DIGITAL EXPRESSIVE MEDIA for SUPPORTING EARLY LITERACY
through CHILD-DRIVEN, SCAFFOLDED PLAY

by

Ivan Sysoev

B.Sc., Novosibirsk State University (2009)
M.Sc., Novosibirsk State University (2011)
M.Sc., Georgia Institute of Technology (2014)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences
at the Massachusetts Institute of Technology
May 2020

Author
……
Program in Media Arts and Sciences
May 8th, 2020

Certified by
……
Deb Roy
Professor of Media Arts and Sciences

Accepted by
……
Tod Machover
Academic Head, Program in Media Arts and Sciences

DIGITAL EXPRESSIVE MEDIA for SUPPORTING EARLY LITERACY
through CHILD-DRIVEN, SCAFFOLDED PLAY

Ivan Sysoev

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning
on May 8th, 2020,
in partial fulfillment of the requirements for the degree of
Doctor of Philosophy in Media Arts and Sciences

Abstract:

Digital technology holds many promises for supporting early literacy development. To stimulate both learning achievement and children's interest in literacy, it is beneficial for a learning activity to be playful, support children's agency and self-efficacy, and meaningfully connect to their life. However, nearly all current literacy technology, designed within the instructionist paradigm, lack these qualities. This work attempts to address this issue by exploring the design space of technology that is: (1) "child-driven"—allowing initiative and ideas to come from the learner; (2) expressive—fostering the creation of messages or artistic artifacts; and (3) scaffolded—assisting the child, in real time, in accomplishing his/her self-selected goals. Several forms of scaffolding were explored: (1) direct guidance routines with input from the child, (2) facilitating invented spelling, and (3) phoneme-based building blocks aimed at eschewing the orthographic complexities of English. The exploration was conducted through two apps, primarily aimed at phonological awareness development—minimalistic SpeechBlocks I and scaffolded SpeechBlocks II. They were evaluated in four exploratory studies, both in classrooms and homes.

The following was learned: (1) The media sparked intrinsic motivation, supported agency and self-efficacy, and allowed for non-trivial expression; (2) They were used in markedly different ways: from chaotic, impulsive exploration to sophisticated imaginative play; (3) The media encouraged literacy-oriented social play; (4) Real-time, built-in scaffolding was essential in supporting the meaningful participation of early literacy learners. It allowed children to engage in high-level creativity, while simplifying the necessary low-level routine; (5) Different scaffolding types fulfilled different functions, such as responding to children's specific requests and facilitating the search for ideas; (6) The distinction between letter and phoneme blocks was ultimately less important than originally thought. However, onomatopoeic mnemonics (designed for phoneme blocks) were helpful for a certain category of children; (7) Initial phonological awareness and executive function appear to be moderators in how productive children's engagement was with the media. This work can provide insights to researchers, educators, and designers on how to combine children's agency with supportive guidance.

Thesis advisor:
Deb Roy
Professor of Media Arts and Sciences

DIGITAL EXPRESSIVE MEDIA for SUPPORTING EARLY LITERACY
through CHILD-DRIVEN, SCAFFOLDED PLAY

Ivan Sysoev

This dissertation has been reviewed and approved by

Deb Roy, advisor

……

Professor of Media Arts and Sciences
MIT Media Lab

DIGITAL EXPRESSIVE MEDIA for SUPPORTING EARLY LITERACY
through CHILD-DRIVEN, SCAFFOLDED PLAY

Ivan Sysoev

This dissertation has been reviewed and approved by

Mitchel Resnick, committee member

……

LEGO Papert Professor of Learning Research
MIT Media Lab

DIGITAL EXPRESSIVE MEDIA for SUPPORTING EARLY LITERACY
through CHILD-DRIVEN, SCAFFOLDED PLAY

Ivan Sysoev

This dissertation has been reviewed and approved by

Catherine Snow, committee member

……

Patricia Albjerg Graham Professor of Education
Harvard Graduate School of Education

Acknowledgements

Antoine de Saint-Exupéry once said that human beings apart from each other are like a pile of stones, but together they are like stones in a cathedral. Working on this dissertation led me to deeply experience the meaning behind his words. Without the help of many minds, hands, and hearts, this work would have never come to fruition. On these pages, I would like to express my tremendous gratitude to all the people who made it possible.

First, I would like to thank my advisor, Deb Roy. Deb recognized my passion and capacity for children learning research earlier than I did myself. In one of our early group meetings, we hosted guests who talked about their work, at the time, called Global Literacy Collaborative. I listened to them with great interest while thinking how great it must be to do that kind of work. Coming from a very different background, I had never thought that I would be capable of doing anything similar. My interest, however, didn't escape Deb. About a week later he asked whether I would like to apply the methods of our group to the literacy learning domain. His encouragement started the journey which has brought me so much joy. Aside from Deb's constant efforts to discover and support his students' passions, he is also remarkable for always striving to do work that empowers people. Furthermore, I greatly appreciate Deb's erudition and analytical mind.

I would also like to thank my committee members, Mitch Resnick and Catherine Snow. Since I first read Mitch's thought-provoking articles, his philosophy of learning continuously influenced my work. I also benefited from his responsiveness and availability to students, thanks to which I had multiple opportunities to discuss my work with him long before he became my committee member. What I learned from these meetings, as well as from the classes taught by Mitch, played a crucial role in shaping my approach. Finally, I see Mitch as a perfect example of the qualities that a person working with children should have: always gentle, warm, compassionate, and sparkling with curiosity.

Working with Catherine, I constantly enjoyed how she instantly grasped any problem presented to her and immediately returned very useful advice. This was particularly impressive given the difference in paradigms between the "worlds" of literacy learning and constructionism, which often made it difficult to explain my work to either camp. That was never an issue with Catherine. Her right-to-the-point feedback was invaluable for me in navigating the sophisticated world of literacy learning. I also truly enjoyed her candid, but very warm and supportive style. Finally, I really appreciate the amount of effort Catherine put into supporting my work. She even corrected my grammatical mistakes throughout the text—something only a rare committee member would do!

Two other people whose support has been invaluable are Susan Fine and Jim Gray. Susan, a clinical faculty member at Northeastern University, was instrumental in helping organize both the SpeechBlocks I and SpeechBlocks II pilots. It is hard to imagine this work succeeding without her incredible generosity, kind heart, and enthusiasm. Borrowing a phrase from Susan's repertoire, what a fabulous collaborator she was! Jim, who joined our group from Sesame Workshop, brought

with him a great experience of working on technologies designed for playful learning. He was an early enthusiast of the ideas implemented in SpeechBlocks II, such as phoneme blocks, and it was his encouragement and his thoughts that helped me shape these ideas. He also played a key role in organizing and running SpeechBlocks II pilot. Both Susan and Jim invested a lot in this work. I hope its outcome will delight them. I also would like to acknowledge Marianna Walker, a professor at East Carolina University, whose interest in SpeechBlocks I drove our largest study with that app.

