby talk" such as *mama* 'mother', *mashina* 'car', *volk* 'wolf', *kuku* or the verb *daj* 'give' (IMP.2SG). On the other hand, words that already appear from stage 1 (or enter the network in stage 2) and stay throughout all the future stages are mostly prepositions, conjunctions and pronouns, such as *i* 'and', *a* 'and/but', *chto* 'that', *vot* 'here', *on* 'he', *ja* 'I'. Interesting is the change with the verb *datq* 'to give'. It appears as a hub in its perfective form in the first two stages, denoting the actual meaning of *give*, but then becomes a freuqent hub in its imperfective form from stage two onwards, where it is rather used as denoting an order/suggestion/prompt to do something (*daj* >*davaj*). When looking at the visualization for the mothers' hubs in Figure 5.22 on page 90, we can see that they use their hubs more consistently. This can be seen by the fact that many hubs that occur from stage 1 onwards, stay important hubs throughout the future stages as well. Hubs from stage 1 that later loose their status are again content words or names such as *Alja*, *kisa* 'kitten', but also the verb *idi* 'come/go' (IMP.2SG). Hubs that enter from stage 2 or 3 are particles, prepositions, adverbs, conjunctions and pronouns such as *vse* 'all', *tozhe* 'also', *my* 'we', but also other nouns.

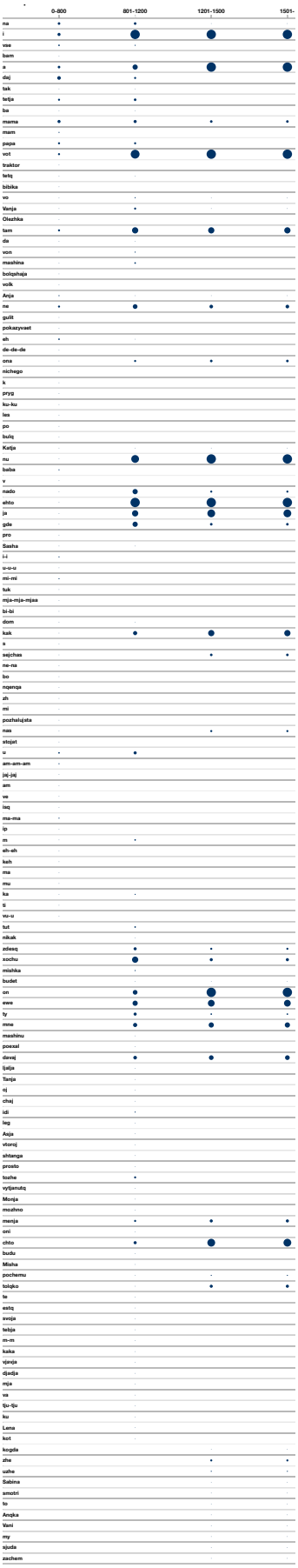

Figure 5.21.: Hubs development RU children

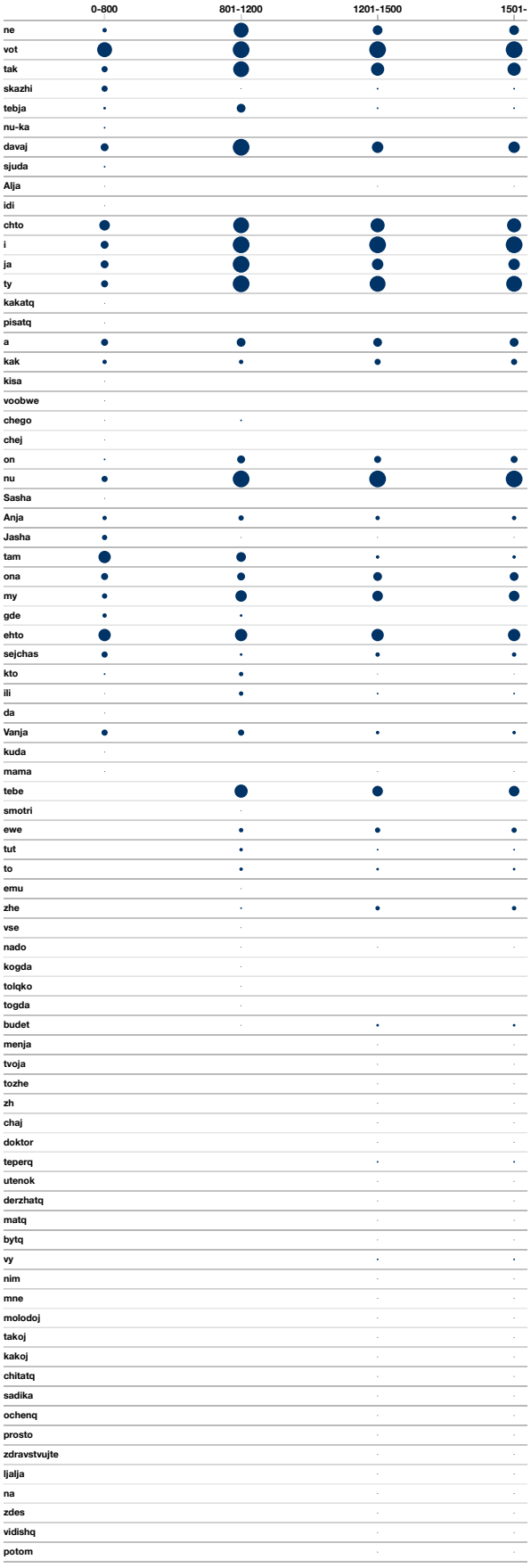

Figure 5.22.: Hubs development RU mmothers

After extracting the most important hubs for various stages defined by the out-degree centralization, our next question was: how do these hubs "behave" in the network, i.e. with which other nodes do they connect? To simulate that, we followed again the information-seeking visualization mantras and loaded the .gexf files into Gephi, in order to make the changes of the hubs visible over time for one Russian and one Chintang child. We used the `.gexf` files we created from the graphs and simulated the development of hubs for the child RUChild2 up to 800 days for all the hubs in her network, and from 800 days onwards for the hub *ne* ('not').[33]

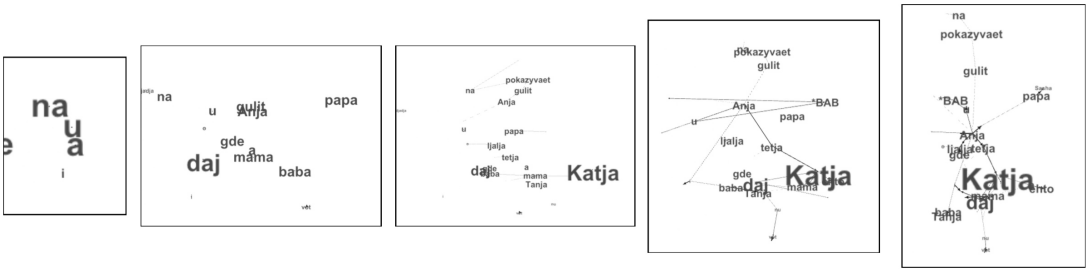

Figure 5.23.: Hubs development stage 1 child RUChild2

In Figure 5.23, we can see that the most important hubs are *mama, papa*, *baba* 'grandmother' and *daj* 'give'.[34] We can also observe a decrease in size for the node *daj* from picture 2 to 3 in Figure 5.23. This is due to the fact that other hubs enter the network, thus spreading the centralization more evenly across the network. Then, when the node *Katia* enters the network, it becomes the most important hub because it connects to many nodes of the network which themselves connect to other important nodes.

