Bochumer
Linguistische
Arbeitsberichte
27

Exploring growing orthographic networks built from longitudinal spelling data of German primary school students

Joshua Wieler

Bochumer Linguistische Arbeitsberichte

Herausgeberin: Stefanie Dipper

Die online publizierte Reihe „Bochumer Linguistische Arbeitsberichte“ (BLA) gibt in unregelmäßigen Abständen Forschungsberichte, Abschluss- oder sonstige Arbeiten der Bochumer Linguistik heraus, die einfach und schnell der Öffentlichkeit zugänglich gemacht werden sollen. Sie können zu einem späteren Zeitpunkt an einem anderen Publikationsort erscheinen. Der thematische Schwerpunkt der Reihe liegt auf Arbeiten aus den Bereichen der Computerlinguistik, der allgemeinen und theoretischen Sprachwissenschaft und der Psycholinguistik.

The online publication series “Bochumer Linguistische Arbeitsberichte” (BLA) releases at irregular intervals research reports, theses, and various other academic works from the Bochum Linguistics Department, which are to be made easily and promptly available for the public. At a later stage, they can also be published by other publishing companies. The thematic focus of the series lies on works from the fields of computational linguistics, general and theoretical linguistics, and psycholinguistics.

Band 27 (April 2026)

Herausgeberin: Stefanie Dipper
Sprachwissenschaftliches Institut
Ruhr-Universität Bochum
Universitätsstr. 150
44801 Bochum

Erscheinungsjahr 2026
ISSN **2190-0949**

Joshua Wieler

Exploring growing orthographic networks built from longitudinal spelling data of German primary school students

---

2026

Bochumer Linguistische Arbeitsberichte

(BLA 27)

Joshua Wieler

joshua.wieler@ruhr-uni-bochum.de

Exploring growing orthographic networks built from longitudinal spelling data of German primary school students

Abstract

Individual differences in language learning arise very early in life (Mani & Ackermann, 2018). Assessing such differences might help us to understand the mechanisms that drive learning on the level of the individual as to potentially provide informed guidance to those who are in need of it. The intention of this study is to analyze the still developing orthographic knowledge of primary school students living in Germany. I utilize the tools of network science to set up individual networks for different children at several points in time. By assessing these networks visually as well as quantitatively, I aim to characterize factors that drive the growth of the orthographic network. Additionally, a thorough assessment of the methodology will be provided, alongside a discussion of the limitations I encountered during the analysis.

1 Introduction

The ability to write down thoughts for others to read at any point in time is said to be a defining element of what makes society (Treiman & Kessler, 2014). It comes as no surprise then that as soon as children enter school, much effort is put into teaching them how to read and spell. Acquiring these skills does not go the same for every child, however, as some learn faster, while others take a little more time and/or are in need of special attention. In this study I aim to explore the differences between early spellers in German primary schools from two complementary angles. Using the toolkit of networks science, I will first construct visual representations of orthographic networks for individual children throughout time. After inspection of the networks, I will conduct a statistical analysis of their properties, for which I will consider findings from the visual analysis.

Recognizing differences between children at a young age might be especially important, because, as studies have found, further development is influenced by such differences (see Kidd et al., 2018, for a review on the matter). Mani and Ackermann (2018), for example, have reported that by 20 months of age, there already exists high variability in the words that different children know. Although the most extreme differences are mitigated somewhat over time (i.e., children will typically learn more frequent words at some point, even though the time of acquisition may differ), Mani and Ackermann argued that the early lexicon still impacts further word leaning.

This notion ties into one account of *preferential attachment* proposed by Steyvers and Tenenbaum (2005). Steyvers and Tenenbaum conducted a computer simulation in which they demonstrated that a network model of a mental lexicon whose growth is primarily driven by early acquired words yields a structure very similar to those observed in semantic networks build from empirical data.

In this study, I aim to extent what was found for word learning to orthography acquisition. One crucial difference is, of course, that children who have just started to learn how to spell already know a lot of words. Also, words that they chose to produce in their writing should at the very least have a conceptual as well as phonological representation in memory. On the other hand, the orthographic representation of these words might not be fully developed yet, if at all, which could lead to errors. Another possibility would be that children simply refuse to produce spellings they are unsure about.

That German orthography mostly derives from a graphematic system which comprises many rules and principles was shown by Eisenberg (2020). Importantly, this system is not arbitrary, but rather systematic. For example, so-called *phonographic spellings* are the result of simply mapping each phoneme of a spoken word to a corresponding grapheme in script. For other words, higher order principles must come into play to produce the correct spelling. These principles, for example, index tense or lax vowels – e.g., vowel-lengthening <h> or consonant-doubling, respectively – or retain the spelling of morphemes throughout different inflections as is the case in *Hund* ([hʊnt], 'dog') and *Hunde* ([hʊndə], 'dogs')[1].

As German orthography is systematic, children might be inclined to primarily produce words that follow similar principles which they have already acquired. Considering findings from the preferential attachment account in semantic network growth (Steyvers & Tenenbaum, 2005), the overarching question of this study then is whether orthographic networks that exhibit higher connectivity in regard to orthographic similarity between words inside the network, also exhibit higher lexical diversity, and whether this leads to increased growth of the orthographic network throughout time.

This report is structured as follows: After the introduction, in chapter 2 I will introduce the Litkey Corpus and provide information on the sample I use in here. The subset will only include

[1]To provide more detail, in *Hund* the final letter <d> is devoiced in speech. In script, the corresponding phoneme /t/ typically maps onto the grapheme <t>. As in the plural form *Hunde*, <d> is not devoiced, the spelling with <d> instead of <t> is retained in both forms of this lemma. This phenomena is referred to as *morpheme constancy*.

data from children who have participated in all ten test points. I furthermore decided to exclude all function words from the subset. The rational leading to this decision will be given in section 2.1.

In chapter 3, I introduce the method to calculate and visualize per child, per test point networks. I will utilize the R library *igraph* (Csardi & Nepusz, 2006) to set up individual orthographic networks for each child that has participated in each of the ten test points. In the end, there will be ten networks per child, one for each test point.

By plotting these networks in sequence I aim to visually outline the development in spelling ability of the children throughout their time in primary school. This will be done in chapter 4. I will observe a selection of the networks in search of patterns that might tell us something about the mechanisms that drive learning.

Finally, in chapter 5, I conduct a statistical analysis of the properties elicited from the individual networks. For example, I will examine the role that a network's size at a certain time in development plays as well as the connectivity inside the network. Development will be operationalized as the network's growth between consecutive test points. Chapter 5 will be followed by a discussion of the results in chapter 6.

Before going into chapter 2, please note that there is an R notebook that incorporates much of this report as well as all code (alongside comments) that explicate and provide further detail on the steps I took in this inquiry.

2 What is the Litkey Corpus?

The Litkey Corpus (Laarmann-Quante et al., 2019), which is the basis of my analysis, is well suited to investigate individual differences between children in early stages of learning to spell. The corpus contains freely written texts of German primary school students, collected at separate test points between 2010 and 2012 by Frieg (2014). At each test point children were instructed to write a story according to a sequence of pictures following specific topics. This fact

should place a constraint on the lexical diversity in the data. For example, a picture story with the topic *Schule* ('school') makes usage of words like *Lehrer* ('teacher') or *Klassenraum* ('class room') more likely than others. As all children received the same instructions, there should be some overlap between the words they used. Because of this, lexical diversity in here is measured by the amount of different word forms (also referred to as types[2]) a child uses, rather than how diverse these types are semantically. For the analysis it is operationalized as the sum of unique types a child has used.

The Litkey Corpus is provided in form of a data frame consisting of 162,535 rows. Each one of these rows represents one word token written by a specific child during data collection. The columns hold information regarding that specific token.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	word_index	orig	target	spellerror	testpoint	text_id	story	child	school	grade	age	sex	multiling	born_ger
2	1	und	und	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
3	2	und	und	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
4	3	gehen	gehen	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
5	4	Eis	Eis	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
6	5	kaufen	kaufen	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
7	6	belt	bellt	1		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
8	7	den	den	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
9	8	Eismann	Eismann	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
10	9	an	an	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
11	10	der	der	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
12	11	Eismann	Eismann	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
13	12	kukt	kuckt	1		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
14	13	an	an	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
15	14	dan	dann	1		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
16	15	haben	haben	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
17	16	sie	sie	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
18	17	Ihren	ihren	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
19	18	Eis	Eis	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
20	19	ge_kricht	gekriegt	1		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
21	20	und	und	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
22	21	sie	sie	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
23	22	leken	lecken	1		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
24	23	daran	daran	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
25	24	und	und	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
26	25	gehen	gehen	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
27	26	felt	fällt	1		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
28	27	mit	mit	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
29	28	dem	dem	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
30	29	Eis	Eis	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
31	30	hin	hin	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
32	31	und	und	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
33	32	wolte	wollte	1		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA
34	33	den	den	0		101-005-2-III-Eis	Eis		5	4	2	8f	TRUE	NA

Figure 1. Excerpt of the Litkey Corpus (Laarmann-Quante et al., 2019).

In Figure 1 above, we see an excerpt of the data. The *orig* column informs of the actual spelling that was produced by the child in that particular instance. Right next to the actual spelling in the *target* column, we can see the correct spelling the child was (probably) going for. Because

[2]Note that in here I differentiate between tokens, types and lemmas; a token refers to specific word form written by a particular child; a type subsumes all instances of a word form; a lemma refers to the dictionary form of a word form. Example: A particular child might have used the word *Hunde* (which means 'dogs') five times in a text. Each one of these five instances counts as a separate token of *Hunde*, though all five of them are of the same type. Furthermore, the child might have used the singular form *Hund* ('dog') three times. Although *Hunde* and *Hund* count as different types, they belong to the same lemma *Hund*.

these spellings are based on judgments made by the authors of the corpus, taking the context in which the word was produced into consideration, Laarmann-Quante et al. (2019) referred to these spellings as the *target hypotheses*. If both the actual spelling and the target hypothesis align, there is no spelling error, so the column *spellerror* to the right of target is annotated with 0. If there was a spelling error, however, as is the case in the sixth row, where the child instead of producing the correct spelling *bellt* ('barks') produced [<*belt*>][3], the spellerror column is annotated with 1.

As can be seen in the *child* column, each word token in that excerpt has been produced by the same child, namely the child with the identification number 5. Right next to the ID, we also get additional information regarding the child, like the school she went to, the grade she was in during this point of testing or whether she was raised mono- or multilingual. Note that, although all tokens in the excerpt above were produced in second grade and at the same test point – as the *testpoint* column tells us –, data collection actually comprised ten total test points across grades two, three and four of German primary school.

Besides information pertaining to the children, the Litkey Corpus also provides linguistic annotations such as word class membership, word frequency and length, number of morphemes, various orthographic neighborhood measures and more.

Without any further processing, this data already gives way to various insights on different levels; one might, for example, be interested in the words themselves, which ones were used and how often; such questions might then be investigated on the data as whole, but also for individual schools, grades and/or children.

2.1 Creating a subset (of only content words)

Because the aim of this study is to characterize individual differences between children throughout time, I decided to only include data of children that have participated in all ten test

[3]Note that I follow Laarmann-Quante et al. (2019) in that I indicate spelling errors by putting them between squared and angled brackets, e.g., [<*belt*>].

points. In the statistical analysis in chapter 4, I will model the development that a child exhibits at a certain test point as a function of properties characterizing the preceding test point, which makes complete data regarding test points a requirement.

Of the 251 children that have contributed to the overall data, 56 children have participated in all ten test points. Removing the data of children who have not participated in all ten test points, substantially reduces the data to 47,145 tokens over all remaining children and test points.

As mentioned before, I decided that the subset should only include content words, yet exclude function words. To understand what lead to this decision, reconsider the intent of this inquiry, namely to analyze the differences in spelling ability and development between individual children over time. With this in mind, the question then becomes: what portion of the data serves best as a stand-in for such differences? I argue that this is a subset that omits function words.

Note that I define function words in accordance with Laarmann-Quante (2021, p. 242) as words that are annotated with the following tag in the *POS* (part-of-speech) column:

> APPO, APPR, APPRART, APZR, ART, CARD, KOKOM, KON, KOUI, KOUS, PDAT, PDS, PIAT, PIDAT, PIS, PPER, PPOSAT, PRELS, PRF, PTKA, PTKZU, PWAT, PWS, VAFIN, VAINF, VAPP, VMFIN, VMINF

With this distinction in place, an almost even split between content and function words in the data that only includes children that have participated in all ten test points can be observed; 24,343 tokens are labeled as content words (~51.6%), while 22,802 words are labeled as function words (~48.4%). From this, it follows that whatever differences there are between individual children, the contribution to these differences derives (almost) equally from content and function words.

Now consider Trautwein (2019), who found that besides the fact that function words make up but a fraction of a student's (and adult's) total vocabulary, the rate at which children acquire new function words is very low when compared to content words, especially nouns. This is not surprising as function words belong to closed word classes, whereas the amount of nouns a person can learn is theoretically infinite. Furthermore, about 82% of the function words a child knows by fourth grade have already been a part of their lexicon in first grade (calculated from table 7.4 in Trautwein, 2019, p. 70), and there is also less variance between individual children. By excluding function words from the sample then, I remove a portion of the data that contributes very little to the overall variance between children, which should in turn make individual differences in the remaining data all the more pronounced.

One might think that omitting function words could lead to an overestimation of observed differences between children, yet I would argue that regardless of whether function words are included or excluded, what differences do arise from the data are always an underestimation of the real differences between children. This has to do with the fact that the Litkey Corpus is, of course, in itself already a subset of a child's lexical knowledge, and one that draws heavily from function words due to it being comprised of written stories in which function words are needed for their syntactic role. Considering Trautwein (2019), who states that function words only make up about 3% of the lexicon of primary school students, the almost even split between content and function words in the corpus is therefore not reflective of a balanced sample, but of one that is biased against content words.

Another concern regards the orthography of function and content words. In his investigation of German (as well as English) graphematics, Berg (2019) found that many principles that constitute the spelling of function words do not hold for content words and vice versa. Function words tend to be shorter and in many cases their spelling is highly individual and, seemingly, less rule-based.

If one takes error rate as a proxy, the fact that content and function words pose different challenges when it comes to spelling may be examined in the Litkey Corpus. Figure 2 below shows that the mean error rate for content words (on the left) is about 23%, while it is about 5% for function words (on the right). Conducting Pearson's Chi-squared test reveals that spelling errors are not uniformly distributed over function and content words at a significant level (χ^2 = 3229.3, p < 0.001). Figure 2 furthermore demonstrates how the distribution in error rates across children exhibits a wider spread for content words than for function words, with an interquartile range of ~0.156 as opposed to ~0.048.