This work also wouldn't be possible without my colleagues. Anneli Hershman's wonderful skills in working with children, parents, and teachers were irreplaceable for all SpeechBlocks I studies. Anneli and Juliana Nazare, who joined us one year later, were the driving forces behind the home studies with SpeechBlocks, from establishing contact with partners and families to developing the curriculum to administering the assessments. Sneha Makini pioneered the picture canvas and the semantic association network that I later used in SpeechBlocks II. In a truly exciting fashion, Sneha's and my research influenced each other. Mina Soltangheis and I jointly developed PlayTrees, the key analytical tool in making sense of SpeechBlocks I data. Marc Exposito helped me to develop a taste for the Unity game engine, and Raphael Schaad introduced me to the basics of design thinking. Juliana, Marc, Sneha, Eric Chu and Lauren Fratamico created elements of SpeechBlocks and most of the supporting infrastructure for the home studies. Sneha contributed to the look-and-feel of late SpeechBlocks I. Martin Saveski, Juanita Buitrago, Jim Gray, and Heather Pierce made important contributions to the home studies. Anneli, Jim Gray, Nazmus Saquib, and Manuj Dhariwal accompanied me on play-testing sessions.

Several East Carolina University students participated in the second home study as coaches. While I cannot disclose their names for the sake of participant identity protection, I would like to express my sincere gratitude to them. I also would like to thank the students from ECU who volunteered to assist with pre- and post-assessments.

Six students from Northeastern University, Allie Shoff, Christine Schlaug, Kathryn Falvey, Anne Mathew, Mackenzie Russell, and Lauren Brinker, were the engines that powered the SpeechBlocks II study. They administered the assessments, took qualitative observation notes, and often played the role of facilitators. Without their help, I would not have had sufficient capacity to handle the overwhelming array of tasks that occur in the classroom. Allie and Christine also generously volunteered to annotate the data from the study.

Several students participated in the apps' development and studies through the Undergraduate Research Opportunities Program (UROP). Abel Tadesse helped to develop an early text recognition module for SpeechStickers. Katherine Wang evaluated off-the-shelf speech recognition tools on children's voices and confirmed their usefulness. Claire Traweek helped to annotate the data from the first SpeechBlocks study.

An important part of SpeechBlocks II is the collection of eighty two animated onomatopoeic mnemonics, or "sound creatures". Colin Campbell helped me to learn the basics of animation, so that I could create some of the "creatures". But the majority of them were animated by three

talented MIT students, Allan Gelman, Danny Gelman, and Jesso Wang, as well as a professional animator, R Ryan Hayes from Blackridge Studio. Additionally, two more animations were done by Abraham Tena (who also designed his “creature”) and Lingxi Li. Through work of these individuals, the onomatopoeic mnemonics sprang into life in a most beautiful way.

Several people helped me to navigate through the jungle of statistics: my colleagues Martin Savesky, Nabeel Gillani, and Brandon Roy, as well as Mohit Karnani from MIT, and professor Howard Cabral from Boston University. They all generously volunteered a lot of their time to answer my (often confused) questions.

I also would like to thank the teachers and staff at the schools and afterschool programs we worked with. To protect the identities of participants in the studies, I won’t state their names here, but my gratitude to them isn’t diminished because of that. I also would like to thank all children and their families who participated in the studies and, in addition to valuable data, brought me so much joy.

A number of people proofread and copy-edited this document, fixing many mistakes that I made due to not being a native speaker of English. My gratitude goes to Catherine Snow, Thomas Hagerty, Anneli Hershman, Juliana Nazare, Bill Powers, Keyla Blanco Gomez, and Olivia Stevens.

This work was influenced by knowledge and creative ideas of many people. Cynthia Breazeal and Tinsley Gallean conducted the above-mentioned talk on Global Literacy Collaborative, which sparked the present endeavour. Sep Kamvar’s work on the Wildflower network of Montessori schools attracted my interest to the Montessori approach, which had a profound influence on my designs. I am also grateful to the Montessori teachers who provided important insights about their approach and allowed me to observe their classrooms. Karen Erickson was my General Exam committee member and contributed greatly to my knowledge about literacy development. Brian Magerko advised my Master’s research at Georgia Tech and helped me develop interest in computational play. Karen Brennan’s ideas about the relationship between agency and structure guided my thinking about automatic scaffolding. Jacqueline Kori-Westlund’s research informed my design of the Mr. Fox interface for speech recognition. Winston Chen introduced me to the Word Wizard app and many other relevant information about the world of commercial literacy apps. Hal Abelson read my early writeups and provided valuable feedback. I had many fascinating discussions with Allen Gorin. Creators of the wonderful Tools of the Mind program, in particular Elena Bodrova, refined my thinking about scaffolding and drew my attention to the issues related to executive functioning. Of course, this work wouldn’t exist without the foundational ideas of Seymour Papert and my compatriot Lev Vygotsky.

My group administrators, Heather Pierce and Keyla Blanco Gomez, made sure that I had everything I needed to successfully complete this research. It is because of their difficult and stressful work that I was able to do my investigations in peace. They also truly were the soul of our group. I also appreciate the tireless work of Russell Stevens, who ensured that our research always had sufficient funding. Keira Horowitz, Sarra Shubart, and Linda Peterson helped me navigate the

academic program. David Bonner from Network and Computing Systems (NeCSys) helped us to facilitate the infrastructure for the second home study.

This work was primarily financed by internal funds of the lab. However, we were at a luxury to do so mainly because of the generous grants that we received from Twitter, as well as financial support from Media Lab member companies. Some funds for the second home study came through the Language, Literacy, and Technology Fund of the Department of Communication Sciences and Disorders at East Carolina University Medical & Health Sciences Foundation. My gratitude goes to these sources for making the present work possible.

It would have been difficult to complete this work without good sources of inspiration. For me, two such sources were the works of Hayao Miyazaki and Ursula Le Guin. Many times, reading or watching them, I thought: if only in my life I could create just one thing this beautiful, it would be a life well-spent. Their characters kept shining like beacons of true humanity. A lot of inspiration also came to me through the thoughtful lectures on meditation by Dada Sadananda. During one of the most stressful periods of my research, I used to listen to them on the bus to and from the school where I ran the study, and these hours of commute turned into the most peaceful and delightful hours of my day. I was later glad to have many conversations with kind Dada Diiptimanananda—a person who I think is like a real-life Joseph Knecht from The Glass Bead Game.