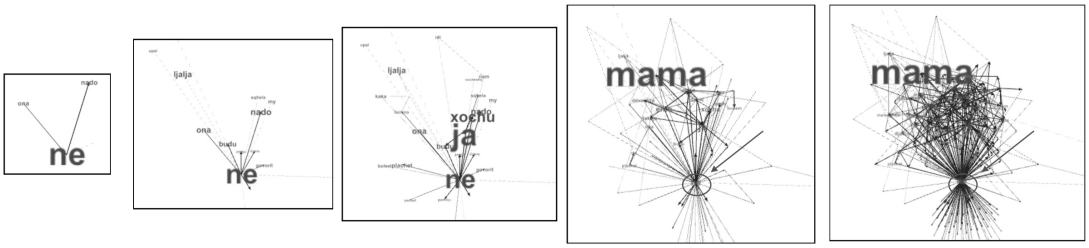

Figure 5.24.: Development hub *ne* child RUChild2

When we look at the development of a specific hub such as *ne* in Figure 5.24, we can also se the decrease in size due to the emergence of new hubs in the network. We see that this node already starts off in the combination *ne nado* 'it's not necessary'

[33]Due to the great amount of data, we decided to concentrate only on one specific hub.

[34]Note that the width of the links is drawn according to its weight (which was computed via the number of times an edge occurs between a staring and a target node). The links in picture a and two in Figure 5.23 are not visible here because their are very weak.

[35] and *ne ona* 'not she'. But from the second picture onwards, after the child has learned the first verbs, this particle from then on mostly connects to other verbs, which is particularly striking in the last picture in Figure 5.24.

When comparing the MLU values for the Chintang children in Figure 5.26, we can see that this parameter fluctuates a lot stronger than with the Russian children. This may be explained by the fact that very flexible word order and in Chintang and also by it being a pro-drop language. At an age between 1000 and 1120 days, the children LDCh1, LDCh2 and LDCh3 have MLU values between 1.9 and 2.1. The MLU of child LDCh4 fluctuates at this age much more, then lowers significantly until the age of 1640 days, and rises again afterwards. Also the MLU values for the Chintang adults do not really show a trend and also fluctuate a lot more than we have seen in the plot for the Russian mothers.

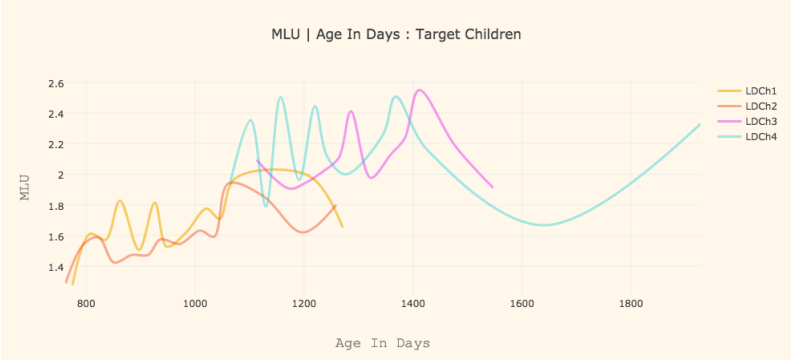

Figure 5.25.: MLU for Chintang children

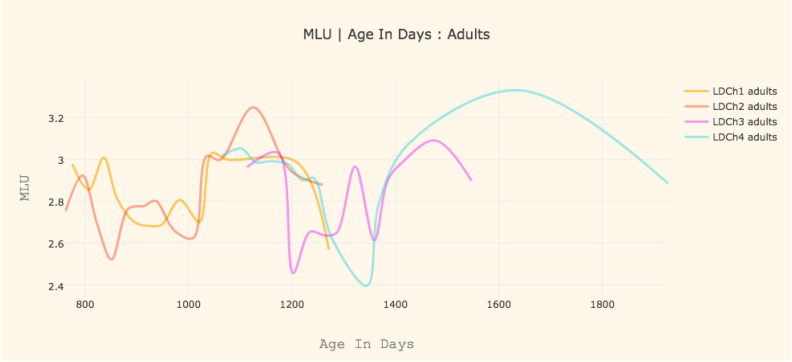

Figure 5.26.: MLU for Chintang adults

When we now also compare the changes in size and average degree of the Chintang children's networks (Figure 5.27), we can see that even though child LDCh4 has a considerably higher MLU, he/she does not have a very much bigger vocabulary (i.e. network size) than child LDCh1 and LDCh2 (in fact his/her network size even

[35]Here with the meaning "stop it".

decreases with time). Child LDCh3, on the other hand, starts off with the lowest vocabulary size, but his/her network grows a lot faster. Another interesting point can be seen when we look at the average degree of the Chintang children's networks in Figure 5.28. Here, child LDCh4's average degree is in a similar range as with child LDCh3, even though the network size of child LDCh4 is a lot smaller at the age of approximately 1380 days. This again indicates that child LDCh3 uses his/her vocabulary more flexibly compared to the other children. Compared to the Russian children, we see again a much stronger fluctuation.

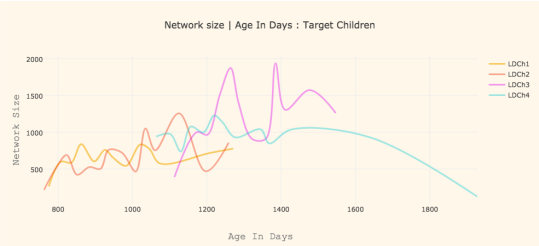

Figure 5.27.: N CTN children

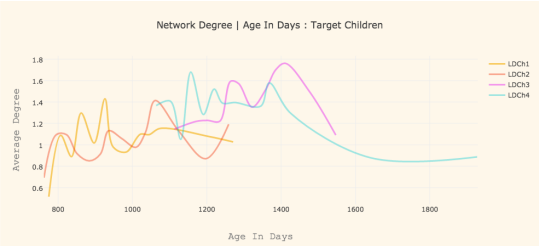

Figure 5.28.: <k>CTN children

When comparing the Network size of the Chintang adults, we can also se (as with the Russian mothers), that the adult speakers do not show a rise in their network size over time. An interesting fact to see is that the adults with the (generally) largest vocabulary (i.e. the network size) are the adults around child LDCh3, who also has the largest vocabulary of all the children. With regards to the network degree, both, the Chintang children and adults show a much higher fluctuation than can be seen for the Russian data (cf. Figure 5.15 - Figure 5.20).

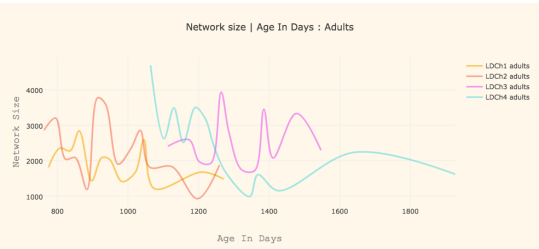

Figure 5.29.: N CTN kids

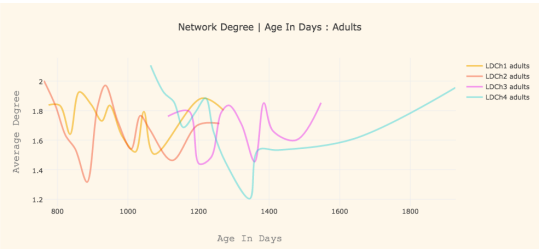

Figure 5.30.: <k>CTN adults

When we now look at how much the out-degree in the children's networks centralizes around specific nodes (Figure 5.31), we can see that child LDCh2 already starts off with a very high out-degree centralization compared to child LDCh1. For the children LDCh1, LDCh3 and LDCh4 the out-degree centralization seems to level off between 0.04 and 0.06, the data of child LDCh2 does not even out yet by the end of the recordings.