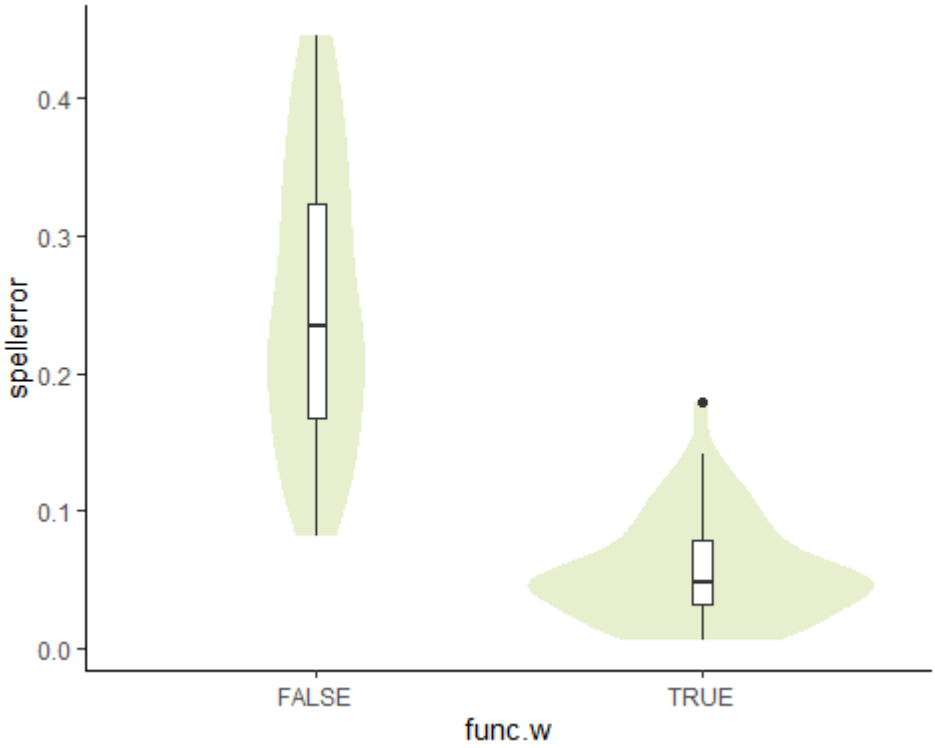


Figure 2. Description of the distribution of error rates across content words on the left of the x-axis and function words on the right.

Looking at this in regards to lexical diversity, Figure 3 presents a similar pattern. Because function words are so ubiquitous in the data and children show comparatively little problems with their spelling, not removing them from the sample would probably lead to less variance in

error rates. To avoid such dilution, I conclude, that in order to best approximate the actual differences between children, creating a sample that disregards function words should be the best course in action.

Further pre-processing of the data involved correcting all-capitalized word tokens so that they match non-capitalized tokens of the same type. I furthermore removed proper nouns as their orthography, similarly to function words, is rather individual and often does not follow more general rules of orthography (Eisenberg, 2020). Besides that, I removed some function words still in the data because of erroneous part of speech annotations. Lastly, I removed data with missing frequency information. This was due to the fact that I intended to include frequency as a controlling variable in the statistical analysis.

After pre-processing was concluded, 21,746 data points remain in the subset of the data.

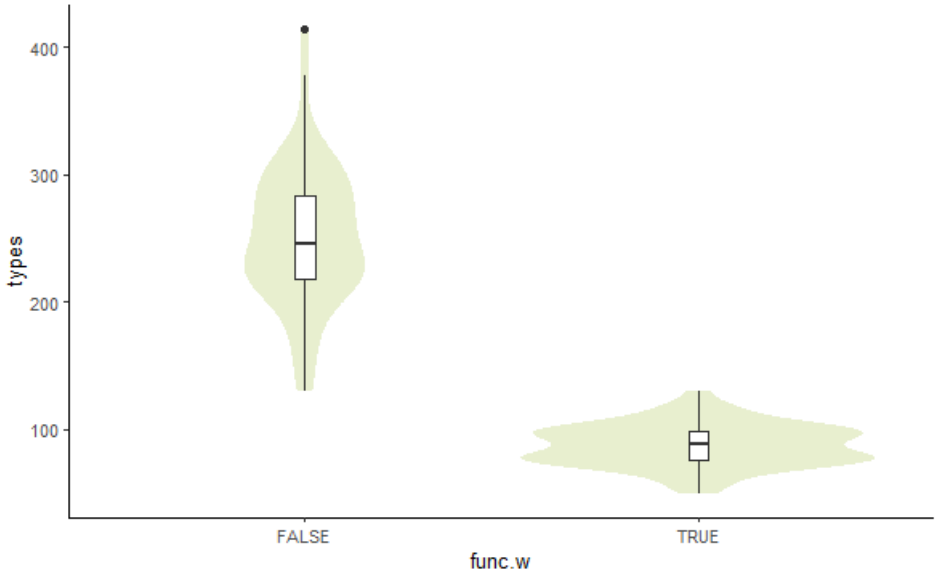


Figure 3. Lexical (type) diversity in the content word and function word subsets of the data.

3 Computation and visualization of the networks

3.1 Setting up the networks

In this chapter I will describe the method to calculate and visualize the networks. Most of the functions used to build the networks are part of the *igraph* R library by (Csardi & Nepusz, 2006). To best showcase the diverging outcome in network visualization depending on child specific properties, for this chapter I handpicked four children from the data, to whom I will refer to by their identification number. Table 1 provides some overt information on these children and their networks:

Table 1. Four exemplary children from the Litkey Corpus considered for the first set of network visualizations.

Child ID	Error rate over all ten test points	Total number of nodes (= individual types) at test point …									
		1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
285	0.09	18	59	77	117	138	162	185	216	247	273
417	0.17	10	27	43	57	76	100	131	158	181	199
064	0.34	15	34	50	74	104	151	188	221	258	287
026	0.47	21	33	44	60	71	90	102	120	131	143

The overall error rates of the children in Table 1 mark the two extremes as well as the first and third quartile boundary within the whole range of error rates in the sample. Concretely, child 285 was the child which exhibited the lowest overall error rate, while child 026 exhibited the highest. I picked these children for demonstrating the visualization, because I assumed that between these two, potential differences should appear the clearest. Children 417 and 064 were added as contextualization and to maybe spot tendencies towards the higher and lower

end of the error rate distribution. From Table 1, for example, we can see that at test point ten child 285 has produced almost twice as many individual types as child 026, even though they started out on similar levels at the first test point. Although this divergence might be a function of spelling proficiency, with child 285, who has higher proficiency, also producing more word types, the examples of children 417 and 064 indicate that this pattern might not be consistent over all children. This will be further investigated in chapters 4 and 5.

To compose a network, one has to first make a decision on the entities that occupy the network. These entities are often referred to as *nodes* (Siew et al., 2019). In here, each node represents a unique word type produced by a child up until a certain test point. In addition to the nodes, one also has to define by what measure these nodes relate to each other. In phonological networks, for example, one such measure is the phonological Levenshtein edit distance (e.g., Vitevitch, 2008; Siew & Vitevitch, 2020). In a network of this kind, all words that differ only by addition, deletion or substitution of a single phoneme are connected to each other. These connections are commonly referred to as *edges* (Siew et al., 2019).

One example for an orthographic network analysis can be found in Trautwein and Schroeder (2018). As in the phonological networks from the studies mentioned above, Trautwein and Schroeder also used an edit distance threshold to set up edges, though in their case this distance was implemented on orthographic instead of phonological representations. The aim of their study was to model lexical development of children throughout time, which is quite similar to what I intend to do in here. One crucial difference, however, is that rather than modeling individual networks from empirical data, Trautwein and Schroeder conducted their investigation on a simulated network that was taken as a stand in for children in general. As a consequence, their network was quite a lot larger (in terms of the total number of nodes involved) than any of the individual networks build in here. For example, the first iteration of the orthographic network in Trautwein and Schroeder (2018) already comprised more than 30,000 word types, while the final network reached about 130,000 individual word types. Compared to the numbers reported in Table 1, the difference in magnitude is rather striking.

The distinction between large, holistic networks and small, individual networks raises one very important methodological concern: while the edit distance threshold of one (addition, deletion or substitution of a letter) works fine for networks as reported in Trautwein and Schroeder (2018), for networks as small as they are in here, implementing the same measure would not yield a lot of connections between word type nodes. In fact, in the sample of the four children picked for demonstrating the network visualization in this chapter, there is not a single instance in which a network at any test point would yield a connectivity exceeding 1% of all possible edges, if the one edit distance threshold would be implemented (see Table 2).

Table 2. Overview of word pairs with an edit distance of one in the exemplary networks. The table also reports the total number of possible edges in the corresponding network (if all nodes were connected to each other).

Child ID	Word pairs with edit distance == 1 // total of possible edges at test point ...									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
285	1 // 153	4 // 1711	7 // 2926	9 // 6786	13 // 9453	27 // 13041	30 // 17020	34 // 23220	39 // 30381	44 // 37128
417	0 // 45	0 // 351	2 // 903	2 // 1596	2 // 2850	3 // 4950	5 // 8515	9 // 12403	9 // 16290	12 // 19701
064	0 // 105	0 // 561	0 // 1225	1 // 2701	3 // 5356	9 // 11325	16 // 17578	19 // 24310	29 // 33153	39 // 41041
026	0 // 210	2 // 528	3 // 946	3 // 1770	3 // 2485	7 // 4005	7 // 5151	11 // 7140	12 // 8515	14 // 10153

Now consider the fact that many analyses of networks elicit information from so-called *largest components* (Siew & Vitevitch, 2019), which comprise the largest connected part of a network. In Vitevitch (2008), for example, this largest component involved about one third of all nodes in the network. In some of the individual networks, on the other hand, if the one edit measure was to be implemented, there would be no components whatsoever, only isolated nodes

(e.g., the test point one networks of children 417, 064 and 026; see Table 2). As a consequence, these networks could not be integrated in the statistical analysis as it would be impossible to calculate some of the typically considered networks metrics.

To account for this problem, in here I make use of a normalized variant of Levenshtein's edit distance derived from Kumar et al. (2022). The idea behind normalization is to use the raw edit distance between words and divide it by the length of the longer word. Doing so will return a value between 0 and 1. This value will furthermore be subtracted from 1, which results in a score that can be interpreted as a measurement of orthographic similarity, with higher scores signifying higher similarity.

Implementing the normalized similarity overcomes another drawback of Levenshtein's edit distance, which is that it does not control for word length. Consider, for example, the distance between *Eis* and *Eismann* (lit. 'ice man', referring to a person, who sells ice cream), which is 4, as well as between *Eis* and *neu* ('new'), which is 3. As the edit distance between *Eismann* and *Eis* is higher than between *Eis* and *neu*, the implication is that the former pair is less similar than the latter, even though *Eismann* fully incorporates *Eis*, while *Eis* and *neu* do not share any letters whatsoever[4]. The reason for this is the short length of both *Eis* and *neu*, as well as the fact that the edit distance between two words can never exceed the length of the longer word in that pair.

After normalization, however, the orthographic similarity between *Eis* and *Eismann* is higher than between *Eis* and *neu*. The calculation for both pairs, in pseudocode, goes as follows:

edit distance(*Eis*, *Eismann*) / word length(longer word(*Eis*, *Eismann*)) = 0.43

edit distance(*Eis*, *neu*) / word length(longer word(*Eis*, *neu*)) = 0

As can be seen, the similarity calculation for the pair *Eis* and *neu* has returned the lowest possible score of 0, which can be interpreted as complete dissimilarity.

[4]Note that capitalized <E> and <e> are taken as different letters, though even if they were not, it would still need three operations to transform *Eis* into *neu* or vice versa.

Compared to the raw edit distance, the normalized score provides more nuance to (dis-)similarities between orthographic forms in that it controls for effects of word length. Additionally, and conveniently, this score brings with it a heuristic to decide which edges should be removed from the network: all pairs of words for which the similarity score does not pass the threshold of 0 will have their connection severed. The effect that the implementation of the similarity score has on the connectivity, which is often referred to as *network density* (Siew et al., 2019), is shown in Table 3.

Table 3. Edges in the exemplary (normalized) orthographic similarity networks. The number in brackets shows the network density, i.e., the relative proportion of existing edges compared to all possible edges.

Child ID	Total number of established edges (and network density) at test point ...									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
285	69 (0.45)	1037 (0.61)	1786 (0.61)	4340 (0.64)	6178 (0.65)	8403 (0.64)	10946 (0.64)	14999 (0.65)	19888 (0.65)	24342 (0.66)
417	20 (0.44)	236 (0.67)	558 (0.62)	1030 (0.65)	1956 (0.69)	3495 (0.71)	6224 (0.73)	8977 (0.72)	11671 (0.72)	14063 (0.71)
064	66 (0.63)	381 (0.68)	848 (0.69)	1799 (0.67)	3729 (0.70)	7750 (0.68)	12371 (0.70)	17188 (0.71)	23225 (0.70)	28472 (0.69)
026	123 (0.59)	310 (0.59)	607 (0.64)	1108 (0.63)	1565 (0.63)	2541 (0.63)	3317 (0.64)	4542 (0.64)	5415 (0.64)	6531 (0.64)

The exact value that characterizes the orthographic similarity between each pair of nodes will be annotated to the corresponding edge in the network object as an edge *weight*. These weights provide a more detailed description of each edge in the network, which will be considered during visualization as well as in the statistical analysis.

With the measure to establish edges implemented, I now turn to the visual attributes of the networks, starting with all exemplary networks for test point one.

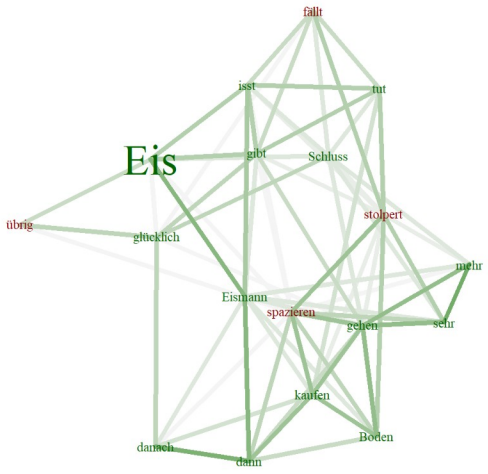
3.2 Visual aspects of the networks

In this section I will highlight some key visual features of the networks. I start with the first test point, depicted in Figure 4. The visual parameters of the networks were chosen in order to quickly convey to the spectator the following information:

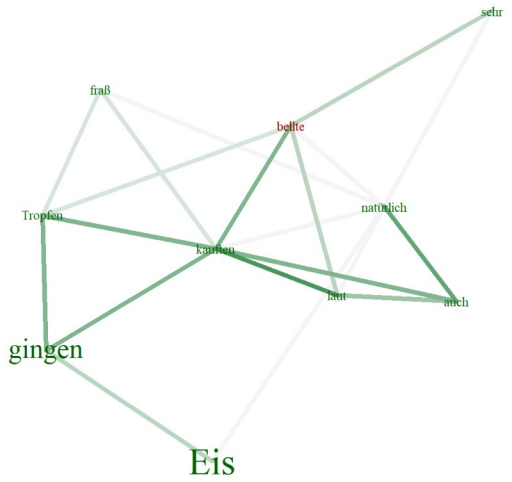
Error rate at a specific test point. The color of the edges were picked from a gradient spanning green (for error rates equal to 0), yellow (error rates equal to 0.5) and red (error rates equal to 1). In the example above, we can see that the networks of children who exhibited lower error rates at test point one, specifically children 285 and 417, appear in a strong green color. The color of the edges in the network of child 064 appear a little muddied, though they still show a hint of green as the error rate is below 0.5. The network of child 026, on the other hand, is rather yellow as the error rate is close to 0.5. Though it is not depicted in these networks, at later test points, we will see that the closer an error rate approximates 1, the stronger becomes the red coloring of the edges.