A lot of support and warmth was also received from my friends. I was very lucky to find so many good friends far from home. Each and every one of my labmates enriched my life in different ways. But I would particularly like to mention Bill Powers, Soroush Vosoughi, Ann Yuan, Collmann Griffin, and Flora Su, who truly were a blessing and kindred spirits. Flora also must be mentioned as the voice of SpeechBlocks, painstakingly recording dozens and dozens of lines for various components of the system.

The very foundations of this work, and my entire life, were laid down by my family. For example, I developed my love of reading thanks to my mother and my grandfather. When I was about three or four, my grandfather, Vladimir Sysoev, often took me on walks and taught me reading by pointing to various signboards and helping me decipher them. He also had a very curious and imaginative mind, and sparked the same qualities in me. My mother, Olga Sysoeva, is largely responsible for my childhood being surrounded by books. We used to have more books than furniture to store them. A few kitchen cabinets became impromptu bookshelves; the books were stored on the shelves in two layers, so that you had to remove the books in front to reach those on the back. A story indicative of my mother's love of literature is that, once, while having a monthly salary of 115 Soviet rubles at the time, she spent 100 roubles to buy a copy of Hemingway on the black market (such books were in short supply during that time). My step-father, Ilya Kochurov, ignited my interest in programming and encouraged me to pursue studies at the Novosibirsk State University, and later abroad. He also suggested several great ideas regarding some design elements of SpeechBlocks, such as its text recognition interface. I'm very lucky to call my mother and step-father my genuine friends—a friendship between generations that is not that common. I also

would like to express gratitude to my childhood teachers—in particular, Valentina Nikitina—who nurtured my love of learning.

We tend to attribute creative accomplishments to specific people. But writing down this long list of names makes me think that the true author of any work is the fundamental unity of all humans, if not all beings on Earth. I am truly proud and happy to be a stone in this great cathedral. I can only hope to serve my function in it well, so that me and my work could be as helpful to someone as all these people were kind and helpful to me.

Contents

Chapter 1. Introduction

1.1. Child-Driven, Expressive, Scaffolded, Digital: A Promising Path to Early Literacy

The land of the written word is full of treasures. How can we help children become native to it? Native not only in the sense of knowing the language, but also in the sense of feeling comfortable and at home, identifying with the place, and wishing to return there again and again? How can we make sure that citizenship in that land is granted to everyone regardless of their social and economic status, race, gender, or country of origin? These questions are unlikely to have a single, simple answer. However, works of prominent educators and advances in technology highlight one promising direction: building upon children's innate curiosity, playfulness, and urge to learn. The power of giving learners agency and making the subject personally meaningful to them was shown by such approaches as Montessori (P. P. Lillard, 1972), Waldorf (Clouder & Rawson, 1998), Reggio Emilia (Edwards et al., 1998), and constructionism (Papert, 1980). In the literacy domain, it is highlighted by researchers of emergent literacy and invented spelling (Bissex, 1980; Richgels, 2001; Strickland & Morrow, 1989). Today, the ever-growing capabilities of mobile digital technology provide a rich soil on which these educational ideas can flourish. At the same time, the ubiquity of mobile devices offers a promise to deliver such a model of learning to children from all strata of society. The aim of this thesis is to explore how ideas of Montessori, Vygotsky, and Papert can be planted onto digital soil in application to early literacy learning.

I approach this problem by means of design exploration: through developing prototypes of early literacy apps and evaluating them with children. I focus on exploration, because this design space is large and has previously received little attention. My work concentrates on a foundational component of early literacy: phonological awareness (PA). My designs were also created having in mind development of another literacy skill, letter-sound-pattern correspondence, but I don't evaluate their efficacy with respect to that skill. My focus is on typically developing children who are actively in the process of acquiring those skills. In the US, this usually happens between the ages of four and six. The approach described in this work is characterized by three key principles, child-driven, expressive, and scaffolded, applied to the digital domain. Below is the reasoning behind why these three principles were pursued.

One of the key aspirations of my work was to design an experience that would be intrinsically motivating. Accomplishing this goal would allow for the possibility of children using the system outside of formal settings, e.g. at home. It could also help them to perceive literacy activities as something joyful, interesting, and motivating. In this work, I approach this goal by making the experience **child-driven**, which means that the initiative and ideas come from the learner. Child-driven experiences can facilitate the sense of a learner's agency and self-efficacy, which have

a strong emotional significance for children (A. S. Lillard, 2016; Strickland & Morrow, 1989). Striving for self-efficacy, the capacity to cause changes in the world, is a manifestation of the intense drive towards self-actualization that Montessori observed in all children (P. P. Lillard, 1972). Speaking broadly, I believe that it is important to support and foster this drive early on to allow for fuller unfolding of human potential. There is another advantage to the child-driven approach—it also allows children to connect their play to their life experiences and interests. This not only motivates them, but also situates new knowledge in the context of existing interests which helps it to be retained (Holt, 1989; A. S. Lillard, 2016).

One of the most sophisticated ways in which the self-efficacy drive manifests itself is the desire for **expression**. This urge can be seen in children's love for drawing, telling stories, building elaborate sand castles, making outfits for their dolls, and crafting. It explains the worldwide popularity of Lego, Knex, Meccano and its relatives[1], and other construction kits. Expression can facilitate learning by letting children get to know the materials, techniques, principles, and ideas involved (Papert, 1980; Resnick, 2017; Resnick & Rosenbaum, 2013). Various forms of child expression serve as languages that can be used for both communication and thinking (Edwards et al., 1998). It empowers children by letting them leave a mark in the world (Strickland & Morrow, 1989). It is also a deeply human tendency; no other species is known for the capability to produce sophisticated forms of expression that are not rigidly prescribed by their genetic makeup. This work takes a stance that expression is a significant part of being human, and it is desirable to let it unfold as early as possible. Its subject is therefore *expressive media*—technologies that foster creation of messages or artistic artifacts.

Children's communicative and expressive intentions are ambitious. But we are looking at children whose literacy skills are not fully developed. In order to express themselves by literacy means, such learners need something akin to training wheels on a bike. The wheels make it possible for a child to ride around "for real" even when their own skill wouldn't permit this. By making riding feasible, the training wheels encourage them to be on the bike more often. Through this, the child learns to balance, and eventually the wheels are no longer needed. This principle is described as "**scaffolding**" in the seminal paper of Wood, Bruner, and Ross (1976); it is in turn based on the theory of socially grounded learning developed by Vygotsky (1978). In this work, we look at scaffolding in a narrow way: as a means to help children accomplish their own self-selected goals. When scaffolding is viewed more broadly, it has other functions, such as helping a child to set up goals in the first place, recruiting their attention to the learning activity, etc. These functions are outside of the scope of the present work.