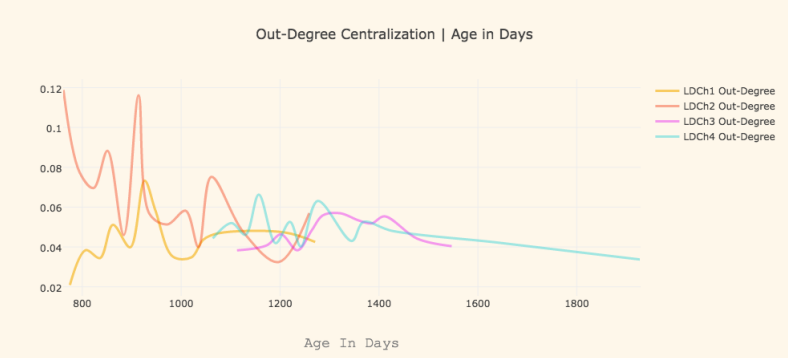

Figure 5.31.: Out-degree for Chintang children

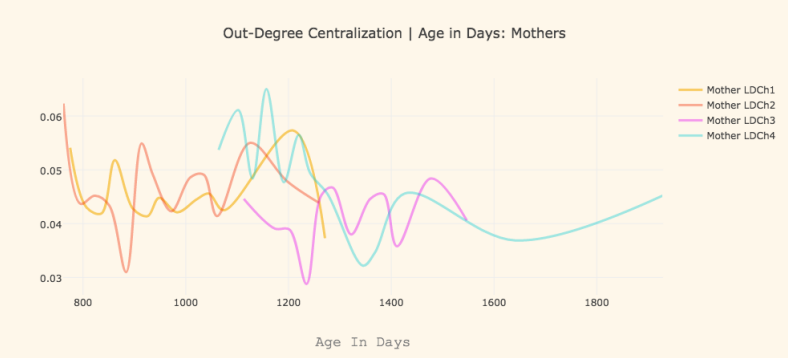

Figure 5.32.: Out-degree for Chintang adults

According to the development in out-degree centralization, we divided the data into the following three stages[36]:

- **Stage 1**
  The age range before the first burst in the data of child LDCh1 (0-820 days).
- **Stage 2**
  The age range including the highest peak and the following decrease (821-1060 days).
- **Stage 3**
  The age range after the last peak in the data of child LDCh2 (1061 days and beyond).

In the corresponding punchcard visualization (Figure 5.33 on page 96) we can see an interesting fact: As opposed to the hubs in stage 1 for the Russian children, the hubs of the Chintang children which already appeared in stage 1 mostly tend to stay hubs also in the future stages. Hubs that only appear in stage 1 are *mimum* 'little', *sune* (name) and *didi* 'elder sister'. When looking at the hubs for stage 2, there are more of them which do not appear in stage 3, mostly denoting names,

[36] We used three stages here and not four as in the Russian data because the hubs from stage 3 to stage 4 changed a lot less than in the Russian corpus.

demonstrative pronouns (both in more complex forms) and numbers. Nodes that become hubs only from stage 3 onwards are mostly verbs and again interjections, demonstrative pronouns and particles. The most frequent hubs in stage 3 are *ba* (proximity marker), *ta* (distance marker), *akka* 'wow' (interjection) and *lo* 'now' (interjection). When comparing the visualization for the adult Chintang speakers (Figure 5.34 on page 97), we see that many words which are hubs in stage 1 remain frequent hubs also in the future stages. These are also mostly demonstrative pronouns, particles, interjections, verbs and nouns, such as *to* (interjection) and *lusakha* 'dance' (IMP.2SG). Also many hubs that enter the networks in stage 2 reappear as hubs in stage 3. Hubs entering the networks in stage 2 stay again mostly and demonstrative pronouns, particles, adverbs, interjections, and in stage 3 words denoting hubs are demonstrative pronouns, verbs, interjections but also nouns (such as *cuwa* 'water', *cama* 'rice' and *gadi* 'vehicle'). This illustrates the importance of demonstrative pronouns in Chintang, which are mostly used to direct somebody's attention to something that is being talked about. Also their syntactic flexibility (i.e. that they can follow words other than nouns) and their use as interjections is mirrored in their frequent appearance as hubs. When we compare this visualization with the Russian adults (i.e. the mothers), we can see that in Chintang there are more different words denoting hubs than in Russian. Again this can be explained by the fact that Chintang has no article system, a very flexible word order, and is also a pro-drop language.[37]

[37]Note that the rise in frequency for the Chintang hubs from stage1 to stage3. This may lead us to think that the adults adapt their language to the children, however, in Chintang culture it's not very typical that the adults speak in a child like manner with the children. This is why this development needs further investigation.