Edge weight. On top of the gradient spanning green, yellow and red, I also added a gradient of paleness to individual edges, informing of the weight (i.e., the orthographic similarity) between each pair of nodes. Stronger colors indicate higher similarity, while more pale appearing edges indicate rather low similarity. Furthermore, I implemented an ordering of all edges so that high similarity edges would be placed more to the front of the network, making them stand out even more. At later test points, we will see that these networks can become rather cluttered. Making higher weighted edges appear in the front alleviates this cluttering to a degree.

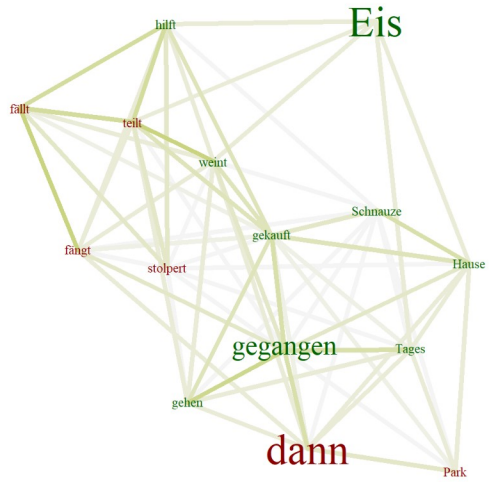
Type frequency of use. When it comes to individual nodes, the relative size of a node's label informs of higher/lower frequency of use up until the current test point. For example, we can see that in all four networks depicted in Figure 4 the children have used the type *Eis* relatively often when compared to other types.



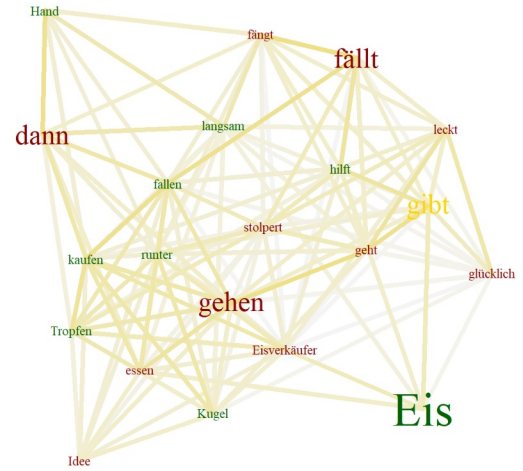
Child: 285
Test point: 1
Error rate: 0.19
Total nodes: 18



Child: 417
Test point: 1
Error rate: 0.08
Total nodes: 10



Child: 064
Test point: 1
Error rate: 0.41
Total nodes: 15



Child: 026
Test point: 1
Error rate: 0.52
Total nodes: 21

Figure 4. Exemplary networks at test point one.

Type specific error rates. Besides type frequency, the nodes also inform of type specific error rates. Each individual type was assigned a color from a gradient similar to that installed for the edges. Looking at the network of child 026, for example, we see that the frequently produced type *Eis* was spelled correctly in every instance, whereas *dann* ('then'), which was also produced relatively often, was never spelled correctly. Additionally, from the yellow coloring of the node *gibt* ('gives'), we see that this word was spelled correctly only half of the time.

In order to compute the layout of the networks, I used the Fruchterman-Reingold algorithm (Fruchterman & Reingold, 1991), which is integrated as a function in the *igraph* library (Csardi & Nepusz, 2006). As described in Fruchterman and Reingold (1991), applying this algorithm makes it so that two competing forces (as well as the number of nodes and the space that is available) decide over the graphs layout. The so-called "repulsive forces" (p. 1132) cause all nodes to be pushed apart, while the "attractive forces" (p. 1132) weakens the repulsive effect between all nodes that are connected via an edge. In other words: while all nodes are being pushed apart, connected nodes are not being pushed as far apart as unconnected nodes. As a consequence, nodes that share a connection should be placed closer to each other in the final layout than unconnected nodes. Furthermore, as the documentation of this algorithm in R says, if the network object contains weights for the edges – which is the case in here –, the attraction between nodes is multiplied by the weight of their connection. So nodes with high weight edges connecting them should be placed even closer to each other. In the next section, I will assess the algorithm in more detail.

3.3. Assessing the Fruchterman-Reingold algorithm

Notice how in every graph in Figure 4 there are some nodes that are placed quite far outside the center, most notably *übrig* ('left' as in 'nothing left') in the network of child 285, as well as *sehr* ('very'), *fraß* ('ate', most commonly with an animal agent) and *Eis* in the network of child 417. What these nodes have in common is that (1) they only connect to few other nodes and (2) their connections are relatively low in weight. As described above, when applying the

Fruchterman-Reingold algorithm the force that pushes all nodes away from each other is weakened between nodes that are connected to each other and even more so when their connection has a higher weight assigned to it. This explains why in the networks of Figure 4 exactly these nodes are pushed farthest from the center, because neither do they connect to a lot of other nodes that would draw them closer nor are their connections very high in weight.

Now consider the node *natürlich* ('naturally') in the network of child 417. It connects to seven other nodes, though six out of these are rather low in weight, as indicated by how pale they appear in the network. The fact that *natürlich* was placed rather central in the graph's layout, even though most of the edges attached to it are very low in weight, suggests that the amount of connections a node has is more important for its placement than the weight of the edges itself. This has implications for the networks of subsequent test points as well, because due to how orthographic similarity was implemented, pairs that consist of at least one relatively long word are more likely to pass the similarity threshold of 0 to establish an edge. As a consequence, longer words added to the graphs will probably establish a higher amount of edges, and even more so when two relatively long words are involved, which will draw longer words more towards the center. In fact, there are hints of this even in the graphs of the first test point. Though there are exceptions, many nodes that mark the corner points of the graphs are words consisting of three or four characters, while in the center, the graphs are predominantly occupied by words with five or more characters.

The implications of this reach even farther than sheer visualization. If longer words actually do connect with more other nodes, they should also yield a higher *degree*, which is a commonly employed statistic in network science, informing of how many connections a node has. Degree can be calculated individually for each node, but also as an average for the whole network. Additionally, *network density*, which I briefly introduced in section 3.1, is calculated by dividing the number of all existing edges in a graph by the number of possible edges. If longer words are more likely to establish an edge, we would expect that those networks that mostly consist of longer words are also more dense and yield higher average degree. This should be kept in mind

and definitely checked for when conducting a statistical analysis later on. For now, however, I turn the qualitative part of the analysis, inspecting and describing the exemplary networks spanning all ten test point.

4 Visual analysis

In this chapter, I will describe and compare the visual appearance of several graphs across all ten test points, especially focusing on those aspects that might point to differences in spelling development. Doing so might help formulate hypotheses for the subsequent statistical analysis of the networks conducted in chapter 5.

By themselves, the networks of test point one do not provide any information regarding the lexical development of individual children. For that, subsequent networks have to be considered. Still, this first set of networks might nevertheless hint at individual differences between children. To give an example, child 417 only produced ten different (content word) types, exhibiting the lowest overall error rate, while 026 produced the most types with the highest error rate out of these four. The notion that error rate might correlate with types produced, could be something to look out for in later graphs. My main interest, of course, is to investigate whether such observations can be made systematically across test points.

Before continuing with the second test point, however, I would like to draw attention to one other example as to demonstrate another potential use case of network visualizations. On the level of individual words, we see that every child has used the word *Eis* in their story. Furthermore, *Eis* seems to be the most frequent word in every story as indicated by the size of its label. Another type that appears in multiple networks is *fällt* (‘falls’ as in ‘the ice cream cone falls to the ground’). This as well as other types that appear across multiple networks are not a coincidence, because remember that the children were given picture stories as a basis for their writing.

As the visual networks show, all children who had used *fällt* in their stories have misspelled it in all instances. Interestingly, the only type misspelled by child 417 was *bellte* (past tense of 'barks'), which – just as *fällt* – contains a doubled consonant. This could point to shared difficulties across children, so a closer look might prove fruitful. For this, consider Table 4.

Exactly half of the errors in Table 4 are due to a misspelled doubled consonant; three errors are due to a false capitalization, and the remainder involves the umlaut <*ä*>, which erroneously has been spelled out as <*e*>.

Table 4. A selection of error tokens taken from the networks of test point one (Figure 4). Errors in bold pertain to consonant doubling.

Child ID	Original spelling	Target hypothesis
285	**fält**	fällt
417	**belte**	bellte
	Bellte	bellte
	Bellte	bellte
	Fällt	fällt
064	**felt**	fällt
026	fellt	fällt
	fellt	fällt
	fält	fällt
	fält	fällt

Now it is, of course, not big enough of a sample to draw conclusions on any of these observations. Furthermore, this study is not about specific graphemic patterns like the doubled consonant (see, for example, Laarmann-Quante, 2021, for an extensive investigation of that matter). However, spotting patterns to follow up with a little more digging that could eventually

lead to hypotheses to test on a larger scale is – as I would argue – exactly what these network visualizations can be in service for.

As in here I am more interested in how the children develop over time, next I will turn to the networks of the second test point and compare them to their predecessors and also across children.

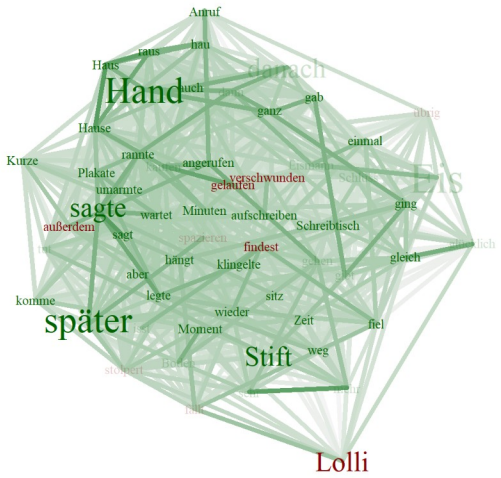
Considering Figure 5, one thing that immediately catches the eye is how many words have been added to the network of child 285 when compared to the others. With 59 total nodes, it is now the largest network out of these four, while its error rate is the lowest at the second test point.

The other children exhibited similar growth (between 12 and 19 new nodes), leading to networks that are rather comparable in size. When it comes to spelling proficiency, child 026 still shows the highest error rate out of these four children, while 417 and 064 have moved closer together in this regard.

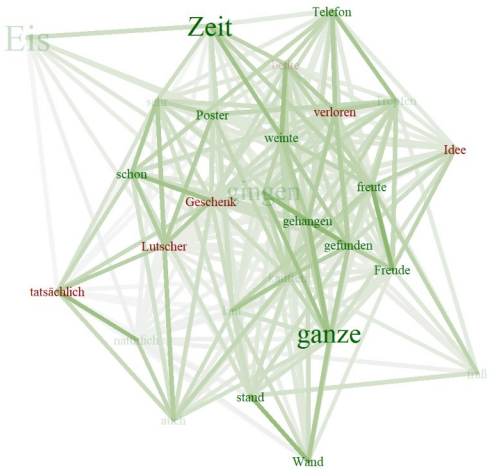
Structurally, most layouts do not seem to exhibit clear patterns between how correct and erroneous nodes are distributed. In the graph of child 285, *Lolli* ('lollipop') and *übrig* stand out the most, with both being errors, but there are also some errors that were placed rather central and amidst correct spellings. This is likely due to their high length, as no new errors placed in the graphs center are shorter than seven characters.

In the graphs of 417 and 064, correct spellings and errors seem to mix. Only the graph of child 026 seems to show somewhat of a divide between correct spellings and errors, which is even more pronounced when taking words from earlier test points into account. This might be something to look out for in subsequent networks.

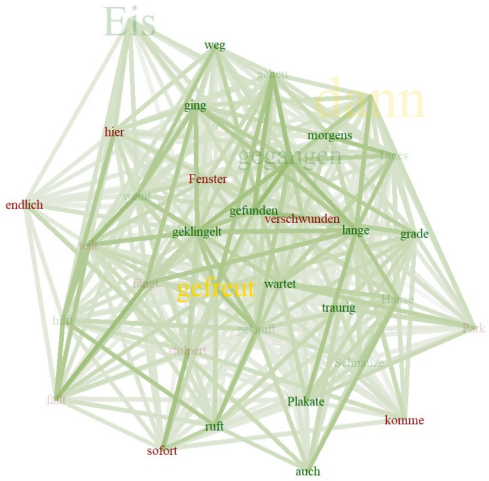
Taking test point three into consideration (Figure 6), we see comparable growth between all four children (between 11 to 18 new nodes). Children 417, 064 and 026 also show growth that is on a similar level with test point two. Child 285, whose network has increased drastically at the previous test point, is now more on par with the others.