Scaffolding can potentially resolve a certain tension between the focus on the child-driven design and the need for explicit, systematic phonics instruction suggested by early literacy research (Beck & Beck, 2013). "Explicit" refers to presenting letter-sound patterns directly, rather than relying on children discovering them. "Systematic" implies a certain order of presenting the material: more

[1] In my childhood, I knew this toy as Конструктор-Строитель (Konstruktor-Stroitel), or "Engineer-Builder", and enjoyed it greatly.

common letter-to-sound correspondences should be introduced before less common ones, monosyllabic words before polysyllabic ones, and so on. In conflict with that, when children learn by building words of their own choice, it is impossible to predict which words they would select to build or which letter-sound combinations they would need. Scaffolding allows explicit presentation of letter-sound combinations that are relevant to the child at the moment. Therefore, it effectively allows for explicit, systematic instruction without the need to resort to a fixed, teacher-driven progression of material. The potential of such methodology is evidenced by the effectiveness (in terms of facilitating phonological awareness development) of invented spelling-based approaches (Richgels, 2001), as well as of the Montessori approach (Franc & Subotic, 2015). A significant part of the Montessori approach to literacy is letting children make words of their own choosing using the so-called Moveable Alphabet, with the support of the teacher. The teacher supports the child interactively, invoking a variety of phonics techniques.

Montessori's Moveable Alphabet routine shows an example of human-provided scaffolding. Scaffolding can also be present in the learning environment, or incorporated within learning materials itself (Bodrova, E., & Leong, D. J., 2001). Advances in "intelligent" user interfaces also allow for a hybrid between the two approaches: scaffolding provided by a machine, but in a human-like fashion. In this work, we focus on scaffolding that is in one way or another incorporated into the medium. Such an approach allows children to receive on-demand, just-in-time guidance directed towards their individual goals without having to rely on one-on-one, in-person interaction with a skilled adult. This offers better scalability and suitability to low-income settings, simplifies collaborative play between children, and makes the media more readily available for use at home. Decreased reliance on skilled adult support might also be important in light of the fact that findings of literacy learning research were slow in making their way into classrooms (Castles et al., 2018), and many teachers still rely on outdated methods of instruction, such as the whole-word approach to early reading.

It is worth noting that the "child-driven" and "scaffolded" principles might appear somewhat at odds with each other. Wood, Bruner, and Ross (1976) argue that one of the key functions of scaffolding is "reduction in degrees of freedom." Therefore, when scaffolding mechanisms are active, the child's agency may be somewhat limited by them. However, in my experience, the opposite was the case: scaffolding enhanced children's agency. The causes of this will be examined.

This work applies the above-mentioned principles to the **digital** domain. Some readers might question this choice: technology is often perceived as addictive and a major distractor of children's attention (Guernsey & Levine, 2015). Indeed, it has been shown that improperly designed literacy technology can do more harm than good (ibid.). On the other hand, some researchers cautioned against ignoring the potential of digital technology for closing literacy gaps (Pratham, 2019; Guernsey & Levine, 2015). In the context of the present work, such technology is particularly relevant because of its potential to support the above-mentioned three principles and bring them to scale. It offers powerful expressive capabilities, beyond those of traditional media. Its interactive qualities, such as instant feedback, can facilitate child-driven play by supporting autonomous

exploration and tinkering. It can make child-driven, scaffolded play more scalable by making the scaffolding machine-provided. Growing "intelligence" of digital devices, such as their ever-improving capability to recognize speech, promises to make interactions with automated scaffolding quite smooth and human-like. At the same time, the proliferation of mobile devices, even among low-income families and in developing countries (Taylor & Silver, 2019), gives an opportunity to deliver playful literacy learning even to disadvantaged children.

To summarize, the confluence of child-driven, expressive, and scaffolded principles applied to digital technology offers a promising, and previously underexplored, approach to early literacy. The aim of the present work is to examine this approach by means of design exploration. It does so through the lens of two early literacy apps that I designed and developed, SpeechBlocks I and II. SpeechBlocks I is a minimalistic platform that implements the expressive and child-driven dimensions. SpeechBlocks II adds additional expressive capabilities, as well as various mechanisms for built-in scaffolding. Several variations of these designs were tested.

1.2. Research Questions

The designs introduced in this work were evaluated with respect to the following questions:

- **How do children engage with digital expressive media for literacy learning?** We can expect different children to have different ways of interacting with the system. What are the main interaction patterns? Do children manage to express themselves in non-trivial ways? Do they experience agency and self-efficacy? Do they interact "minds-on" or "minds-off", following the terminology of Hirsh-Pasek et al. (2015)? How can their experience be characterized: Engaging? Frustrating?

- **How do children engage with built-in scaffolding?** Does it help or hinder their expression? Does it conflict with their agency or perhaps support it? Is it sufficiently flexible? Does it help children to be sufficiently autonomous?

- **Which type of building block is optimal for such media?** Since English doesn't have one-to-one correspondence between letters and phonemes, children could be given the opportunity to build words either out of letters or out of phonemes. There were reasons to believe that phonemes could be a more appropriate construction material. This question is examined in the present work.

- **Do the media benefit child learning?** In particular, do they help children develop phonological awareness?

In addition to these general questions, a mass of specific observations were collected during the studies regarding which designs seemed to work and which seemed not to work, and what

improvements could be made. They can be aggregated as an answer to the fifth, practical question:

- **What are the recommendations for implementing the media and the scaffolding mechanisms?** These recommendations are, of course, given for a particular technological, social, and cultural context. Nevertheless, other researchers and developers may find the observations applicable to the contexts with which they find themselves working.

1.3. Target Population and Setting

The media discussed here were designed for normally developing children who are actively in the process of building their phonological awareness skill. Although it may have potential for children with language and literacy disorders, this was not investigated in the current work. Some readers might find the focus on normally developing children strange: don't they successfully learn literacy anyway? Even if this were entirely the case, I think there would still be value in making the process more playful, engaging, and fostering children's creativity. In actuality, there are further reasons why this target population is still relevant.

First, there are significant gaps, based on social and economic status, in access to quality resources for literacy learning, and, as a result, in literacy outcomes. These gaps exist even in developed countries, such as the United States. The causes of the gaps are complex and cannot be addressed by technology alone, but researchers and practitioners mention the potential of technology to narrow these gaps, both in developed countries (Guernsey & Levine, 2015) and in the developing world (Pratham, 2019).