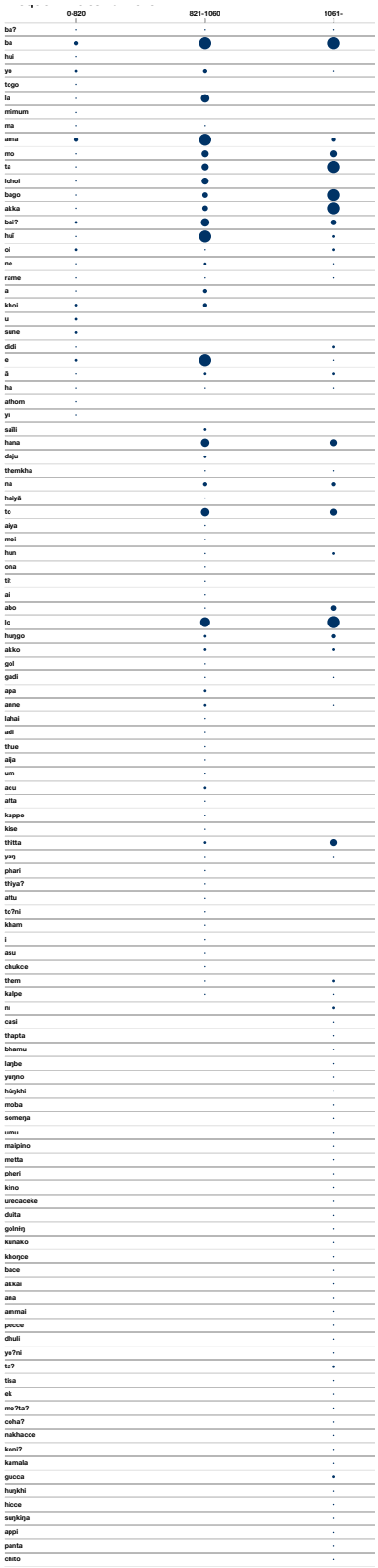

Figure 5.33.: Hubs development CTN children

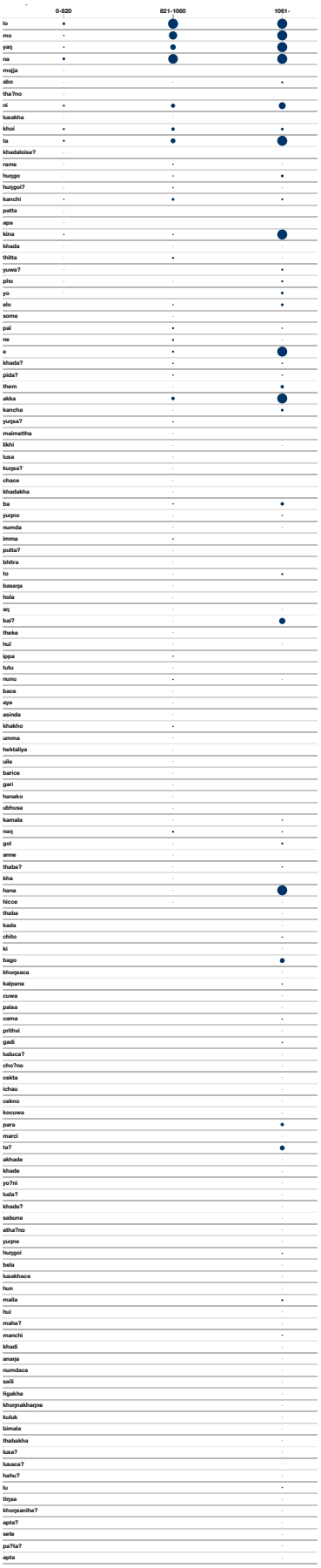

Figure 5.34.: Hubs development CTN adults

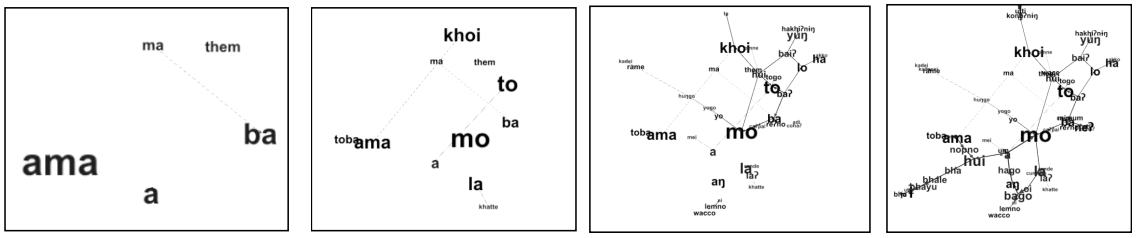

Figure 5.35.: Hubs stage 1 child LDCh1

When we look at LDCh1's network growth of stage 1 in Figure 5.35, we can also see an decrease in size for the nodes *ama* 'mama' and *ba* (marker for proximity), as soon as the nodes *mo* 'down', *to* 'up' and *khoi* 'where' enter the network. Other nodes that are added to the networks until the last picture in Figure 5.35 denote mostly names, demonstrative pronouns and interjections. Again, a great number of demonstrative pronouns and interjections in Chintang can be explained with the typological feautres mentioned above.

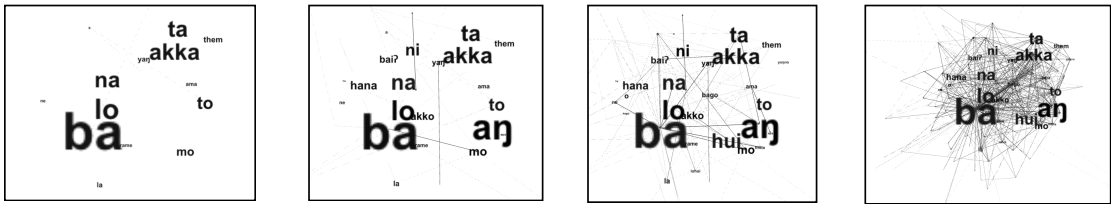

Figure 5.36.: Development hub *ba* LDCh1

With respect to the growth of a specific node, we can see that the node *ba* initially connects to few other demonstrative pronouns and then gradually also combines with nouns and particles, such as the pragmatic particle *ang*.

5.4.4. Discussion

In our study we have shown that our multilingual database allows us to conduct research on less studied languages, by applying analysis methods that can be used language independently. Following the work of Ke 2007, Ke and Yao 2008, as well as Adamo and Boylan 2008, we have investigated the global networks structure of Russian and Chintang lexical co-occurrence networks which also showed small-world characteristics in their global structure. With respect to the development of nodes denoting hubs, we were able to show that hubs in Russian tend to develop from nouns and proper names to mostly conjunctions, pronouns and prepositions. In Chintang, the most frequently used hubs are demonstrative pronouns, interjections and also pronouns and verbs. This can be explained by the fact that Chintang is a polysynthetic language, meaning that words are composed of many morphemes and

that there are many "sentence-like" words (i.e. single words that would correspond to whole sentences in more isolating languages). Given that there are languages in the ACQDIV database that share this feature (e.g. Turkish and Inuktitut), future research to apply network analysis on a morphological level would be a very interesting topic. Also, experimenting with using association measures and various context windows to create the lexical networks, instead of only using direct co-occurring words, might be an interesting direction for further research. What is more, following Shneiderman and Keim et al.'s *visualization mantra*, we added to our statistical analysis a visual approach, where we used the graph toolkit Gephi in order to make the development of lexical networks visible and also explorable. The many more features Gephi has to offer should also be further investigated in future resarch.

6. Conclusion

The various historical developments in the field of data visualization have changed the way we work with data to a great extent. Increasingly available computational power, but also new approaches to research in general have constantly shaped the way how we collect, process, analyze and also visualize data. New software tools and libraries have facilitated the way in which we can collect and manipulate data from various fields, but the growing number of data visualizations that are created with these tools also require a deeper reflection with respect to what benefits visualizations can bring. Benefits not only in the sense of presenting research results, but also in the sense of being part of the whole research process. By experimenting with various visualization forms in order to visualize language development over time, we followed a *visual analytics* approach by using interactive visualizations in order to acquaint ourselves better with a specific part of our ACQDIV data, namely Russian verbs and their development over time. We also studied theoretical concepts from data visualization more closely in order to decide upon and create effective explorative data graphics. Insights from research in human cognitive perception, as well as Gestalt theory has helped us in creating visualizations where patterns, which we first did not know how to visualize, suddenly became visible and analyzable. However, our work has also shown that visualizing linguistic data includes many challenges when trying to abstract linguistic information, which is inherently already an abstract concept by itself. Furthermore, the sheer amount of data linguists are often facing when working with textual data has proved to be challenging for many of the available data visualization softwares. This is, on the one hand, due to the fact that many visualizations are designed for illustrative (i.e. explanative) purposes, where the important parts of the data are already known. However, we hope to have shown that visualizations can also be used as a means of exploring the data for interesting, initially unknown patterns. Visualizing our data has not only helped us in detecting annotation and coding problems, it has even led us from one research question to a completely different research field. By purely experimenting with a network representation of our data, and by the collaborative work of many experts from various fields such as linguistics, statistics and computer science, we only discovered through this process new possibilities of how to combine expertise

from different fields into one project. Furthermore, the following quote by Ke [2007], summarizes very well our approach to combining child language acquisition research, data visualization and network theory:

> All in all, network research will bring a new perspective to linguistics, provide a new methodology to carry out quantitative analyses and suggest new questions and insights; at the same time, studies of networks about language will bring up new challenges to network research in general, and enrich the field with abundance of empirical data and questions. This is a cross-fertilization area worthy exploring (Ke [2007, 25]).