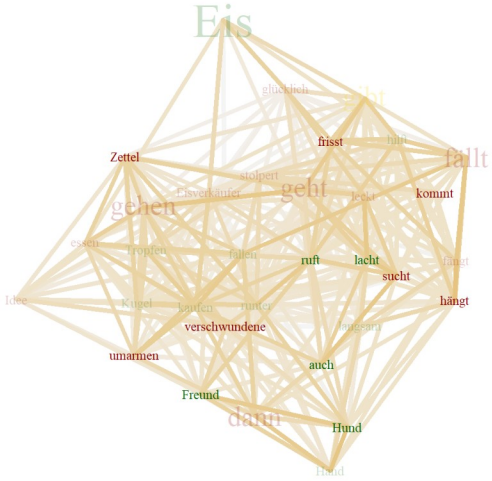
Child: 285
Test point: 2
Error rate: 0.12
Total nodes: 59
New nodes: 41



Child: 417
Test point: 2
Error rate: 0.26
Total nodes: 27
New nodes: 17

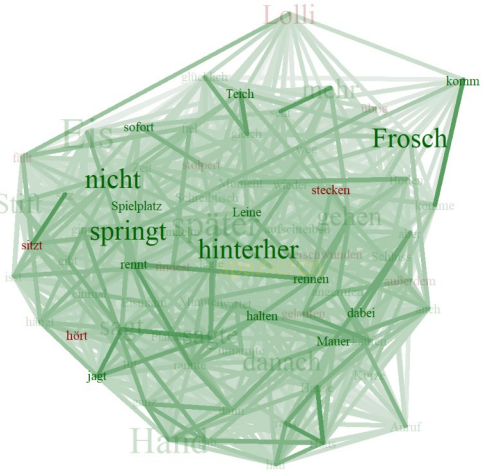


Child: 064
Test point: 2
Error rate: 0.29
Total nodes: 34
New nodes: 19

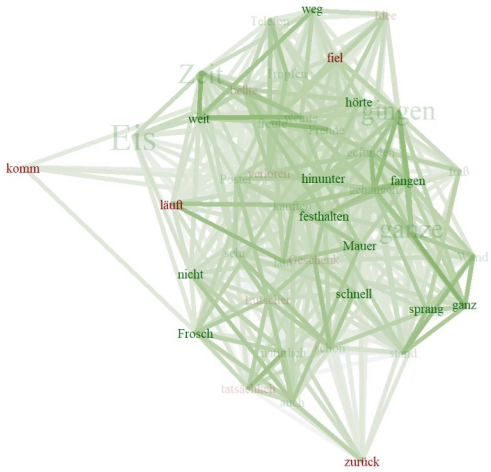


Child: 026
Test point: 2
Error rate: 0.62
Total nodes: 33
New nodes: 12

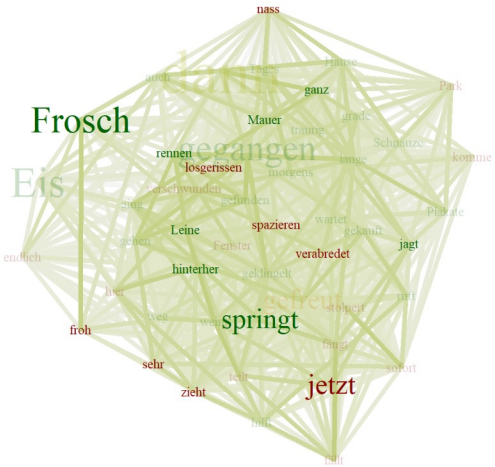
Figure 5. Exemplary networks at test point two.



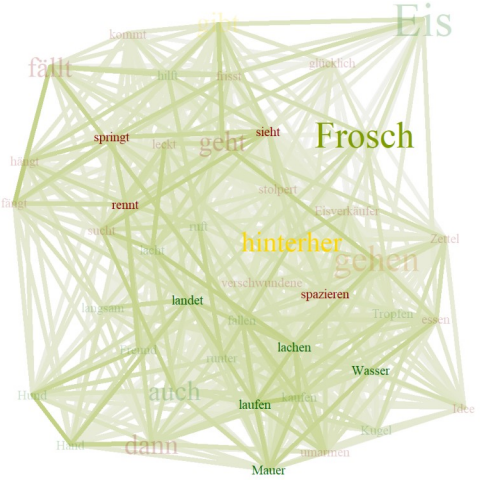
Child: 285
Test point: 3
Error rate: 0.12
Total nodes: 77
New nodes: 18



Child: 417
Test point: 3
Error rate: 0.25
Total nodes: 43
New nodes: 16

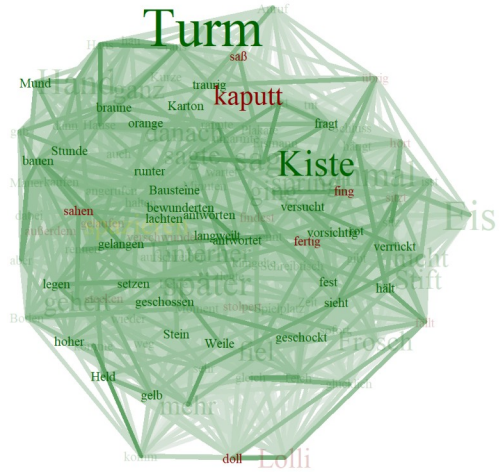


Child: 064
Test point: 3
Error rate: 0.39
Total nodes: 50
New nodes: 16

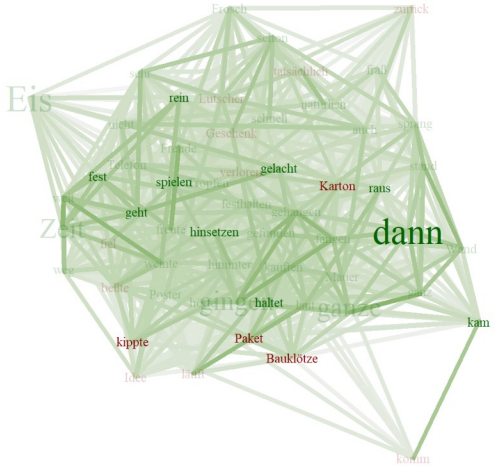


Child: 026
Test point: 3
Error rate: 0.38
Total nodes: 44
New nodes: 11

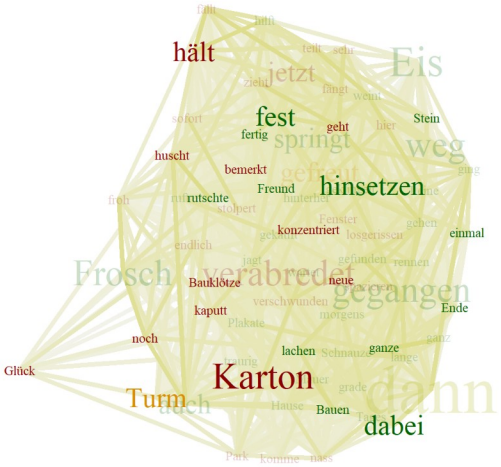
Figure 6. Exemplary networks at test point three.



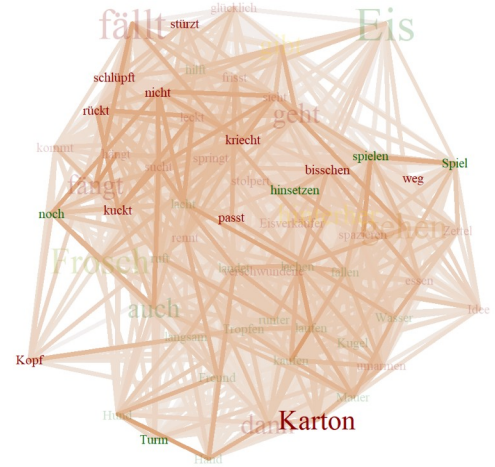
Child: 285
Test point: 4
Error rate: 0.13
Total nodes: 117
New nodes: 40



Child: 417
Test point: 4
Error rate: 0.25
Total nodes: 57
New nodes: 14



Child: 064
Test point: 4
Error rate: 0.46
Total nodes: 74
New nodes: 24



Child: 026
Test point: 4
Error rate: 0.75
Total nodes: 60
New nodes: 16

Figure 7. Exemplary networks at test point four.

Child 285 shows the lowest error rate with about 12% errors, which is the same as before. 417 also shows consistency in error rate, while 064 has produced errors at a higher and 026 at a lower rate than before, with both of them sitting just under 40% erroneous spellings.

The placing of nodes in the graph of child 026 again slightly hints at a division between correct spellings and errors, more so when words from earlier test points are taken into consideration. New errors seem to be placed a little closer together and on top of errors from earlier test points, while the same seems to be the case for correct spellings. Of course, as was pointed out above, the placing of the nodes should be affected by word length, though in this particular layout we can see that apart from *spazieren* ('to take a walk'), newly added correct as well as erroneous spellings are quite similar in their number of characters. Under this premise, and considering that the other factor that plays into the computation of the layouts is weight (i.e., orthographic similarity), the separation of correct spellings and errors in here might indicate that child 026 has difficulties in specific orthographic environments.

The other networks do not show clear patterns. In those of child 417 and 064 we can see that some errors are pushed quite far outside the network, though others are placed more towards the center.

Figure 7 depicts the networks of test point four. As was the case between the first and second test points, between the third and forth, child 285 once again shows a rather large increase in total nodes when compared to the other children. In terms of error rate, child 285 is still very consistent, producing about 13% errors at this test point. Many new correct spellings have been placed towards the center of the network, especially longer words. Strikingly, this child at this test point showed more difficulties with the spelling of shorter words than with longer words.

Child 417 has been rather consistent as well, with about 25% errors for the last three test points. When it comes to total nodes, 417 actually has the smallest network with the smallest growth at test point four.

In contrast, 064 and 026 exhibit more fluctuation in error rates across test points. Both have produced a higher rate of errors compared to the preceding test point, with child 026 producing about 75% errors.

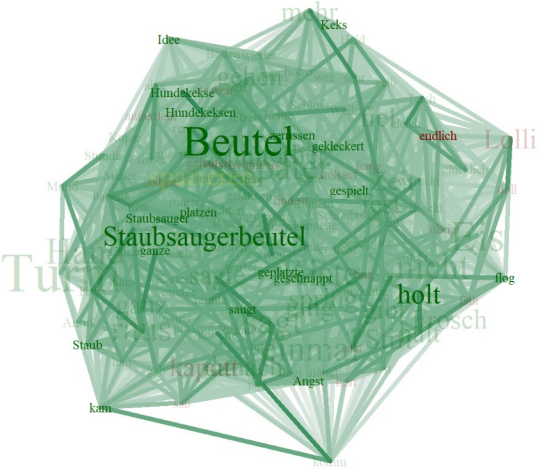
In the network of child 026, most new errors are placed close to each other and on top of a cluster of previous errors. The network of child 064 might now also show a slight division between correct spellings and errors, though not as clearly.

Moving on to test point five in Figure 8, child 064 shows the largest growth, while the network of child 026 has grown the least.

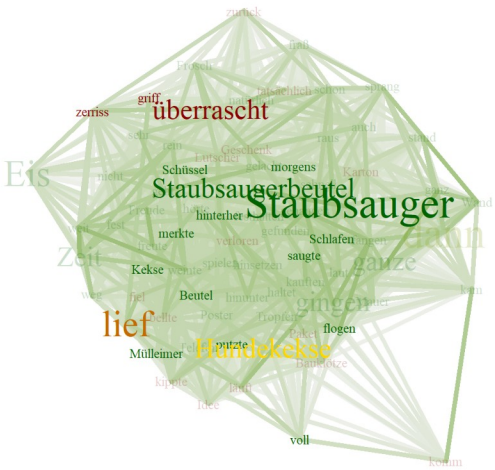
In terms of error rate, child 285 continues to make the least amount of errors. Only one new error was added, namely *endlich* (‘finally’). In all networks that add more than one new error, correct spellings as well as errors do not appear to build any exclusive communities.

There are no clearly observable structural patterns in the networks aside from longer words being drawn more towards the center.

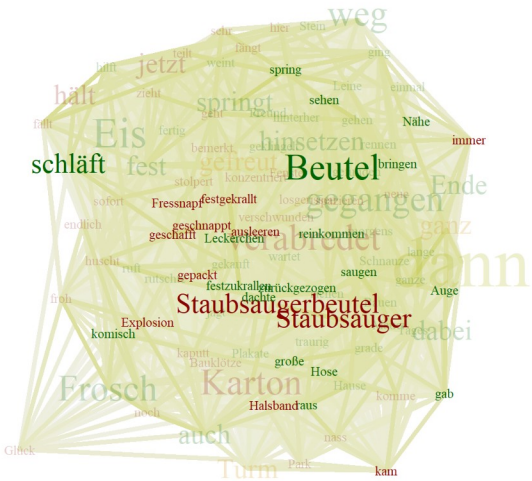
Child 064 once again shows high growth at test point six (Figure 9). Their network now almost has as many nodes as that of child 285. The networks of 417 and 026, on the other hand, are about one third smaller than those of the other two. Interestingly, the growth and consequently size of the networks does not seem to be affected by error rate, at least not when solely looking at these children. While in a sample that involves all 56 children, this outcome might be different, up to this point, we can put on record that on an individual level there are children that exhibit higher growth paired with lower error rates (child 285), higher growth paired with higher error rates (child 064), lower growth with lower error rates (child 417), as well as lower growth with higher error rates (child 026). Maybe this points to different writing strategies that the children (consciously or unconsciously) adhere to; it could, for example, be the case that child 417 only uses words of which they are relatively sure how to spell, while child 064 does not have much regard for the correct spelling.



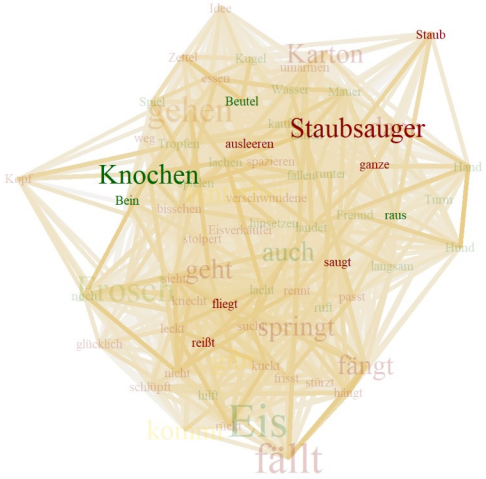
Child: 285
Test point: 5
Error rate: 0.03
Total nodes: 138
New nodes: 21



Child: 417
Test point: 5
Error rate: 0.29
Total nodes: 76
New nodes: 19

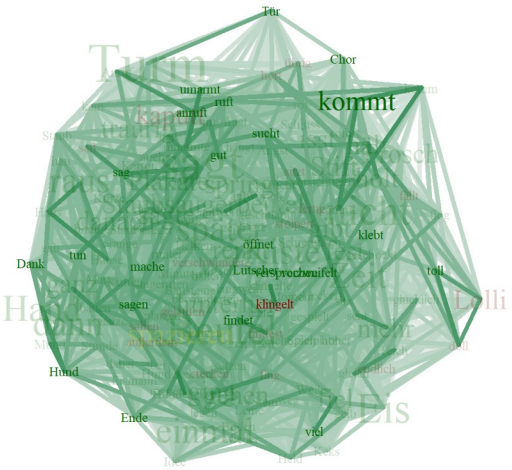


Child: 064
Test point: 5
Error rate: 0.44
Total nodes: 104
New nodes: 30

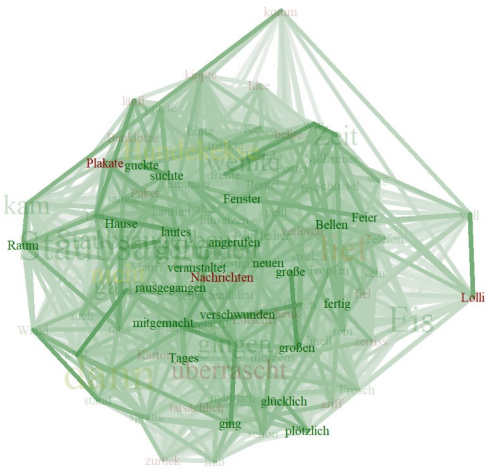


Child: 026
Test point: 5
Error rate: 0.6
Total nodes: 71
New nodes: 11

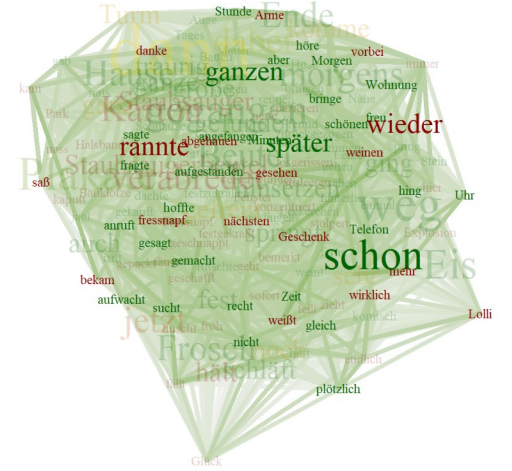
Figure 8. Exemplary networks at test point five.