Second, there is a large "gap" between children's attitudes to literacy and what would be desirable. The amount of reading for pleasure among younger people is declining (at least in the US; see Ingraham (2018)), and children perceive these activities as "not cool" (Goodwyn, 2014). This is tragic, since literacy plays an important role in the formation of a person and a citizen. Literacy changes the organization and interaction of modules within the brain (Wolf, 2008), thus directly affecting the way we think. There is some evidence that literacy facilitates development of both verbal and non-verbal intelligence (Cunningham & Stanovich, 1998; Ritchie, Bates, & Plomin, 2015), as well as creativity (Anderson, 2006; Meline, 1976; Ritchie et. al., 2013) in a way that non-print media don't. Reading volume is strongly associated with the acquisition of foundational facts about the world, an association that has not been demonstrated for such media as television (Cunningham & Stanovich, 1998). Literacy is also likely to play an important role in the cultural and moral development of a person. An important potential advantage of an approach emphasizing children's intrinsic motivation is that it may be able to help them form a positive attitude towards literacy activities.

In addition to the target audience, there is a question of target setting. My aspiration was to develop designs that would be eagerly used by children at home. Achieving such a goal could

create an important extra channel of literacy acquisition, complimentary to school. However, the present approach might be useful for classroom environments as well.

An important potential for the designs discussed here is to be used as a component of larger, human-machine literacy learning setups. One such setup, involving a literacy expert (called Family Learning Coach) remotely and asynchronously interacting with children's devices, was investigated by my colleagues (Hershman et al., 2017; Nazare et al., 2018). Such setups, however, are outside of the scope of the present thesis.

1.4. Method

To my knowledge, there was no previous technology for early literacy learning that incorporated all four principles (digital, expressive, child-driven, and scaffolded). Therefore, studying this approach went hand-in-hand with the development of such technology, and design exploration became the main mode of the present work. Throughout the course of the work, a question arose multiple times on whether to continue such exploration or to focus on a rigorous experimental evaluation. I made the choice in favor of broad exploration, since it appeared to me that more could be learned from it than from focusing prematurely on rigorous assessment of a specific design.

The methodology of this work is therefore not the one of experimental research, but one of Design-Based Research (DBR) (Barab & Squire, 2004). This methodology is developed within the context of learning sciences and oriented towards creating interventions that work in real-world settings. The primary goal of DBR is not to test hypotheses, but "to look at multiple aspects of a learning design and develop a profile that characterizes the design in practice" (Barab & Squire, 2004). To paint a rich picture of such a profile, DBR simultaneously looks at multiple variables, including outcome variables (e.g. difference in test results), climate variables (e.g. interaction between learners), and system variables (e.g. sustainability, transferability to other contexts). While hypothesis testing involves fixed procedures, DBR allows for iterative adjustment of the procedures and the design based on their performance in practice. Issues that arise in the environment are prescribed to inform the development of the theory. Another difference of DBR compared to hypothesis testing is embracing "messiness" and richness of real-life settings as opposed to striving for an environment where as many variables as possible can be controlled. While DBR doesn't allow one to make statements with the same level of quantitative rigour as hypothesis testing, it allows one to work with complex webs of interconnected factors, which is often the case in education research, and was certainly the case in the present work. DBR has many similarities with formative evaluation (e.g. as in Fisch & Truglio (2014)), both in ideas and methodology. However, formative evaluation is oriented towards assessing the value of a particular product with the purpose of improving it in later iterations, whereas Design-Based Research aims at developing theories that generalize beyond a particular design.

A combination of play-testing and larger, longer-term studies was used to advance the design exploration. Play-testing sessions were conducted in order to determine which design directions were promising. They used intermediate versions of SpeechBlocks or auxiliary programs written to mock up a particular feature. The programs used in the play-testing sessions were typically simplistic, in order to enable rapid iteration over multiple designs. In each session, a few (usually three to six) children used each program, and observations of their play guided development of designs for the next play-testing sessions. Design decisions refined over the course of multiple play-testing sessions were then used to set up the larger studies, each lasting several months, which were intended to study each design in detail and provide more solid evidence regarding its performance. Four such studies were conducted: three with SpeechBlocks I and one with SpeechBlocks II. One SpeechBlocks I and one SpeechBlocks II study were conducted in classrooms in order to enable direct observation of children. They involved children aged 4 to 5. The other two SpeechBlocks I studies were conducted in homes. They were led by my colleagues and conducted with the primary purpose of studying the Family Learning Coach architecture. They involved 5 to 10 and 5 to 8 year old children. The upper part of this age range was older than ideal, but it was chosen in order to make sure that children would be able to use SpeechBlocks independently. At the time, independent use of SpeechBlocks was more difficult, because the scaffolding mechanisms weren't yet available or were limited. For me, these studies provided an opportunity to look at SpeechBlocks in the home context. The primary focus of this thesis is on the classroom study with SpeechBlocks II which is the culmination of all preceding design and research work; other studies will be analyzed in less detail.

Two types of data were used for the analysis. First, different versions of SpeechBlocks were instrumented to collect detailed logs of everything that happens in the digital realm: children's taps and drags (including sequences of coordinates on the screen), animations, state transitions, etc. These logs are sufficient to completely reconstruct the digital facet of the sessions. However, they are oblivious to everything that happened outside the screen: children's stated goals and intents, their interactions with peers and adults, their verbal and nonverbal expressions, etc. This context turned out to be crucial in accurately interpreting how the media were used. Information about it was gathered differently in different studies. In the SpeechBlocks I classroom study, we used video recording. In the SpeechBlocks II classroom study, we used observation notes collected by trained observers. In the home studies, some amount of context could be inferred from the exchange of messages between the literacy experts (the coaches) and the parents.

The statistical analysis used in this work is exploratory. Although p-values are reported to highlight particularly strong and interesting patterns, these patterns should not be considered statistically significant findings. This is because no prior hypotheses were formulated before the start of the analysis, and no attempts were made to address the multiple comparisons issue (Reinhart, 2015). The purpose of the analysis is to show potentially interesting phenomena and to form hypotheses. Confirmation or rejection of these hypotheses remains a matter of future studies.

1.5. Key Learnings

Below are the key learnings derived from the four studies.

As anticipated, **the media sparked intrinsic motivation to play and supported senses of agency and self-efficacy.** Although it hasn't been explicitly investigated, it is plausible that these factors help children establish a positive emotional connection with literacy activities. It also **allowed children to express themselves in non-trivial ways,** which might stimulate development of their creative skills, allowing to combine this process with early literacy learning.

There were markedly different ways of using the media. In the context of SpeechBlocks I, the following types emerged:

- **Impulsive Exploration**—characterized by seemingly chaotic actions driven by short-term rewards;
- **Word Crafting**—focused on building words and word collocations as an end in itself;
- **Proto-narrating—**focused on telling simplistic stories with SpeechBlocks or about words built in SpeechBlocks;
- **Remixing and Rhyming—**focused on morphing words into other words;
- **Communicative Play** —focused on making SpeechBlocks speak on a child's behalf;
- **Using the App as a Reference** —either to copy words from it, or to copy environmental words into the app to sound them out.