This quote illustrates very well the reciprocal influence of theoretical concepts and practical application, where the combination of new theories can lead to new applications, but experimenting with new applications can also result in extending (or also challenging) already existing theories.

Working for ACQDIV and the Visual Linguistics project has shown me many times how multifaceted the challenges in these two research fields are. Especially the field of child language acquisition (which was completely new to me when I started my thesis) turned out to be more challenging, but also more fascinating than I could have ever imagined. Therefore, I would like to close this thesis with a quote from Hoff [2013]), which was one of the first sentences I have read when starting this work:

> **Language acquisition is the New York City of the field of cognitive science: If you can make it there, you can make it anywhere** (Hoff [2013, 8]).

References

M. Adamo and S. Boylan. A network approach to lexical growth and syntactic evolution in child language acquistion. *Unpublished manuscript*, 2008.

L. A. N. Amaral, A. Scala, M. Barthelemy, and H. E. Stanley. Classes of small-world networks. *Proceedings of the national academy of sciences*, 97(21):11149–11152, 2000.

A.-L. Barabási and R. Albert. Emergence of scaling in random networks. *science*, 286(5439):509–512, 1999.

A.-L. Barabási and E. Bonabeau. Scale-free networks. *Scientific American*, 288(5): 50–59, 2003.

M. Bastian, S. Heymann, M. Jacomy, et al. Gephi: an open source software for exploring and manipulating networks. *ICWSM*, 8:361–362, 2009.

M. Bayer. Sqlalchemy-the database toolkit for python. *URL http://www. sqlalchemy. org/. Last accessed on the 20th of March*, 2016.

N. Beckage, L. Smith, and T. Hills. Small worlds and semantic network growth in typical and late talkers. 2011.

B. Bickel, S. Stoll, M. Gaenszle, N. Rai, E. Lieven, G. Banjade, T. Bhatta, N. Paudyal, J. Pettigrew, I. Rai, et al. Audiovisual corpus of the chintang language, including a longitudinal corpus of language acquisition by six children, paradigm sets, grammar sketches, ethnographic descriptions, and photographs. *DOBES Archive*, 2011.

M. Bostock. D3. js. *Data Driven Documents*, 2012.

S. Brain and L. Page. The anatomy of a large scale hypertextual web search engines. *Computer Networks and ISDN System30 (1-7)*, pages 107–117, 1998.

M. Breckon. Visuwords: Online graphical dictionary. *Reference Reviews*, 29(1): 36–36, 2015.

N. Bubenhofer. Projekt — visual linguistics, 2016. URL `http://www.cl.uzh.ch/de/research/visuallinguistics.html`. [Online; accessed 02-February-2016].

N. Bubenhofer, K. Rothenhäusler, K. Affolter, and D. Pajovic. Geocollocations: Visual analytics with a humanist's eye. 2016. submitted.

S. K. Card, J. D. Mackinlay, and B. Shneiderman. Using vision to think. In *Readings in information visualization*, pages 579–581. Morgan Kaufmann Publishers Inc., 1999.

C.-h. Chen, W. K. Härdle, and A. Unwin. *Handbook of data visualization*. Springer Science & Business Media, 2007.

C. M. Collins. A critical review of information visualizations for natural language, 2005.

B. Corominas-Murtra, S. Valverde, and R. V. Solé. The ontogeny of scale-free syntax networks through language acquisition.

G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5):1–9, 2006.

S. De Deyne and G. Storms. Word associations: Network and semantic properties. *Behavior Research Methods*, 40(1):213–231, 2008.

S. De Deyne, S. Verheyen, and G. Storms. Structure and organization of the mental lexicon: a network approach derived from syntactic dependency relations and word associations. *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*, page 47, 2016.

M. S. Dryer and M. Haspelmath. Wals online. *WALS Online. Max Planck Institute for Evolutionary Anthropology, Leipzig*, 2013.

L. Euler. Solutio problematis ad geometriam situs pertinentis. *Commentarii academiae scientiarum Petropolitanae*, 8:128–140, 1741.

M. Filiouchkina. How tense and aspect are acquired: a cross-linguistic analysis of child russian and english. *Nordlyd*, 32(1), 2005.

J. A. Fodor and Z. W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28(1-2):3–71, 1988.

D. Freudenthal, J. M. Pine, J. Aguado-Orea, and F. Gobet. Modeling the developmental patterning of finiteness marking in english, dutch, german, and spanish using mosaic. *Cognitive Science*, 31(2):311–341, 2007.

M. Friendly and D. J. Denis. Milestones in the history of thematic cartography, statistical graphics, and data visualization. *U RL http://www. datavis. ca/milestones*, 2001.

N. Gagarina. The acquisition of aspectuality by russian children: the early stages. *ZAS Papers in Linguistics*, 15:232–246, 2000.

E. Gegov, F. Gobet, M. Atherton, D. Freudenthal, and J. Pine. Modelling language acquisition in children using network theory. 2011.

R. Goldstein and M. S. Vitevitch. The influence of clustering coefficient on word-learning: how groups of similar sounding words facilitate acquisition. *Frontiers in psychology*, 5, 2014.

S. T. Gries and S. Stoll. Finding developmental groups in acquisition data: variability-based neighbour clustering. *Journal of Quantitative Linguistics*, 16 (3):217–242, 2009.

A. Gvozdev. Formirovaniye u rebenka grammaticheskogo stroya russkogo yazyka (2 parts) izd. *Akademii Padag. Nauk RSFSR, Moscow*, 1949.

A. Hagberg, D. Schult, and P. Swart. Networkx: Python software for the analysis of networks. Technical report, Technical report, Mathematical Modeling and Analysis, Los Alamos National Laboratory, 2005. http://networkx. lanl. gov, 2005.

M. Hearst. *Search user interfaces*. Cambridge University Press, 2009.

D. H. Hepting. The history of a picture's worth. `http://www2.cs.uregina.ca/~hepting/research/web/words/history.html`, 2008. [Online; accessed 12-March-2016].

T. T. Hills, M. Maouene, J. Maouene, A. Sheya, and L. Smith. Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychological Science*, 20(6):729–739, 2009.

E. Hoff. *Language development*. Cengage Learning, 2013.

C.-c. Huang. *Child language acquisition of temporality in Mandarin Chinese*. na, 2006.

R. F. i Cancho and R. V. Solé. The small world of human language. *Proceedings of the Royal Society of London B: Biological Sciences*, 268(1482):2261–2265, 2001.

R. F. i Cancho, R. V. Solé, and R. Köhler. Patterns in syntactic dependency networks. *Physical Review E*, 69(5):051915, 2004.

R. Ihaka and R. Gentleman. R: a language for data analysis and graphics. *Journal of computational and graphical statistics*, 5(3):299–314, 1996.

W. H. Inmon. *Building the data warehouse*. John wiley & sons, 2005.

A. A. Karavanov. *Vidy russkogo glagola - značenie i upotreblenie: praktičeskoe posobie dlja inostrancev, izučajuščix russkij jazyk*. Russkij jayzk. Kursy, 2008.