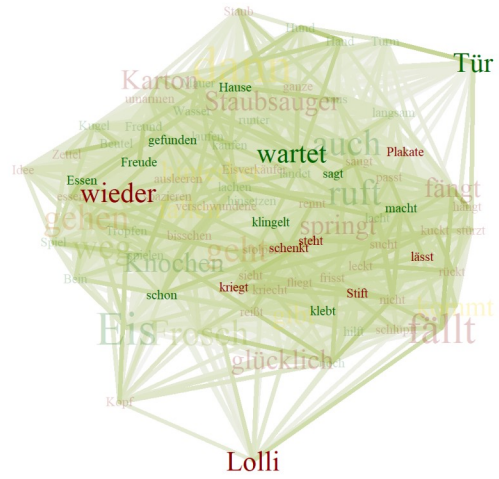
Child: 285
Test point: 6
Error rate: 0.03
Total nodes: 162
New nodes: 24



Child: 417
Test point: 6
Error rate: 0.14
Total nodes: 100
New nodes: 24

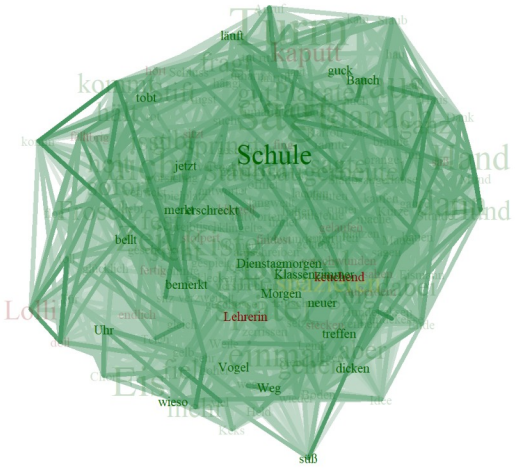


Child: 064
Test point: 6
Error rate: 0.28
Total nodes: 151
New nodes: 47

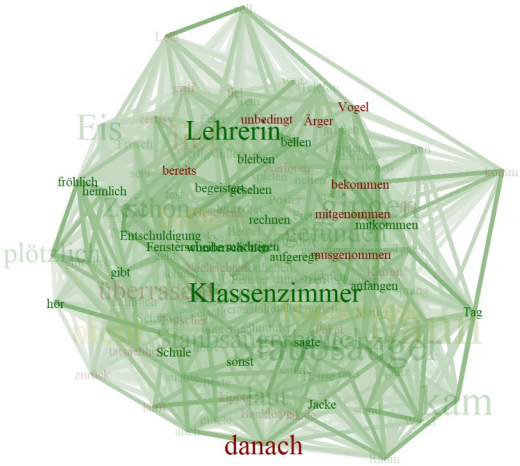


Child: 026
Test point: 6
Error rate: 0.37
Total nodes: 90
New nodes: 19

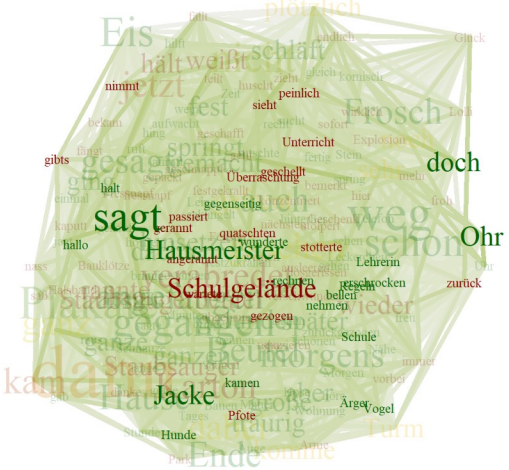
Figure 9. Exemplary networks at test point six.



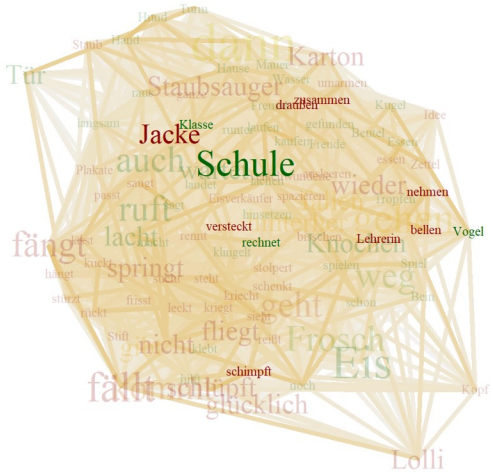
Child: 285
Test point: 7
Error rate: 0.05
Total nodes: 185
New nodes: 23



Child: 417
Test point: 7
Error rate: 0.2
Total nodes: 131
New nodes: 31



Child: 064
Test point: 7
Error rate: 0.36
Total nodes: 188
New nodes: 37



Child: 026
Test point: 7
Error rate: 0.61
Total nodes: 102
New nodes: 12

Figure 10. Exemplary networks at test point seven.

At test point seven depicted in Figure 10, child 417 has shown their highest growth yet, with 31 newly added nodes. In contrast, child 026 exhibits the lowest growth with 12 new nodes. Their network is now about 22% smaller than the second smallest network.

At test point eight (Figure 11), child 026 once again shows the lowest amount of newly added nodes, widening the gap to the other children. In terms of error rate, 026 continues to show high fluctuation across test points, this time yielding about 36% errors, which is a substantial improvement of about 25% when compared to the previous test point, yet about the same when compared to test point six (37% errors).

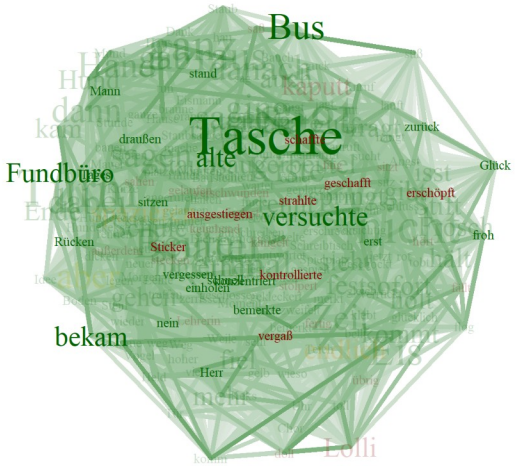
Children 285, 417 and 064 are almost on par when it comes to newly added nodes, though they differ in that 064 still adds a relatively high amount of new errors.

The second to last test point (Figure 12) again shows relatively high growth for children 285 and 064. The network of child 026 has once more grown the least, while the growth of 417 sits somewhat in between both extremes.

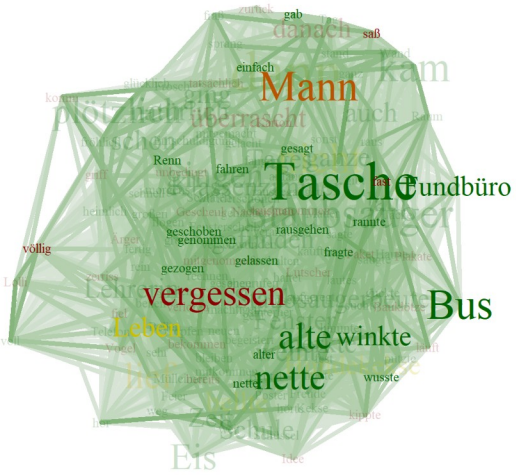
Finally, at test point ten in Figure 13, we see that many trends from former test points have continued. Child 285 and 064 once again show higher growth than the others, with 285 adding mostly correct spellings, while 064 has added a more even mix of correct spellings and errors.

Child 026 has added the least amount of new nodes to their network, which puts it at a total size of 143 nodes, which is about half of the networks of children 285 and 064, and about 28% smaller than the second smallest network in this sample.

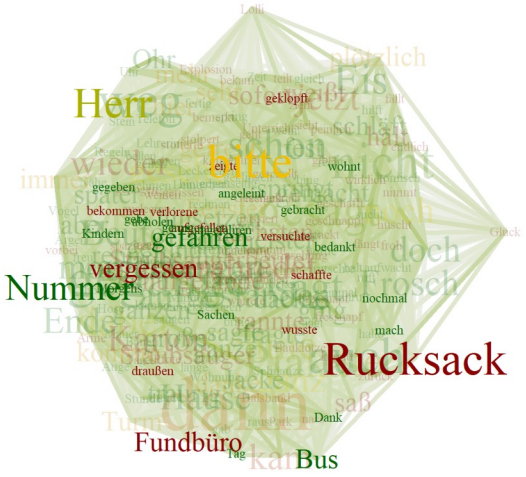
Previously I pointed out that, by only looking at these four children, error rate does not seem to predict network growth. While arriving at the last test point has not necessarily changed this notion, it certainly is worth mentioning that throughout all ten test points, the child with the highest error rate has very consistently added the least amount of new nodes to their network. Investigating this on a broader sample might clarify, if there is a systematic component in there or not.



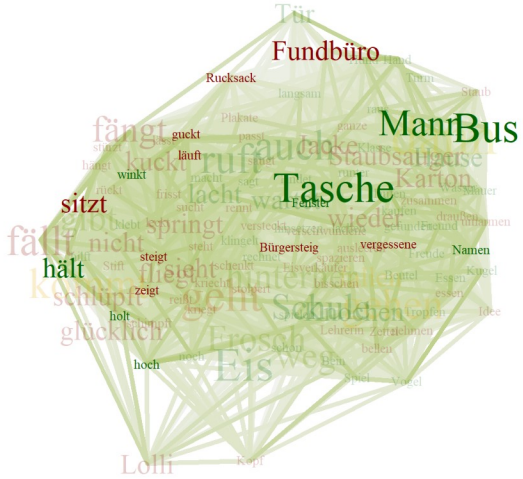
Child: 285
Test point: 8
Error rate: 0.15
Total nodes: 216
New nodes: 31



Child: 417
Test point: 8
Error rate: 0.19
Total nodes: 158
New nodes: 27

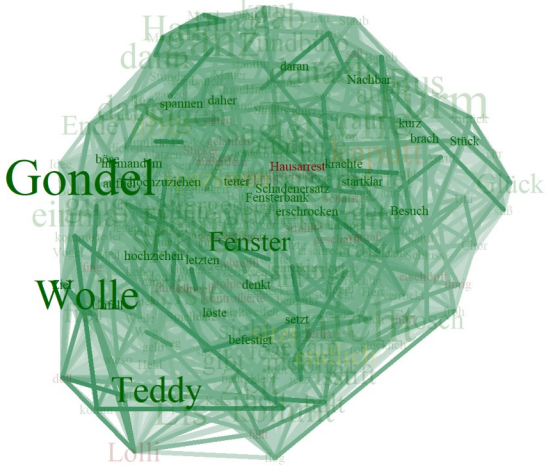


Child: 064
Test point: 8
Error rate: 0.35
Total nodes: 221
New nodes: 33

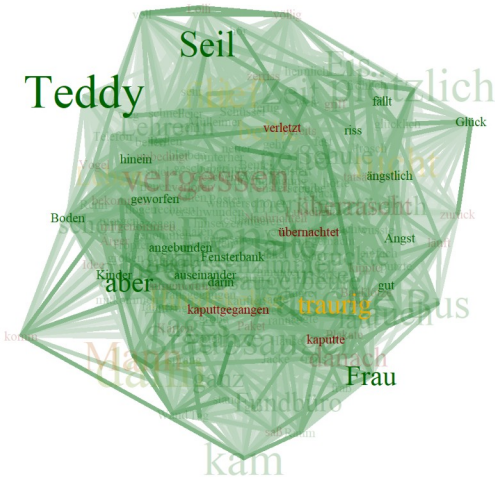


Child: 026
Test point: 8
Error rate: 0.36
Total nodes: 120
New nodes: 18

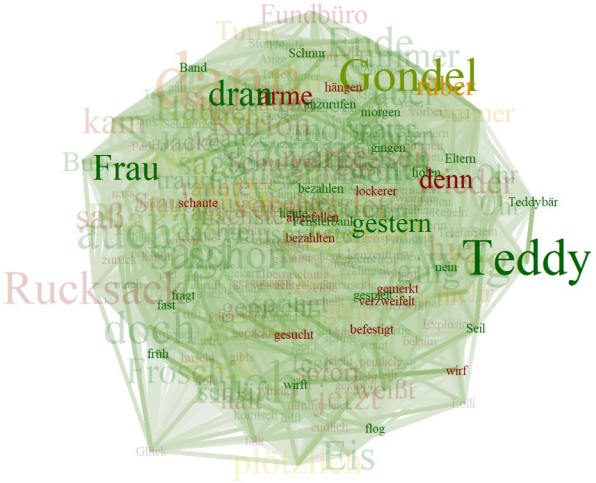
Figure 11. Exemplary networks at test point eight.