In SpeechBlocks II, three types of play emerged:

- **Word crafting**;
- **Imaginative play**—focused on the creation of scenes and stories using imagery and enaction;
- **Impulsive exploration**.

The difference between the observed play types can be attributed to the different designs of the media, to the different setting and age of participants, and to the different study settings.

The media encourages various forms of social play centered on word-making. Such play allows children to learn from each other, to be an inspiration for each other, and to maintain mutual engagement.

Real-time, built-in scaffolding for making child-selected words is essential for maintaining meaningful participation of early literacy learners. While simple forms of play (e.g. tinkering with letter patterns and remixing words) can initially be engaging for children, they lose their appeal as their novelty subsides. More sophisticated and truly expressive forms of play require making real words. Without being able to engage in these forms of play, children (particularly 4 to 5 year olds) become frustrated and disengaged. Human-provided scaffolding was observed to become a bottleneck if there was more than one child per adult scaffolder. Real-time built-in scaffolding enabled children to fluently respond to emerging ideas and thus facilitated engaging, flow-like play. It was instrumental in supporting imaginative play and word crafting. As a result, it contributed to children's sense of agency, rather than inhibiting it. It also facilitated idea borrowing, mutual help, and some forms of shared play. It appeared that at a low level, writing was quite routine and mechanical for most children. Scaffolding simplified the routine processes for children while allowing them to engage in high-level creativity.

Different types of word scaffolding were observed to have different functions that complemented each other:

- **Responding to specific requests**—when children have a concrete idea of what they would like to make.

- **Facilitating search for ideas**—when children are facing a blank canvas or would like to elaborate their creations, but don't know how.

- **Being a fall-back option**—when children experience difficulties with more sophisticated technology.

In the SpeechBlocks II study, **usage of letter vs. phoneme blocks turned out to be a less important factor than it was originally thought**. This happened because most words were created via direct guidance mode, whereas the differences between the two block types were relevant only in the open-ended mode. There is currently no evidence suggesting the advantage of phoneme blocks. Thus, it appears reasonable for a designer to adhere to the more conservative option of letter blocks. However, **onomatopoeic mnemonics facilitated block finding for some children, although not for everyone**. The factors mediating whether or not onomatopoeic mnemonics are advantageous for children are currently unclear.

Analysis performed with SpeechBlocks II shows that **initial phonological awareness and executive function** (the ability to engage and disengage mental resources on various tasks at will) **appear to be moderating factors in how productive children's engagement is with the media.** Imaginative play was associated with good initial phonological awareness (PA) and good

executive function (EF), whereas impulsive exploration was associated with being low on both factors. Children with high PA and EF achieved a good level of autonomy and focused on the principal activity towards the end of the study, whereas children low on these factors exhibited a lot of unproductive behaviors. Analysis of PA gains suggests that children with high initial PA benefitted from their play with the media, whereas those with lower PA might not have. The same relationship is true for EF, although it is less pronounced. Further research is needed to determine whether this issue can be mitigated by a form of scaffolding that more flexibly adapts to a child's skill level. These observations also suggest the necessity of providing other forms of scaffolding, aside from help with word building, such as recruitment into the activity, frustration management, direction maintenance, and modeling. Currently, these forms of scaffolding need to be provided by humans.

Aside from the general learnings, the value of the present work for a designer also lies in specific observations concerning the application of various technologies to scaffolding. The reader is encouraged to visit section 6.5 for these specific details.

1.6. Structure of this Document

The following is a brief outline of what lies on the forthcoming pages. First, the literature background that informs the present work is layed out. The description of the designs follows. Since iterative design evolution took place, there is a circular relationship between the designs and their evaluation. However, for the reader's convenience, the corresponding sections are separated. The design section sometimes looks ahead and refers to some findings that are described in the evaluation sections. In addition, learnings from the play-testing sessions are placed in the design section. This is because these sessions directly informed the design, while being too informal to be considered a true evaluation. Making SpeechBlocks work involved not only coming up with a design, but also solving several algorithmical and machine learning problems. These problems and their solutions are described in the section titled *Algorithms and Models*. Some of them required their own evaluation, which is placed in the same section. The following two sections describe the studies conducted with SpeechBlocks I and SpeechBlocks II respectively. The studies' setup, data collection, and analysis techniques are elaborated, followed by a description of the results. In the case of SpeechBlocks I studies, the focus was on play types, evidence of agency, self-efficacy, and ownership of work, engagement, social play, and the need for built-in scaffolding. Many of the observations for SpeechBlocks I hold for SpeechBlocks II. Therefore, in the description of the results pertinent to the latter, the focus is on specifics of SpeechBlocks II—types of play in the presence of the new features, children's interaction with built-in scaffolding, and analysis of different block types (letters vs. phonemes, with or without onomatopoeic creature). The focus of the analysis is on the last study. The final section restates the key learnings in light of the specific data, fleshing them out with more detail. It also describes unresolved questions and directions for future work.

Chapter 2. Background

This chapter reviews the literature that informed the present work. Although the last century saw tremendous spread of literacy worldwide, there are still many pressing issues with access to good literacy education - both in the developing world, and in many developed countries. Experts consider that well-designed digital technology for literacy learning can help in addressing these issues. However, to be beneficial, this technology needs to reflect the findings of early literacy research, as well as an understanding of how young children learn.

Literacy is a multifaceted skill, but its facets can arguably be roughly divided into those pertaining to encoding/decoding and those pertaining to linguistic comprehension. Encoding/decoding describes the ability to convert the sounds of speech into the symbols of written text, and back. A foundation of this skill is the ability to recognize the sound structure of words, called phonological awareness. It is complemented by the ability to convert the letter patterns into sound patterns. Other capacities, such as short-term memory and executive function, are also important for encoding/decoding. The non-straightforward orthography of English is challenging for literacy learners. Although the discourse on early literacy is dominated by the study of reading, there are good reasons for supporting the development of writing as well, including the fact that writing facilitates children's capacity for self-expression.

Self-expression is tightly connected to many forms of play. The term "play" encompassess a great range of interrelated behaviors, which are not yet fully understood. However, we can confidently say that play is associated with exploration, discovery and learning. In particular, it is shown that play has a crucial role in early childhood learning. This makes the current trend for gradual displacement of play with regimented activities in kindergartens quite unfortunate (Hirsh-Pasek et. al., 2009). Play requires a good amount of agency on the side of the child. Agency and structure were often considered antithetical, though a more nuanced view suggests that agency and structure can mutually reinforce each other. Scaffolding can provide a form of structure that can be beneficial to agency. The concept of expressive digital medium offers promise in applying ideas of child-driven play to early literacy learning. Several works, either prior or in parallel to the current one, explored this topic. However, they differ from the current work in several aspects, such as provision of open-ended built-in scaffolding.