J. Ke. Complex networks and human language. *arXiv preprint cs/0701135*, 2007.

J. Ke and Y. Yao. Analysing language development from a network approach*. *Journal of Quantitative Linguistics*, 15(1):70–99, 2008.

D. Keim, G. Andrienko, J.-D. Fekete, C. Görg, J. Kohlhammer, and G. Melançon. *Visual analytics: Definition, process, and challenges*. Springer, 2008.

D. A. Keim, J. Kohlhammer, G. Ellis, and F. Mansmann. *Mastering the information age-solving problems with visual analytics*. Florian Mansmann, 2010.

A. Kibort. Aspect. grammatical features. `http://www.grammaticalfeatures.net/features/aspect.html`, 2008. [Online; accessed 13-March-2016].

R. Kimball and M. Ross. *The data warehouse toolkit: the complete guide to dimensional modeling*. John Wiley & Sons, 2011.

J. M. Kleinberg. Authoritative sources in a hyperlinked environment. *Journal of the ACM (JACM)*, 46(5):604–632, 1999.

M.-J. Kraak. *RTF] Mapping Time: Illustrated by Minard's Map of Napoleon's Russian Campaign of 1812*. 2014.

S. Krämer. *Operative Bildlichkeit. Von der ‚Grammatologie 'zu einer ‚Diagrammatologie '? Reflexionen über erkennendes ‚Sehen '*. na, 2009.

S. Krämer. Diagrammatisch. `https://rheinsprung11.unibas.ch/fileadmin/documents/Edition_PDF/Ausgabe05/Glossar_Kraemer.pdf`, 2013. [Online; accessed 13-March-2016].

E. Kruja, J. Marks, A. Blair, and R. Waters. A short note on the history of graph drawing. In *Graph Drawing*, pages 272–286. Springer, 2001.

W. F. Leopold. 49. speech development of a bilingual child. 4 vols, 1939.

W. F. Leopold. *Speech Development of a Bilingual Child. A Linguist's Record.…: Sound-learning in the First Two Years. II*. Northwestern University Press, 1947.

M. P. Lewis, G. F. Simons, and C. D. Fennig. *Ethnologue: Languages of the world*, volume 16. SIL international Dallas, TX, 2009.

H. Liu and J. Cong. Language clustering with word co-occurrence networks based on parallel texts. *Chinese Science Bulletin*, 58(10):1139–1144, 2013.

V. Lyding. Visualizing linguistic data: From principles to toolkits for doing it yourself. `http://www.eurac.edu/en/research/autonomies/commul/projects/Documents/Linfovis/AVML-workshop-York.pdf`, 2012. [Online: accessed 12-March-2016].

B. MacWhinney. The childes project: Tools for analyzing talk. *Mahwah, NJ & London: Lawrence Erlbaum*, 2000a.

B. MacWhinney. *The CHILDES project: The database*, volume 2. Psychology Press, 2000b.

A. Masucci and G. Rodgers. Network properties of written human language. *Physical Review E*, 74(2):026102, 2006.

C. R. B. F. McCarthy, Lauren. *Make: Getting Started with p5.js*. Maker Media, Inc, 2015.

W. McKinney. *Python for data analysis: Data wrangling with Pandas, NumPy, and IPython*. " O'Reilly Media, Inc.", 2012.

A. Mehler, B. Frank-Job, P. Blanchard, and H.-J. Eikmeyer. Sprachliche netzwerke. In *Netzwerkanalyse und Netzwerktheorie*, pages 413–427. Springer, 2010.

A. Mehler, A. Lücking, S. Banisch, P. Blanchard, and B. Job. *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*. Springer, 2015.

I. Meirelles. *Design for information: an introduction to the histories, theories, and best practices behind effective information visualizations*. Rockport publishers, 2013.

R. Mihalcea and D. Radev. *Graph-based natural language processing and information retrieval*. Cambridge University Press, 2011.

S. Milgram. The small world problem. *Psychology today*, 2(1):60–67, 1967.

G. A. Miller. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41, 1995.

S. Moran and R. Wright. Phonetics information base and lexicon (phoible). *Online: http://phoible. org*, 2009.

S. Moran, R. Schikowski, D. Pajovic, H. Casim, and S. Sabine. The acqdiv database: Min(d)ing the ambient language. In *The ACQDIV Database: Min(d)ing the Ambient Language, 2016. Proceedings of LREC 2016.*, to appear.

A. E. Motter, A. P. de Moura, Y.-C. Lai, and P. Dasgupta. Topology of the conceptual network of language. *Physical Review E*, 65(6):065102, 2002.

J. E. Murdoch. *Antiquity and the middle ages*. Macmillan Reference USA, 1984.

J. Nichols, A. Witzlack-Makarevich, and B. Bickel. The autotyp genealogy and geography database: 2013 release. *Zurich: University of Zurich*, 2013.

A. Ninio. Syntactic networks, do they contribute valid information on syntactic development in children?. comment on. *Physics of life reviews*, 11:632–634, 2014.

D. A. Norman. *Things that make us smart: Defending human attributes in the age of the machine*. Basic Books, 1993.

M. R. Quillian. Semantic memory, 1968.

R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2008. URL `http://www.R-project.org`. ISBN 3-900051-07-0.

R. Rollinger and C. Ierna. Christian von ehrenfels. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall 2015 edition, 2015.

L. Ryan. *The Visual Imperative: Creating a Visual Culture of Data Discovery*. Morgan Kaufmann, 2016.

M. Scaife and Y. Rogers. External cognition: how do graphical representations work? *International journal of human-computer studies*, 45(2):185–213, 1996.

Y. Shirai and R. W. Andersen. The acquisition of tense-aspect morphology: A prototype account. *Language*, pages 743–762, 1995.

B. Shneiderman. The eyes have it: A task by data type taxonomy for information visualizations. In *Visual Languages, 1996. Proceedings., IEEE Symposium on*, pages 336–343. IEEE, 1996.

D. I. Slobin. *The crosslinguistic study of language acquisition*, volume 4. Psychology Press, 2014.

A. E. Smith and M. S. Humphreys. Evaluation of unsupervised semantic mapping of natural language with leximancer concept mapping. *Behavior Research Methods*, 38(2):262–279, 2006.

R. V. Solé, B. Corominas-Murtra, S. Valverde, and L. Steels. Language networks: Their structure, function, and evolution. *Complexity*, 15(6):20–26, 2010.

R. Spence. *Information visualization*, volume 1. Springer, 2001.

C. Stern and W. Stern. Die kindersprache. *Leipzig: Barth*, 1907.

M. Steyvers and J. B. Tenenbaum. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive science*, 29(1): 41–78, 2005.

S. Stoll and B. Bickel. Capturing diversity in language acquisition research. *Language Typology and Historical Contingency: In Honor of Johanna Nichols. Amsterdam: John Benjamins*, pages 195–216, 2013.

S. Stoll and M. Roland. Audio-visional longitudinal corpus on the acquisition of russian by 5 children. 2008.

S. Stoll, E. Lieven, H. Winskel, and P. Pradakannaya. Studying language acquisition cross-linguistically. *South and Southeast Asian psycholinguistics*, pages 19–35, 2014.

S. E. Stoll. *The acquisition of Russian aspect*. PhD thesis, University of California, Berkeley, 2001.

T. Taylor. Principles of data visualization - what we see in a visual. `http://www.fusioncharts.com/whitepapers/downloads/Principles-of-Data-Visualization.pdf`, 2014. [Online; accessed 7-March-2016].