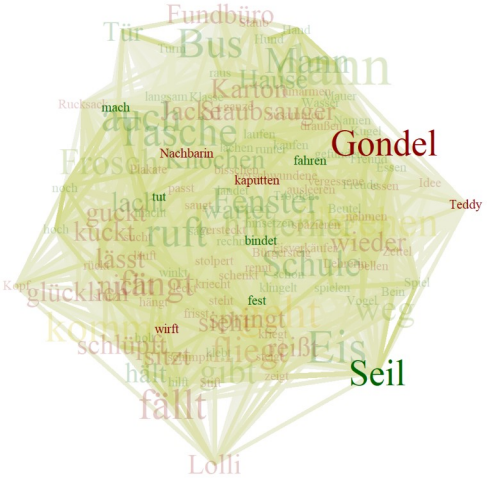
Child: 285
Test point: 9
Error rate: 0.02
Total nodes: 247
New nodes: 31



Child: 417
Test point: 9
Error rate: 0.12
Total nodes: 181
New nodes: 23

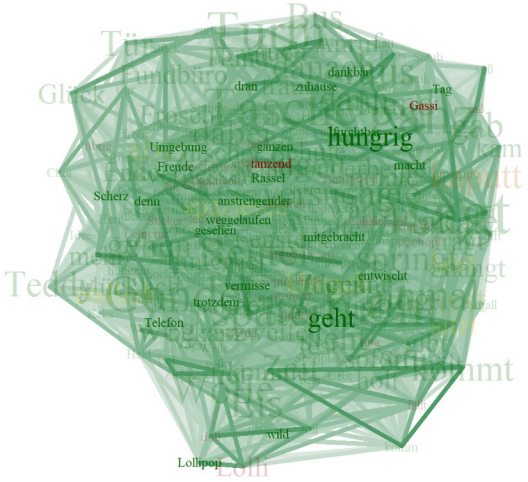


Child: 064
Test point: 9
Error rate: 0.27
Total nodes: 258
New nodes: 37



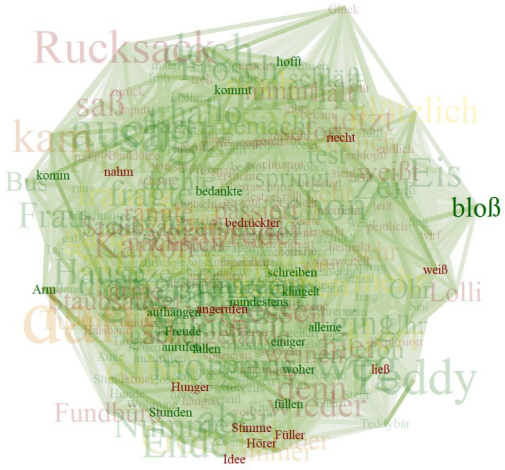
Child: 026
Test point: 9
Error rate: 0.43
Total nodes: 131
New nodes: 11

Figure 12. Exemplary networks at test point nine.

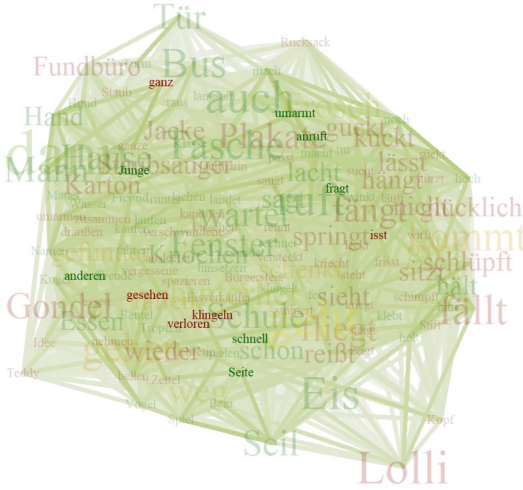


Child: 285
Test point: 10
Error rate: 0.06
Total nodes: 273
New nodes: 26

Child: 417
Test point: 10
Error rate: 0.06
Total nodes: 199
New nodes: 18



Child: 064
Test point: 10
Error rate: 0.3
Total nodes: 287
New nodes: 29



Child: 026
Test point: 10
Error rate: 0.36
Total nodes: 143
New nodes: 12

Figure 13. Exemplary networks at test point ten.

5 Statistical analysis

The visual inspection of the networks in chapter 4 has shown that there are differences in growth among children. In this section then, making use of mixed effect models, the aim is to investigate whether there are any variables that can predict the differences in growth that exist across children.

5.1 Variables and hypotheses

The statistical analysis will be conducted on aggregated data including all ten test points of the 56 children in the content word sample. The number of new words that have been added to a network at a given test point is taken as a proxy for development and therefore serves as the dependent variable. For brevity, I will refer to this variable as *growth*.

For calculation, I chose a linear mixed model with random intercepts for each one of the 56 children in the data. The full model specification, in pseudocode, reads as follows:

> growth ~ error rate + test point + average frequency + text length (scaled) + network size (scaled) + density + average weight + average path length + network size (scaled):test point + (1|child)

It is important to clarify that almost all independent variables refer to the test point prior to the test point to which the dependent variable growth corresponds to (the only exception is *text length*; more on that shortly). The reason for letting the statistics of prior test point networks predict growth at the next test point is, of course, that I aim to investigate the development of children throughout time.

As to text length, because this variable is highly correlated with the dependent variable growth ($r = 0.8$, $p < 0.001$), including it as a predictor poses the possibility of overfitting the data. It is included anyway, because it should serve as a measure of motivation and/or concentration of a given child at the day of testing. It is entirely possible, that the children are differently

motivated by the content of a picture story or that they are not as productive, because they had a bad day. Text length is therefore included to control for this kind of variability.

Besides text length, *average frequency* has also been added as controlling variable. *Test point* was added mainly to test for an interaction with *network size*, which refers to the total number of nodes (at the previous test point). If there is some kind of ceiling effect to the development exhibited in this sample, we could see that larger networks do not grow as fast at later test points than they did at earlier test points.

I include error rate as an independent variable to look for effects of proficiency. Though the visual analysis in chapter 4 hinted at inconsistencies in this regard, I hypothesize that more proficient spellers should exhibit higher growth overall.

Density, *average weight* and *average path length* all pertain to the connectivity inside the networks and are directly computed from the corresponding network objects. I will provide details on these variables shortly, but in order to do so, we first have to reconsider the orthographic similarity metric to constitute edges.

Inspecting the Fruchterman-Reingold layout in section 3.3 has hinted at the possibility that longer words could be at an advantage when connecting to other words. Indeed, testing for a correlation between average length of all words in a network and the corresponding network's density, yields a significant positive correlation ($r = 0.45$; $p < 0.001$). This means that networks whose words are longer on average are also more connective.

This find is of immense importance for the interpretation of most characteristics extracted from the networks, not just density. *Average degree*, for example, measures how many connections every node in a network has on average. This metric already highly depends on the size of the network, because a network that comprises of more nodes in total can also build more connections among these nodes. With the orthographic similarity metric, another influence gets thrown into the mix, because even when it is the case that two networks are of the same size, the network that contains more long words should yield higher average degree. So

average degree in here is very likely confounded by the network's size as well as the average word length inside that network. Calculating Pearson's correlation coefficient between average degree and network size reveals a very strong correlation ($r = 0.998$, $p < 0.001$), as was expected. The correlation with average word length is also significant, though not as strong ($r = 0.26$, $p < 0.001$).

Now consider that network density can be calculated directly from average degree. To do so, one must simply divide average degree by n-1 (with n being the total amount of nodes in a network). This suggests that in a model that controls for network size (by including it as an independent variable), the inclusion of average degree would be redundant if density was included as well. Of course, network size could be excluded instead of average degree, yet I would argue that the size of the network is a more "pure" metric, which makes interpretation easier. Average degree will therefore not be included in the model.

Density is positively correlated with average word length, as was shown further above. If the effect of word length was controlled for, what information does a variable like density provide, and how could network growth be affected by this?

To get a grasp of this, imagine two orthographic networks that consist of the same amount of words, and on average all these words have the same length. If one of these networks showed higher density than the other, this would mean that the words in it have established a higher proportion of edges among each other. Considering the orthographic similarity metric, a find like this would translate to more pairs of nodes passing the similarity threshold of 0 than in the other network, so more pairs are at least a little bit similar. The last part is important, because density does not factor in weight, so it does not directly inform of how similar the words in the network are, as supported by a weak correlation between density and average weight ($r = -0.11$; $p < 0.05$).

Although weak, the negative coefficient is yet another pointer towards the fact that longer words are more likely to establish edges with other words. The relationship is as follows: denser

networks usually contain more long words, which are more likely to connect to other words, though these connections are often rather low in weight.

Back to the question what density might be able to tell us (if average word length was controlled for): in the example of the two same sized, same average word length networks, the one with higher density might point to more words connecting to each other not so much because of chance, but because the child has produced a larger proportion of words that share specific letters or orthographic patterns. If this was the case, it could be interpreted in a way that the child does not diverge from familiar orthography that much, restricting their own writing. Regarding growth between test points, the hypothesis could then be that higher density inhibits growth, because the child would be hesitant to add words to the network that are too different from those already inside the network.

However, the claim that density indicates a more restricted usage of orthography does not stand on the most solid grounds, because even if word length was controlled for, there probably is a lot of chance involved still, simply due to the mass of words (and word forms) that exist in German. So effects of density, if there are any, have to be viewed with caution.

Another metric commonly looked at when analyzing networks is average path length, which measures how many edges the nodes in a network have to traverse to reach every other node. To give an example, in a network in which all nodes are connected to each other, the average path length equals 1. Path length is often investigated in light of a theory called *spreading activation*, which assumes that the activation of a node co-activates other nodes, first and foremost those closest to it (Siew, 2019). In turn, nodes that are farther away from the initially activated node do not get activated as much, if at all.

In networks that have weights assigned to edges, the path length between two nodes is calculated as the sum of the weights of all edges traversed. Furthermore, if the target node can be reached via multiple paths that have the same absolute length (i.e., the same number of

edges traversed), the path with the lowest sum of weights is taken as the path length. So average path length is always based on the shortest paths between nodes.

Considering that edge weight in here is based on orthographic similarity, and higher values indicate higher similarity, plus the fact that the R function to calculate this measure chooses the path with the lowest sum of weights, finding an interpretation of this metric is not straightforward. First of all, because path length is a sum of weights, the path length between two connected nodes *A* and *B* will be smaller than between nodes *A* and *C*, which are only indirectly connected via node *B*. Consequently, path length cannot be interpreted in a way that higher values indicate higher similarity. Yet the fact that the lowest sum of weights is chosen when several paths are available also prohibits the interpretation that low path lengths signify high similarity, because in some instances, the sum of weights between indirectly connected nodes can be lower than the weight of an edge between directly connected nodes. To give an example, let us once again consider the network corresponding to the first test point of child 026 in Figure 4.

In that network, the nodes *kaufen* (‘to buy’) and *glücklich* (‘happy’) are only indirectly connected via the node *langsam* (‘slow’). Calculating the path length and taking weight into consideration yields a value of ~0.25. Now consider that the weight between directly connected nodes *kaufen* and *Eisverkäufer* (‘ice-cream seller’) is 0.33.

So the path length value between *kaufen* and *glücklich* is lower than the similarity between *kaufen* and *Eisverkäufer*, which shows that lower values cannot be taken as indicating higher similarity. To add to the confusion, the shortest path length between *kaufen* and *Eisverkäufer* that is calculated by the function in R is not even the direct path, but an indirect path that goes through the node *stolpert* (inflected form of ‘to stumble’).

All of this makes the average shortest path length as calculated in R way to convoluted to try to make inferences. However, there is a way to modify the calculation in order to make interpretation easier. Because weights range between 0 and 1, calculating their product instead

of their sum penalizes indirect paths by returning values that get increasingly smaller the more edges are involved. In this way, path length values can be interpreted as measures of similarity even between unconnected nodes, with higher values indicating higher similarity. To achieve this, first all weights in a network object have to be transformed to their natural logarithm multiplied with -1. This yields higher values for initially lower weights. After that, taking the sum of these values and transforming it back by usage of the exponential function, the product of the initial values is returned.

The fact that the log transformed values are higher for small initial values also makes it so that in most cases direct paths should be chosen over indirect paths, although this is not guaranteed. Theoretically, very small edge weights could return high enough logarithms so that the sum of an indirect path might be smaller, provided that the initial weights of the involved edges are high enough. There are a few examples of this in the data, for instance in the first network of child 136: In that network, the weight between *suchen* ('to seek') and *leckt* (inflected form of 'to lick') is 0.167, yet the shortest path (after log transforming) between these nodes makes a detour, traversing the node *guckt* (inflected form of 'to watch').

In this case the indirect connection between *suchen* and *leckt* via the node *guckt* would be chosen as the shortest path over the direct connection. While this certainly does not make interpretation easier, one could argue that the indirect path here is a more apt representative of the orthographic relation between *suchen* and *leckt*, because the direct connection was established only due to the shared *<c>* in both words, which might just as well be attributed to chance. In the indirect path, however, there is one bigram (<*uc*> in suchen|guckt) as well as one trigram (<*ckt*> in guckt|leckt) involved. So those instances where an indirect path is chosen over a direct path might be taken as a correction for the very liberal method to establish edges. Granted, as was the case for density, this conclusion assumes a lot and should be regarded with caution.

For now, average path length will be taken as a proxy for the similarity structure in a network. Higher values indicate that the words inside a network are more similar orthographically. Besides average path length, the linear model will also include the average weight of all edges inside a network. Average weight differs from average path length in that it only considers direct connections. Both these measures are positively correlated ($r = 0.29$, $p < 0.001$), which lends support towards the assumptions that average path length is related to orthographic similarity, yet considering that this measure was directly derived from weights, the effect appears to be rather small, so there must have been an effect caused by the indirect paths.

Looking at the violin plots for average weights and average path length across all networks (Figure 14) reveals that considering indirect connections yields a lower score on average. Additionally, while average weights have more outliers above the mean, for average path length, the opposite is the case, with a distribution that is skewed more towards lower values.

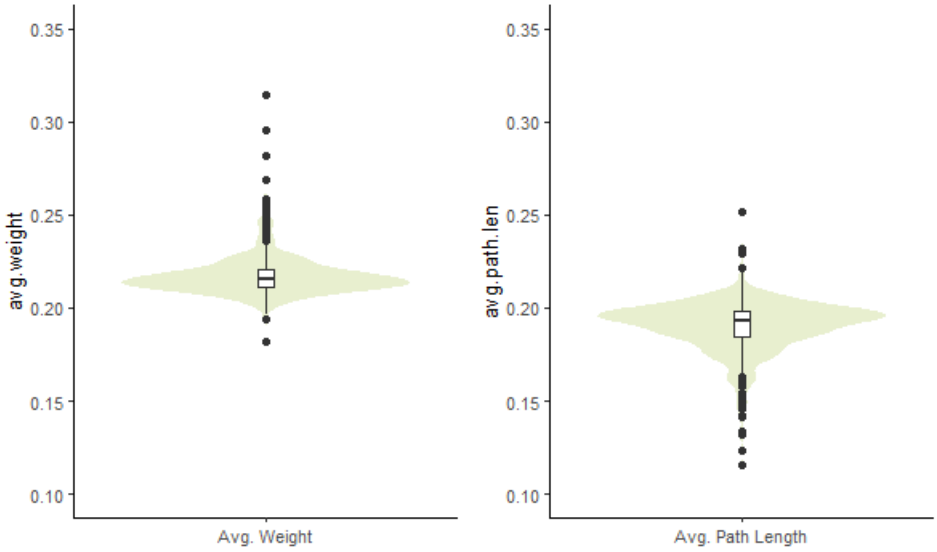


Figure 14. Violin plots for average weight and average path length across all networks.