2.1. The Worldwide Gaps in Literacy Achievement

The worldwide situation for literacy can be likened perfectly to a half-empty/half-full glass. While tremendous progress has been achieved in the last century and continues to advance at a fast pace, there are also pressing literacy needs to address in different regions of the world (Roser & Ortiz-Ospina, 2016). More than one billion people, or 13.75% of the world's population, are illiterate. The majority of the illiterate population is concentrated in developing countries, particularly

Sub-Saharan Africa, where illiteracy among youth (15-24 years old) in some countries reaches 60% (Roser & Ortiz-Ospina, 2016). Furthermore, while such countries as India, Uganda and Zambia nominally achieve moderate rates of youth literacy, they struggle with providing an adequate quality of education, particularly in rural and under-resourced urban areas. For example, in 2018 in rural India, 55.5% of 3rd-graders were unable to read a 1st-grade-level paragraph, and 12.1% didn't even recognize letters (Pratham, 2019). Since Indian public school regulations require teachers to cover precisely the curriculum prescribed for a particular grade level, many of these children fall behind and lose touch with what's going on in the classroom (ibid.).

Even in such a developed and prosperous country as the United States, some researchers diagnose a "quiet crisis" of literacy achievement (Guernsey & Levine, 2015). Large achievement gaps exist between various population groups (Vasilyeva & Waterfall, 2011). For instance, in 2011, the achievement gap between students coming from the upper and the lower 10 percent of family income distribution was equal to 1.3 standard deviations, equivalent to 3-6 years of additional schooling (Reardon, 2011). This gap has significantly increased since 1970. After accounting for income differences, there are still large gaps based on race. Interestingly, children from certain low-income minority communities possess oral language (e.g. storytelling) skills superior to their middle-income white peers, but schools fail to capitalize on these skills (Rhyner, 2009; Vernon-Feagans et. al., 2001).

One factor contributing to the income-based reading gap is so-called summer reading loss. During the summer, reading skills of children from families with high and middle socioeconomic status (SES) remain the same or improve, but reading skills of low-SES children become measurably worse (Allington & McGill-Franzen, 2003). This gap amounts to about 3 months of schooling every year (Cooper et. al., 1996), and some researchers have attributed up to 80 percent of reading achievement gaps between more and less privileged students to this phenomenon (Hayes & Grether, 1983). The primary explanation of this phenomenon is the lack of a supportive and stimulating literacy environment outside of school (e.g. at home) (Allington & McGill-Franzen, 2003). Since the causes of summer reading loss lie beyond schools' domain, addressing it requires innovative forms of intervention.

2.2. Potential of Mobile Digital Technology for Narrowing Literacy Gaps

Attitudes towards digital, and particularly mobile, technology in the literacy world have been ambivalent (Guernsey & Levine, 2015), partly because technology is perceived as a major distractor of children's attention, and partly because the majority of available products were designed without taking into account the findings of learning and literacy research (Vaala, Ly, & Levine, 2015). An improperly designed literacy technology - e.g. one that emphasizes nonessential features to draw children's attention while distracting them from the literacy task - can do more harm than good (Bus et al., 2015). Nevertheless, some researchers recognize the potential of

mobile digital technology and view it as instrumental in addressing the aforementioned literacy gaps (Pratham, 2019; Guernsey & Levine, 2015). Such technology offers multiple affordances (Guernsey & Levine, 2015; Shuler, 2009):

- The cost of mobile devices is dropping rapidly, and they are becoming accessible even to low-income families. Thus, well-designed literacy apps provide a new opportunity to reach underserved populations.
- It has been observed that the interface of mobile devices is quite natural and intuitive for children, allowing for autonomous interaction with the devices. In a classroom context, it offers a potential to reduce teachers' workload.
- Relatively autonomous, one-on-one interactions with the devices provide ample opportunities for personalized learning. The rapidly growing "intelligence" of mobile devices (owing to advances in machine learning) promises sophisticated and increasingly human-like interaction between the device and the child.
- Since apps can be quite engaging and attractive for children (Goodwyn, 2014), learners may regularly interact with the technology outside of the classroom. This can contribute to enriching domestic literacy environments and addressing such phenomena as summer reading loss.
- Mobile devices can interact with their surroundings (e.g. through cameras) and are easy to carry around. This provides excellent opportunities for grounded learning.
- Being connected, mobile devices can be a good platform for social learning.
- Digital technology provides rich data on how children interact with learning tools. This data can offer insights to teachers and specialists in language and literacy disorders.
- Interactive and multimedia capabilities of digital platforms provide new affordances for learning materials design.

2.3. Components of Literacy Skill

In order to see how technology can benefit early literacy learning, let us look at the components of literacy ability. This ability is multifaceted and involves knowledge (explicit or implicit) of phonetics, phonology, orthography, morphology, text structure, vocabulary, pragmatics of language use, etc. (Moats, 1999). However, many literacy researchers agree that on a fundamental level, these multiple facets can be divided into those pertaining to the code or written language and those pertaining to linguistic comprehension (Hoover & Gough, 1990). This work focuses on facilitating literacy skills of the first type in a meaningful and grounded, open-ended context.

Aside from a few exotic examples, all current writing systems are to some degree based on speech. The relationship between text and speech is particularly direct in alphabetic and syllabic forms of writing. Correspondingly, both neurological (Wolf, 2008, pp. 145–155) and behavioral (Ehri, 2005) data indicate that written words are linked in human memory to their pronunciations, not directly to meanings. Experienced readers recognize familiar words as a whole (ibid.). But prior to achieving this familiarity, and in order to deal with novel words, readers must internalize the rules of how elements of the written code are linked to sounds (ibid.). This path begins with recognizing the sound structure within spoken words (not yet involving their written form) - a skill called phonological awareness (PA). This skill is not trivial, since sounds within words are coarticulated and not easily separable from one another. Another piece of the puzzle is knowledge of how sounds correspond to letters or letter combinations (called graphemes). In languages such as English, pronunciation of graphemes is highly context-dependent, so learners have to acquire larger and larger "sight chunks", such as morphemes (ibid.). They also have to utilize analogical reasoning to infer pronunciation of a novel word from similar words (ibid.). The learner should also be able to perform the routine of blending: combining separate sounds into a word.