L. Tesnière. *Eléments de syntaxe structurale*. Librairie C. Klincksieck, 1959.

A. L. Theakston, E. V. Lieven, J. M. Pine, and C. F. Rowland. The role of performance limitations in the acquisition of verb-argument structure: An alternative account. *Journal of child language*, 28(01):127–152, 2001.

M. Tomasello. Do young children have adult syntactic competence? *Cognition*, 74 (3):209–253, 2000.

E. R. Tufte and P. Graves-Morris. *The visual display of quantitative information*, volume 2. Graphics press Cheshire, CT, 1983.

E. R. Tufte and D. Robins. Visual explanations, 1997.

J. W. Tukey. Exploratory data analysis. 1977.

Z. Vendler. Verbs and times. *The philosophical review*, 66(2):143–160, 1957.

M. S. Vitevitch. What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, 51(2):408–422, 2008.

D. J. Watts and S. H. Strogatz. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442, 1998.

J. A. Wise. The ecological approach to text visualization. *Journal of the Association for Information Science and Technology*, 50(13):1224, 1999.

J. A. Wise, J. J. Thomas, K. Pennock, D. Lantrip, M. Pottier, A. Schur, and V. Crow. Visualizing the non-visual: spatial analysis and interaction with information from text documents. In *Information Visualization, 1995. Proceedings.*, pages 51–58. IEEE, 1995.

H. B. Woolf. *The Merriam-webster dictionary*. Pocket Books, 1974.

Y. Zhang, X. Jin, X. Shen, J. Zhang, and E. Hoff. Correlates of early language development in chinese children. *International Journal of Behavioral Development*, 32(2):145–151, 2008.

M. Zortea, B. Menegola, A. Villavicencio, and J. F. d. Salles. Graph analysis of semantic word association among children, adults, and the elderly. *Psicologia: Reflexão e Crítica*, 27(1):90–99, 2014.

A. Age Ranges Russian Children

	Age Ranges				
Age Stage	**RUChild1**	**RUChild2**	**RUChild3**	**RUChild4**	**RUChild5**
1	1;08.10 - 1;09.09	3;01.15 - 3;02.17	3;01.08 - 3;02.02	1;11.28 - 2;00.30	1;05.00 - 1;06.19
2	1;09.15 - 1;10.13	3;02.25 - 3;03.23	3;02.06 - 3;02.25	2;01.06 - 2;02.27	1;08.15 - 1;09.26
3	1;10.24 - 1;11.18	3;04.24 - 3;06.10	3;03.01 - 3;03.23	2;03.04 - 2;03.23	1;10.26 - 2;02.23
4	1;11.24 - 2;00.15	3;06.17 - 3;07.24	3;04.08 - 3;04.24	2;03.30 - 2;04.19	2;03.05 - 2;05.02
5	2;00.22 - 2;01.12	3;08.07 - 3;09.06	3;04.28 - 3;05.18	2;04.26 - 2;06.05	2;05.11 - 2;06.28
6	2;01.20 - 2;02.18	3;09.14 - 3;10.14	3;05.25 - 3;06.17	2;06.12 - 2;07.06	2;07.11 - 2;08.27
7	2;02.28 - 2;03.20	3;10.24 - 4;00.12	3;06.27 - 3;07.14	2;07.19 - 2;08.28	2;09.11 - 2;11.11
8	2;03.27 - 2;04.28	4;00.20 - 4;01.12	3;07.19 - 3;08.07	2;09.14 - 2;10.19	2;11.21 - 3;01.24
9	2;05.04 - 2;06.13	4;01.16 - 4;02.08	3;08.16 - 3;09.14	2;11.00 - 3;00.05	3;02.00 - 3;02.28
10	2;06.21 - 2;07.29	4;02.15 - 4;03.13	3;09.20 - 3;10.24	3;00.16 - 3;02.01	3;03.04 - 3;04.04
11	2;08.03 - 2;08.23	4;03.22 - 4;04.16	3;10.31 - 4;00.12	3;02.13 - 3;03.15	3;04.11 - 3;05.08
12	2;09.01 - 2;09.24	4;04.22 - 4;05.10	4;00.20 - 4;01.12	3;03.26 - 3;04.23	3;05.19 - 3;06.18
13	2;10.00 - 2;10.22	4;05.21 - 4;06.19	4;01.16 - 4;02.15	3;05.08 - 3;06.06	3;06.21 - 3;07.11
14	2;11.07 - 2;11.26	4;06.26 - 4;07.24	4;02.28 - 4;04.02	3;06.14 - 3;09.12	3;07.13 - 3;08.05
15	3;00.06 - 3;01.02	4;08.07 - 4;09.06	4;04.09 - 4;04.28	3;10.01 - 3;11.30	3;08.12 - 3;09.05
16	3;01.09 - 3;02.01	4;09.18 - 4;10.25	4;05.03 - 4;06.11	4;01.00 - 4;02,08	3;09.16 - 3;10.06
17	3;02.14 - 3;03.04	4;11.01 - 4;11.26	4;06.19 - 4;07.17	4;02.24 - 4;03.22	3;10.14- -3;11.04
18	3;03.18 - 3;04.18	5;00.07 - 5;01.03	4;07.24 - 4;08.28		3;11.17 - 4;00.16
19	3;04.30 - 3;05.23	5;01.17 - 5;02.10	4;09.06 - 4;10.02		4;00.24 - 4;01.14
20	3;06.06 - 3;06.27	5;02.16 - 5;03.06	4;10.25 - 4;11.19		4;01.24 - 4;01.34
21	3;07.10 - 3;08.02	5;03.11 - 5;04.10	4;11.26 - 5;00.27		
22	3;08.10 - 3;09.01	5;04.14 - 5;05.18	5;01.03 - 5;02.03		
23	3;09.21 - 3;10.23	5;05.26 - 5;06.30	5;02.10 - 5;03.00		
24	3;11.01 - 3;11.21	5;07.10 - 5;08.13	5;03.06 - 5;03.29		
25	3;11.27 - 4;00.17	5;08.21 - 5;11.22	5;04.10 - 5;05.01		
26	4;00.24 - 4;01.19	6;00.27 - 6;05.13	5;05.18 - 5;06.30		
27	4;01.26 - 4;02.16	6;06.10 - 6;07.22	5;07.10 - 5;08.13		
28	4;02.25 - 4;03.13		5;08.21 - 6;00.27		
29	4;03.27 - 4;04.15		6;01.29 - 6;06.10		
30	4;05.00 - 4;05.25		6;07.11 - 6;07.21		
31	4;06.03 - 4;07.04				
32	4;07.11 - 4;08.10				
33	4;08.21 - 4;08.31				