5.2 Results

For model calculation, the text length and network size variables have been standardized as to bring them on a scale that is more in line with the other variables. The final model was build from a first model that also included average word length as an independent variable. It furthermore included interactions between error rate and network size as well as error rate and average weight, though none of these yielded significant effects and were dropped by comparing larger and reduced models via likelihood ratio tests. Though the final model struggles with high correlations between some of the variables, which could be solved by dropping test point and network size as well as their interaction as predictors, doing so would result in a lower R^2 and higher AIC, rendering an even further reduced model significantly inferior to the final model. The effects of the final model are reported in Table below.

Table 5. Fixed effects output of the final regression model. Significant effects are bold

	Coefficient	**t**	**p**
Intercept	-1.63651	-0.163	0.871
Error rate	-2.90249	-1.688	0.092
Test point	0.05613	0.204	0.839
Average frequency	1.70470	1.911	0.057
Text length (scaled)	**8.69033**	**30.571**	**< 0.001 *** **
Network size (scaled)	1.32744	0.798	0.425
Density	**32.52595**	**2.519**	**0.012 ***
Average Weight	**123.64918**	**2.083**	**0.038 ***
Average path length	**-161.80005**	**-2.326**	**0.021 ***
Test point X Network size (scaled)	**-0.52799**	**-3.748**	**< 0.001 *** **

The residuals of the final model approximate a normal distribution (Figure 15), which is a desired outcome.

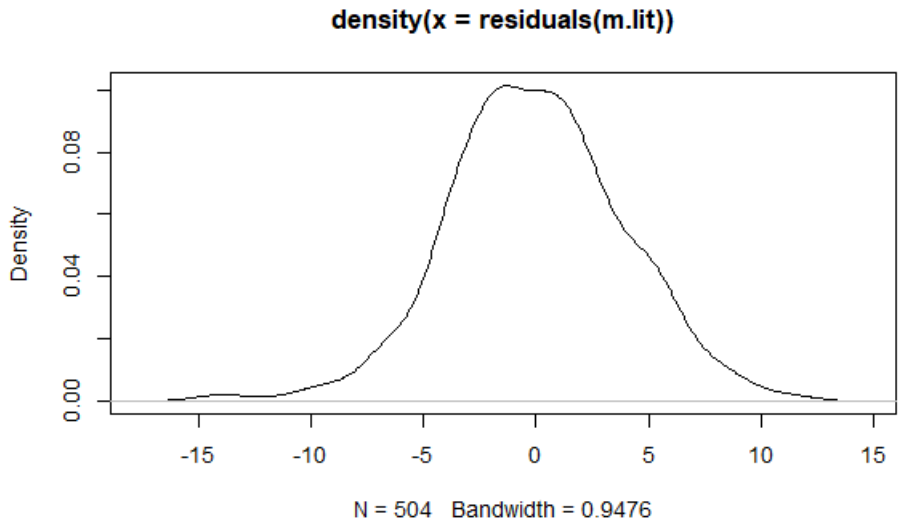


Figure 15. Distribution of the final model's residuals.

The model yields a high (marginal) R^2 of ~0.72, which can be ascribed mainly to the effect of text length. As was expected, text length shows very high predictive power for growth ($t = 30.571$; $p < 0.001$; also see Figure 16), though it does not lead to overfitting the data.

Besides text length, there are three other significant main effects, namely density ($t = 2.519$; $p < 0.05$), average weight ($t = 2.083$; $p < 0.05$) and average path length ($t = -2.326$; $p < 0.05$). Furthermore, the model includes one highly significant interaction between test point and network size ($t = -3.748$; $p < 0.001$). I will come back to these further below.

Interestingly, error rate is not a significant predictor of growth at the next test point, which suggests that children are not necessarily held back by lower orthographic proficiency. Children who make spelling errors probably do so without knowing of their errors, so the mere act of writing out a false spelling should not be a limiting factor. Still, it could be the case that beforehand, orthographic representation are not accessed as fast for less proficient spellers,

maybe due to less consistent representations. This notion is not supported by the results of the model, however.

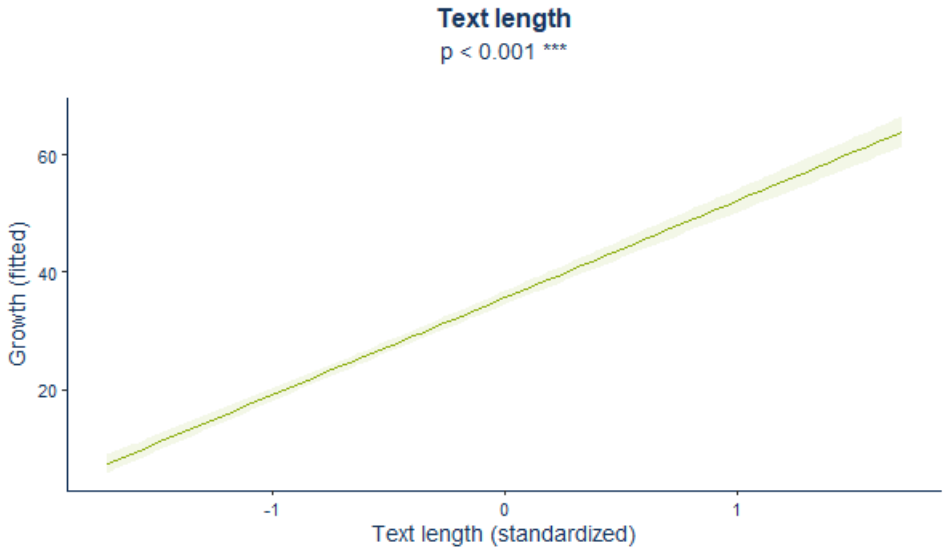


Figure 16. Main effect of text length.

Additionally, the size of a network is not significant predictor of growth either, which stands against the "rich get richer" hypothesis. When considering the highly significant interaction between network size and test point, however, a closer look is warranted. To better understand the interaction, consider Figure 17.

As can be seen, at earlier test points, larger networks tend to grow slightly faster, though this effect is markedly reversed at later test points. This suggests that children who have smaller networks early on will be able to catch up later. Looking at this effect might also evoke the notion of a ceiling to growth, though here this is unlikely, because remember that the network at the final test point represents only a tiny fraction of a child's total vocabulary. Rather, the fact that the growth of larger networks declines at later test points might be caused by the picture stories, as these likely put some restrictions on the words a child might make use of. Time limitations during the test might be a factor as well.

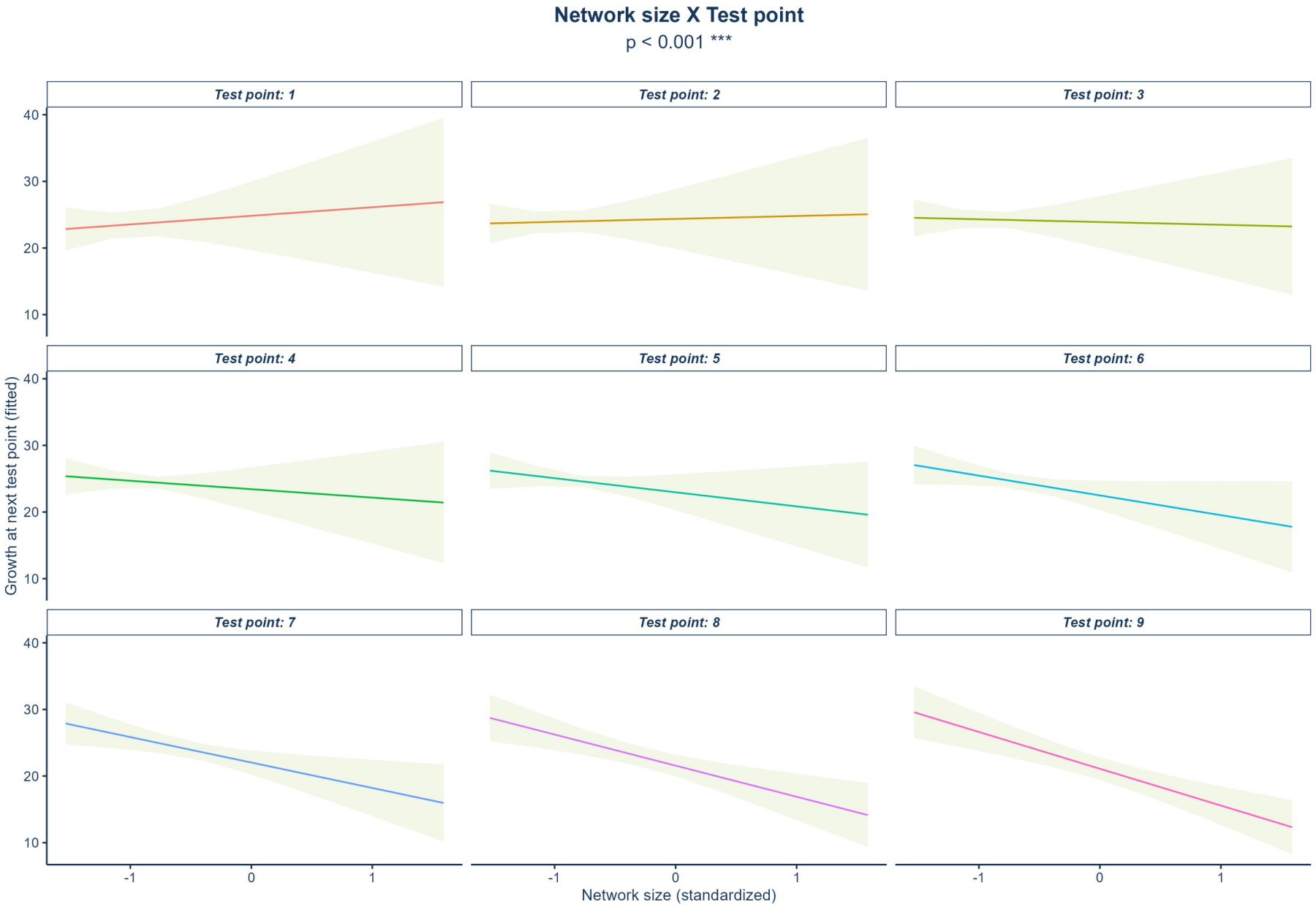


Figure 17. Interaction between test point and network size.

The other main effects are not easily construed, because of what was discussed in section 4.3. Showing the plots illustrates the relationship between the dependent and independent variable, yet in order to pin down what this means, we must again consider the orthographic similarity and its role in constituting edges in the networks.

Average weight and average path length. Figure 18 shows that networks with higher average weights between directly connected nodes exhibit higher growth at the next test point. Crucially, this does not factor in nodes that do not share a connection, so average weight only represents part of a network's similarity structure. Now compare this to the effect of average path length depicted in Figure 19.

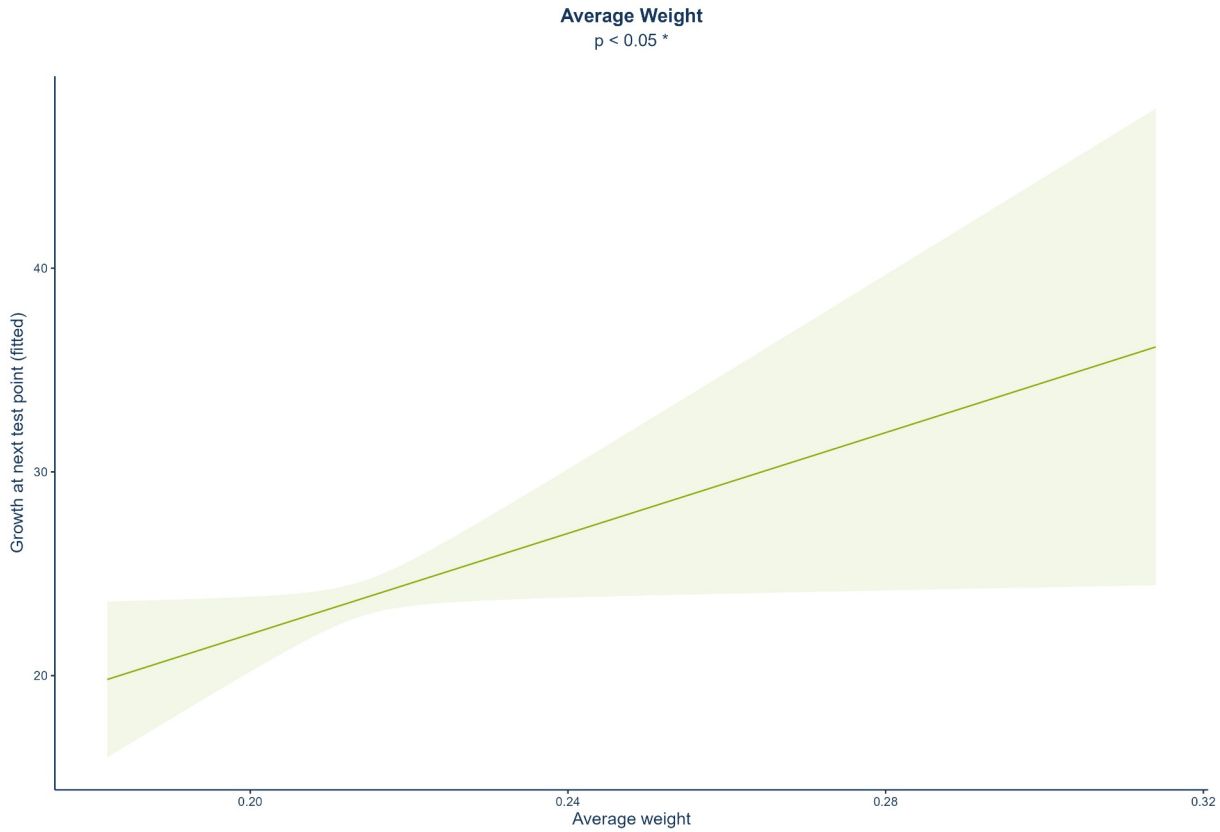


Figure 18. Main effect of average weight.