Table A.1.: Age Ranges Russian Children

B. Networks: Degree Distributions

B.1. Degree Distributions CDS All Languages

B.1.1. Cummulative Degree Distribution of CDS Networks

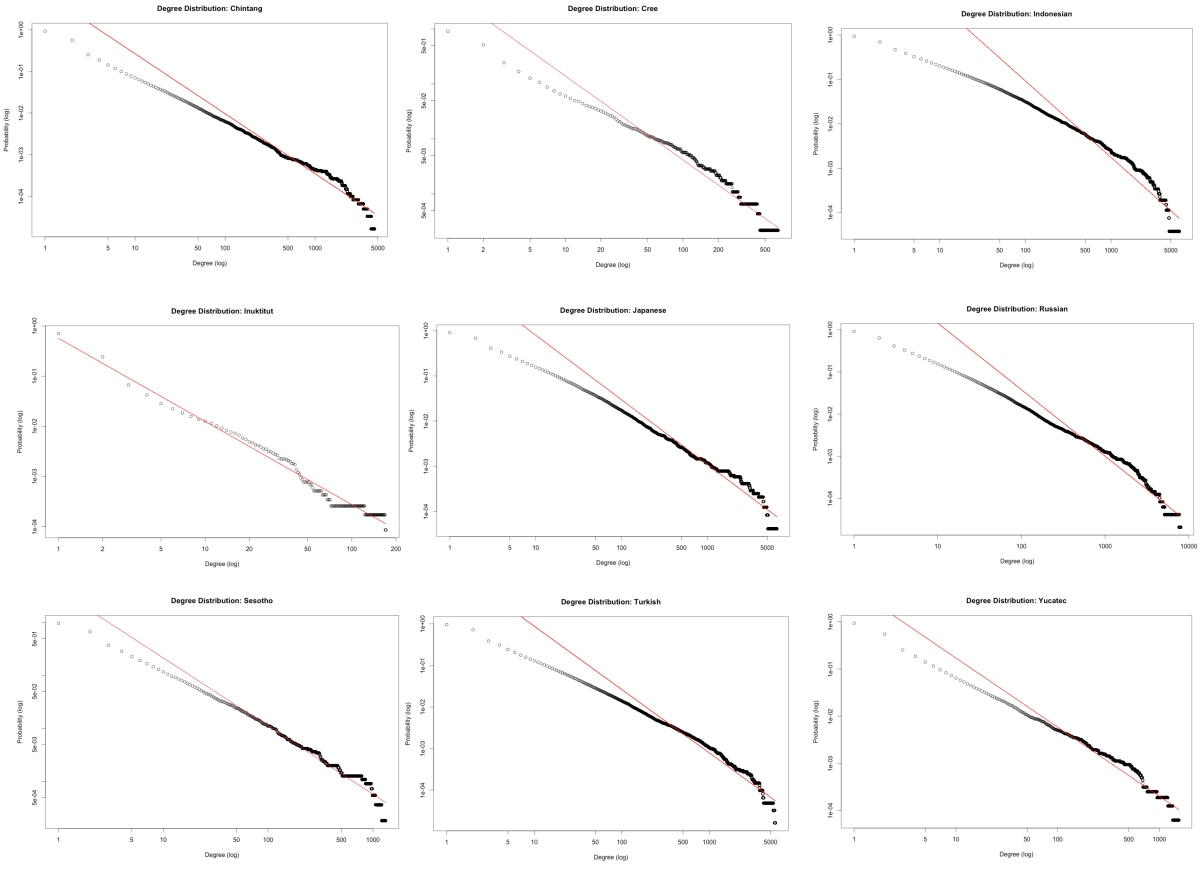

Figure B.1.: Fitted power-law distribution for all languages

B.1.2. Cummulative Degree Distribution of RU Networks (children)

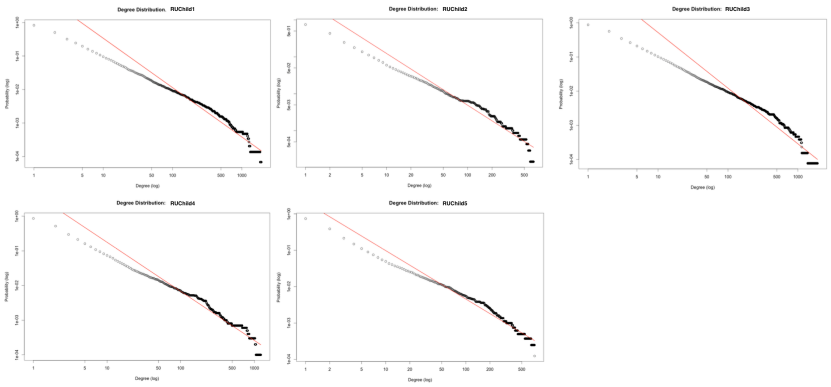

Figure B.2.: Fitted power-law distribution for Russian children

B.1.3. Cummulative Degree Distribution of CTN Networks (children)

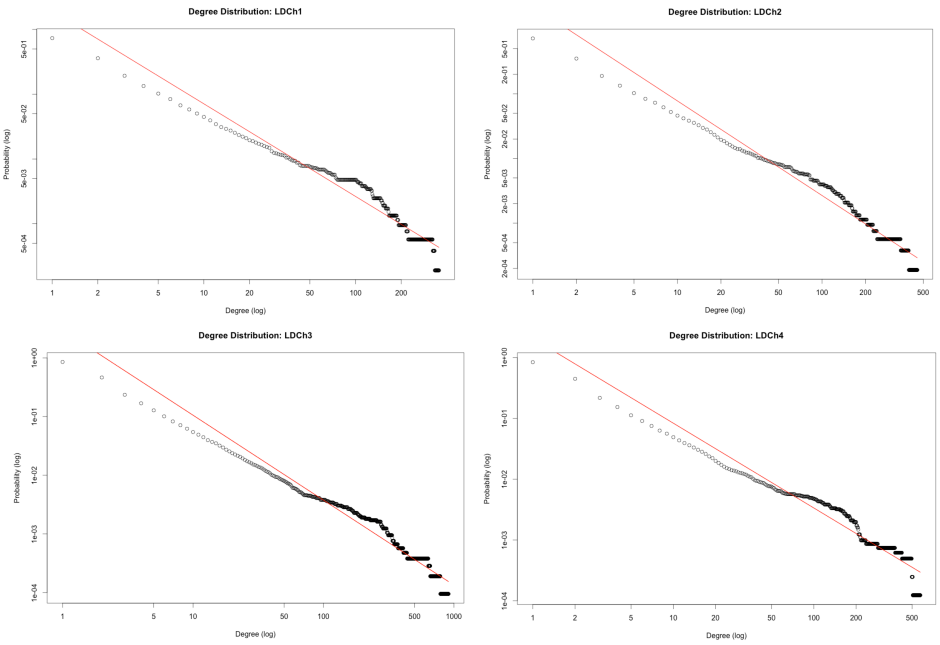

Figure B.3.: Fitted power-law distribution for Chintang children

C. Networks: ACQDIVIZ Package

The `ACQDIVIZ` package contains two main folders: `acqdiviz_web`, which contains the interactive visualizations from chapter 4, as well as `networks`, which contains the various Python scripts and .gexf and .csv files generated for the studies in chapter 5.

Universität Zürich[UZH]

Philosophische Fakultät
Studiendekanat

Universität Zürich
Studiendekanat
Rämistr. 69
CH-8001 Zürich
www.phil.uzh.ch

Selbstständigkeitserklärung

Hiermit erkläre ich, dass
die Masterarbeit von mir selbst und ohne unerlaubte Beihilfe verfasst worden ist und ich die Grundsätze wissenschaftlicher Redlichkeit einhalte (vgl. dazu:
http://www.lehre.uzh.ch/plagiate/20110314_LK_Merkblatt Plagiat.pdf).

Zürich, den 22.3.2016

Ort und Datum Unterschrift

Figure C.1.: Selbstständigkeitserklärung