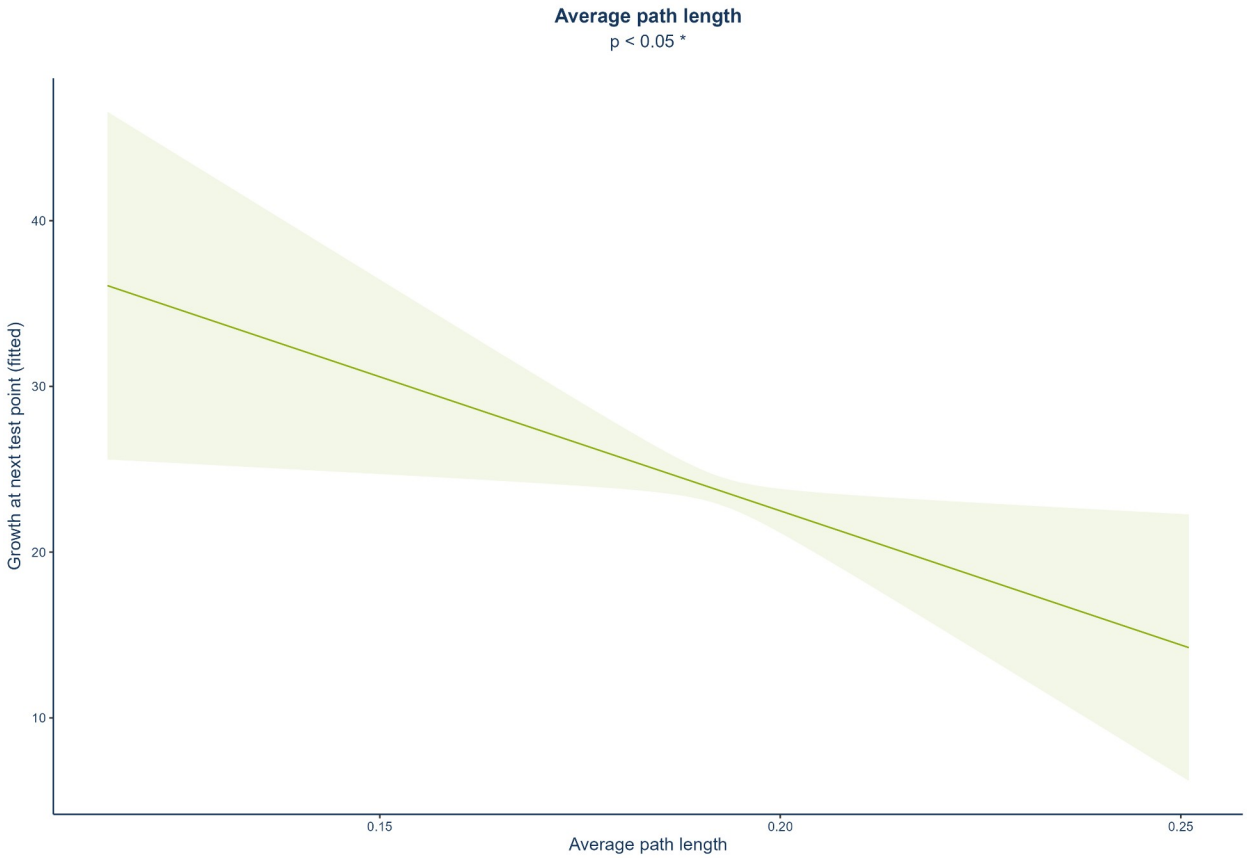


Figure 19. Main effect of average path length.

Interestingly, although average path length can be seen as an extension of average weight, the effect it has on network growth goes into the opposite direction: higher average path lengths predict lower growth at the next test point. Remember that, other than average weight, average path length also factors in nodes that are only connected via an indirect edge. As most of the weights between directly connected edges are retained as path lengths, it must be the influence of indirect connections which causes this reversal of the effect.

In section 5.1, I suggested that high average path length could point to a network that is more homogeneous in terms of orthography. In other words: the child uses a lot of words that are spelled rather similarly. If this was the case, the negative correlation with growth at the next

test point might indicate that children stick to familiar orthography, which limits the vocabulary that is available for usage and therefore leads to a lower amount of new nodes added.

In contrast, as average weights are calculated from directly connected nodes only, high values do not necessarily point to networks that are orthographically homogeneous. While directly connected nodes are fairly similar in networks with high average weights, indirect paths might either lead through nodes with higher or lower similarity. The latter case would yield lower scores in path length, pointing to lower orthographic similarity between unconnected nodes. Because of this, the fact that the effects of average weight and average path length go into different directions must not be a contradiction.

Nonetheless, finding the reason for why high similarity between directly connected nodes leads to increased growth at the next test point is difficult. It could be the case that strong orthographic connections motivate the usage of other word forms that are spelled similarly, though from the analysis in here, it remains unclear whether this is the case.

Density. Figure 20 shows the effect of density. To reiterate, a network's density informs of how connected all nodes inside a network are, with higher values pointing to higher connectivity. The model shows that denser networks exhibit higher growth at the next test point. As was shown in section 5.1, density and average weight are weakly and negatively correlated. Between density and average path length, however, there actually is a strong, positive correlation ($r = 0.64$, $p < 0.001$).

So while denser networks tend to be lower in average weight between directly connected nodes, because of the fact that these networks have a relatively low amount of indirect paths, the average path length should not deviate from average weight that much. So high average weights remain relatively high as average path lengths in dense networks. Low density networks on the other hand have more indirect paths, so average path length might deviate more from average weight. As indirect connections usually have rather low scores as compared

to direct connections, average path length should drop more for less dense networks. This is supported by the high positive correlation between density and average path length.

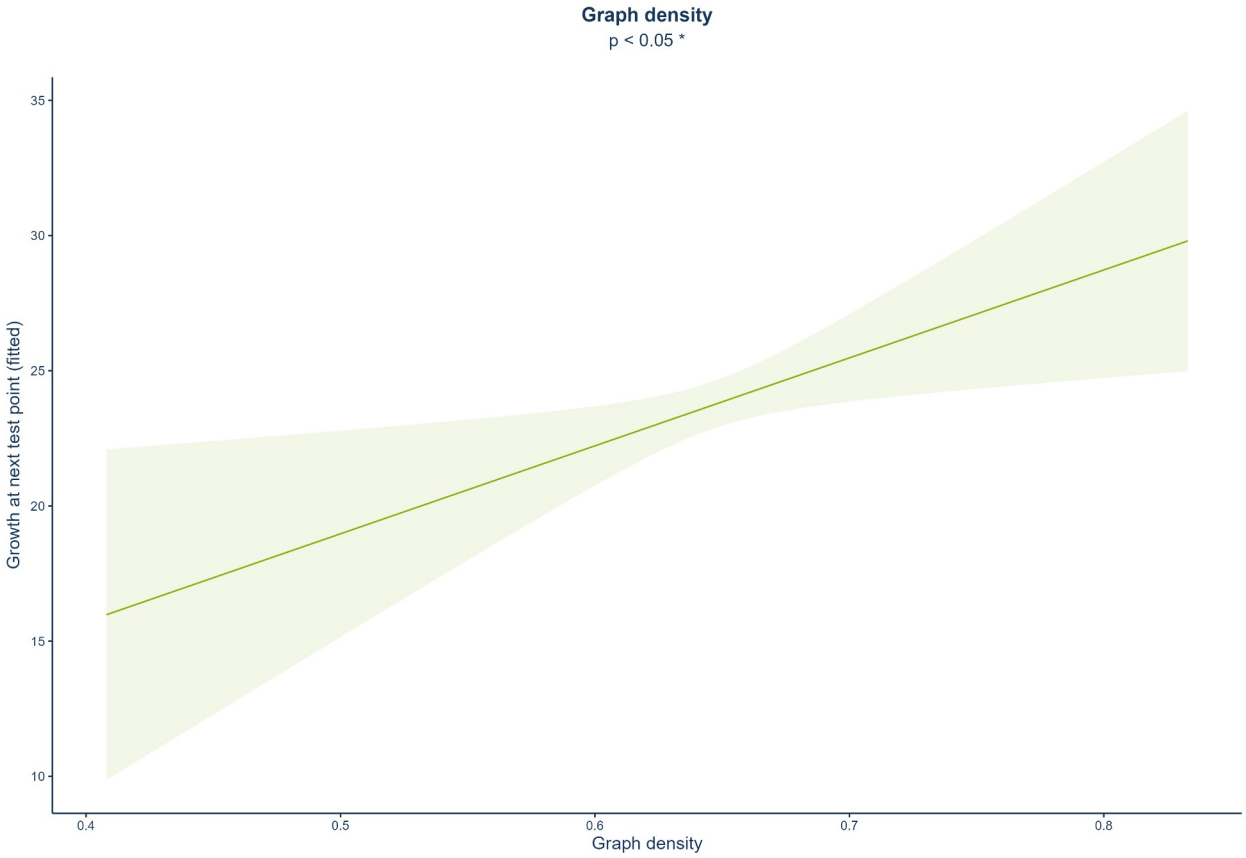


Figure 20. Main effect of density.

Accordingly, in section 5.1 I hypothesized that higher density and average path length might both inhibit growth at the next test point, though the results of the model show that this is not the case, and that the effect of density actually goes in the opposite direction. As was pointed out in that section as well, in contrast to average path length, density does not involve a measure of orthographic similarity, so a dense network can be both, orthographically diverse as well as homogeneous. This could explain why density and average path length exhibit different effects on growth.

6 Discussion

The statistical analysis in section 5.2 revealed that network growth is best predicted by sheer output at a given test point. Children who compose longer texts, also add more new words to their networks. Besides this effect, growth is also predicted by network characteristics, such as density and average path length. However, integrating these effects into any theoretical framework poses serious problems, on which I will elaborate going forward.

As was shown, all measures elicited from networks are dependent on the method that was applied to define edges. In here, this method involved a normalized orthographic similarity score. This method was chosen over the more commonly employed one-edit distance metric, because otherwise the networks of individual children would have been to sparse to analyze. Assessing the resulting networks, however, revealed that the normalized metric had a bias in that longer words were more likely to connect to other words in the network. In many cases, two words were connected just because they shared a single letter. These connections were often rather low in orthographic similarity and might just as well be attributed to chance. As this affects the statistics of the networks, any conclusions drawn from their analysis should be regarded with caution.

Because average path length factors in indirect connections and it would in some cases choose an indirect path to represent the similarity between two nodes, even though a direct path was available, this statistic might involve a correction for the randomness of the orthographic similarity metric to define edges. Assuming then that high average path lengths stand for more homogeneous networks in terms of orthography, the hypothesis that higher average path lengths would inhibit growth at the next test point was supported by the model. However, I would not advice to take this finding at face value, but rather as a catalyst to test this hypothesis anew employing metrics that are not this fuzzy.

The best place to make adjustments is probably the orthographic similarity metric itself. While the one-edit distance metric was too rigid to set up individual networks, the metric employed in here was arguably too liberal, introducing some amount of chance to the edges. One solution could be to determine a higher threshold than 0 to establish an edge between words, though finding a sensible value is not trivial.

Instead of setting a higher threshold, all connections between words who are only connected because of one overlapping letter could be severed. Formally, this would pertain to words whose edit distance equals the length of the longer word minus 1. Cutting those connections should reduce randomness. One could go even further by only connecting word pairs who share at least one bigram. In any case, the effect on network statistics have to be thoroughly evaluated.

Another aspect of this study involved the inspection of visual networks. While observations regarding the structure of the networks are certainly influenced by the orthographic similarity metric, the visualizations still provide easy-to-read information about each child throughout the ten test points. For example, on the level of individual nodes, inspecting the visual networks can point to specific domains where the children are struggling with orthography. Furthermore, because the overall error rate is depicted for each test point, looking at networks throughout all test points outlines development in spelling proficiency.

To summarize, network analysis allows us to access data from different angles, though in order to make meaningful inferences, one has to thoroughly consider the method to build the networks. While the literature on network science is rapidly expanding, it is not always feasible to apply established methods, because the data basis might not allow for it. In this work, the statistical analysis was run on rather obscure measures, so its results are unreliable. Still, the assessment of the different building blocks that went into the construction of the networks could be used to refine the methods applied and run another analysis.

References

Berg, K. (2019). *Die Graphematik der Morpheme im Deutschen und Englischen*. Berlin, Boston: De Gruyter. https://doi.org/10.1515/9783110604856

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

Eisenberg, P. (2006). *Das Wort. Grundriss der deutschen Grammatik* (5th ed.). J.B. Metzler.

Frieg, H. (2014). *Sprachförderung im Regelunterricht der Grundschule: Eine Evaluation der Generativen Textproduktion* (Doctoral dissertation, Ruhr-Universität Bochum).

Fruchterman, T. M. J., & Reingold, E. M. (1991). Graph drawing by force-directed placement. *Software: Practice and Experience, 21*(11), 1129–1164.

Kidd, E., Donnelly, S., & Christiansen, M. H. (2018). Individual Differences in Language Acquisition and Processing. *Trends in Cognitive Science, 22*(2), 154–169. https://doi.org/10.1016/j.tics.2017.11.006

Kumar, A. A., Lundin, N. B., & Jones, M. (2022). Mouse-mole-vole: The inconspicuous benefit of phonology during retrieval from semantic memory. *Proceedings of the Annual Meeting of the Cognitive Science Society*. https://escholarship.org/uc/item/4164178j

Laarmann-Quante, R. (2021). *Prediction of spelling errors in freely-written texts of German primary school children*. (Doctoral dissertation, Ruhr-Universität Bochum, Universitätsbibliothek). https://doi.org/10.13154/294-8266

Laarmann-Quante, R., Ortmann, K., Ehlert, A., Masloch, S., Scholz, D., Belke, E., & Dipper, S. (2019). The Litkey Corpus: A richly annotated longitudinal corpus of German texts written by primary school children. *Behavior Research Methods, 51*, 1889–1918. https://doi.org/10.3758/s13428-019-01261-x

Mani, N., & Ackermann, L. (2018). Why do children learn the words they do?. *Child Dev Perspect*, *12*, 253–257. https://doi.org/10.1111/cdep.12295

Siew, C. S. Q. (2019). spreadr: An R package to simulate spreading activation in a network. *Behav Res, 51*, 910–929. https://doi.org/10.3758/s13428-018-1186-5

Siew, C. S. Q., & Vitevitch, M. S. (2019). The phonographic language network: Using network science to investigate the phonological and orthographic similarity structure of language. *Journal of Experimental Psychology: General, 148*(3), 475–500. https://doi.org/10.1037/xge0000575

Siew, C. S. Q., & Vitevitch, M. S. (2020). An investigation of network growth principles in the phonological language network. *Journal of Experimental Psychology: General, 149*(12), 2376–2394. https://doi.org/10.1037/xge0000876

Siew, C. S. Q., Wulff, D. U., Beckage, N. M., Kenett, Y. N. (2019). Cognitive network science: a review of research on cognition through the lens of network representations, processes, and dynamics. *Complexity*, 2108423. https://doi.org/10.1155/2019/2108423

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78. https://doi.org/10.1207/s15516709cog2901_3

Trautwein, J. (2019). The mental lexicon in acquisition. Assessment, size & structure. (Doctoral dissertation, Universität Potsdam). https://doi.org/10.25932/publishup-43431

Trautwein, J., & Schroeder, S. (2018). Orthographic similarities in the developing mental lexicon. Insights from graph theory and implications for orthographic development. *Frontiers in Psychology, 9*, 2552. https://doi.org/10.3389/fpsyg.2018.02252

Treiman, R., & Kessler, B. (2014). *How children learn to write words*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199907977.001.0001

Vitevitch, M. S. (2008). What can graph theory tell us about word learningand lexical retrieval? *Journal of Speech, Language, and Hearing Research, 51*, 408–422. http://dx.doi.org/10.1044/1092-4388(2008/030)