Reading places demands on both long-term and short-term memory. Text comprehension is severely impeded if brain resources are consumed entirely with retrieval of associations between letter patterns and sound, so this process needs to become rapid and automatic. While in languages with complex orthography (such as English) early reading performance is predicted most strongly by phonological awareness, in languages with more consistent orthography (such as Finnish) performance is predicted most strongly by the rapid automatic naming skill (Georgiou, Parrila, & Papadopoulos, 2008). On the short-term memory side, the demand is dictated by the need to keep all decoded sounds of a word in memory before blending them together (Beck & Beck, 2013).

Another important skill involved in literacy learning (and likely in any kind of learning) is cognitive regulation (Wolf, 2008). The present work employs one conceptualization of such skills, called executive function, which refers to the child's capacity to maintain focused attention, inhibit impulses and switch between tasks (Wright & Diamond, 2014). Segers et al. (2016) show that executive functioning is essential in early reading development. Furthermore, Davidse et al. (2011) show that executive function moderates how much children benefit from book exposure at a young age. In a particularly relevant study, Kegel et al. (2009) show that self-regulation skills can be a determinant of whether or not five-year-old children benefit from their play with early literacy software. It is important to note that there is a correlation between socioeconomic status and executive function (Lawson et al., 2018).

The lack of a consistent correspondence between letters and phonemes in English is a major obstacle for literacy learners (Seymour, Aro, & Erskine, 2003). There were at least eight different attempts to circumvent this problem by introducing a simplified orthography as a temporary learning step (Sandel, 1998). One remarkable example is the initial teaching alphabet (ITA), which devised a symbol for every phoneme while trying to stay as visually similar to English as possible. Early experiments with ITA showed that children exhibited a variety of learning gains while using the

tool and were able to transition to conventional orthography relatively easily afterwards (Sandel, 1998). However, after a brief boom in the sixties, ITA failed to gain prominence (Richgels, 2001). Possible reasons for this failure might include the reluctance of educators to use such an unconventional tool and motivational challenges resulting from mismatch between ITA and texts normally found in the child's environment. It has also been argued that most of the advantages of ITA can be retained by simply coloring letters in accordance to underlying phonemes (J. K. Jones, 1968). In a vein similar to ITA, Falbel (Falbel, 1985) used one-to-one correspondence between phonemes and representative graphemes in a digital system for writing (discussed in more detail in the section 2.6). This led to spellings like TOKING BLAWKS for "talking blocks". Other researchers tried to reinforce the connection between graphemes and sounds by integrating pictorial mnemonics into the shapes of graphemes, so that they would remind the child of the sounds. Some notable works, such as Dekodiphukan (Baratta-Lorton, 1985), Lively Letters[2], Reading Genie [3] and Leapfrog (Smith, 2003), utilize the onomatopoeic principle, while other works utilize the rebus principle. Several studies showed a positive effect of rebus-principle mnemonics (De Graaff, Verhoeven, Bosman, & Hasselman, 2007; DiLorenzo, Rody, Bucholz, & Brady, 2011; Ehri, Deffner, & Wilce, 1984; Roberts & Sadler, 2019; Shmidman & Ehri, 2010), but unfortunately, I was unable to find any studies examining onomatopoeic mnemonics. In addition, to my knowledge, there are no studies examining the use of such mnemonics in animated form, in a digital environment or in an expressive medium. One contribution of the present work is to address these gaps.

A note needs to be made on so-called "reading wars". In the past, a fierce debate about how children should be taught literacy (in English) occurred between proponents of two schools of thought: phonics and whole-language (Castles et al., 2018). Phonics emphasizes explicit teaching of phoneme-to-grapheme correspondence, whereas the whole-language approach emphasizes immersing the child into a literacy-rich environment and allowing the child to discover its principles on his/her own. Among reading researchers, the "reading wars" are currently over: it has been shown that explicit and systematic phonics instruction is highly advantageous for breaking the code of written language. At the same time, it has been acknowledged that literacy is not limited to the written code, and for such areas as reading comprehension, immersion into a literacy-rich environment is likely the best approach (Wolf, 2008, pp. 145–155). However, these findings have made slow progress into policy and practice, keeping the "reading wars" ongoing in these domains. Note that the present approach focuses on the relationship between letter and sound patterns. Its scaffolded version, where the grapheme-phoneme association is presented explicitly, can be viewed as a variation of phonics. A potential advantage of the current approach is that it automatically delivers phonics-based learning to the child, without having to rely on appropriate training of teachers.

The discourse in the field of early literacy research appears to be dominated by the study of reading, with much less attention given to early writing. However, reading and writing are interconnected: they both rely on phonological awareness and knowledge of phoneme-grapheme

[2] https://www.readingwithtlc.com/lively-letters/
[3] http://wp.auburn.edu/rdggenie/

correspondence. Some practitioners, such as Montessori, suggested writing should be the first gateway into early literacy. Montessori argued that writing is, in fact, cognitively simpler, as it doesn't require the child to process the perspective of another person. Other researchers highlight more advantages of the writing practice. Writing taps into children's desire to express themselves, which is common in young children, and is the key to the popularity of such toys as LEGO. It is noted that children often use writing for aesthetic purposes (Strickland & Morrow, 1989). Writing provides the child with agency and may give them a sense of acquiring one of the "special powers" of adults. For instance, Nancy Pfrang (ibid.) suggests that teachers and children can write invitations encouraging various people to come to the classroom and tell stories. Responses to these invitations provide children with visible evidence of the power of the written word. Sulzby, Teale and Kamberelis (ibid.) note that children use writing as a sign of their power and a way to leave a lasting mark in the world. For example, writing one's own name is an important way to claim ownership of things. Bissex (Bissex, 1980) makes the same observation in the case of her own child who was an early spellier. She also describes her child using writing to make signs and regulations for adults, for instance: DO NAT DSTRB GNYS AT WRK ("do not disturb genius at work"). The practice of writing can also more easily connect to the children's interests and their lives by giving initiative to the child. Kelly and Safford (2009) provide a wonderful example of how a topic that students are passionate about can motivate them to venture to the limit of their literacy skills. According to Bodrova & Leong (2006), writing supports the development of children's thinking, by acting as an external mediator. Finally, writing with invented spelling doesn't require mastery of sophisticated letter-to-sound rules.

Practitioners argue that it is counter-productive to prematurely aim for the correct orthography while teaching writing (Strickland & Morrow, 1989). Moreover, the ability to correctly spell a few words might be misleading: children can memorize exactly how these words look, with no generalization to other words at all (ibid.). On the other hand, early on, children often invent their own ways to spell words. These invented spellings are fascinating because of their internal logic, which is often much more elegant and straightforward than conventional spelling. Invented spelling builds upon the development of phonological awareness (Read, 1971), and encouraging invented spelling has been shown to efficiently support the development of this skill (Richgels, 2001). Recently, researchers applied Vygotskian scaffolding (see section 2.5) to invented spelling and claimed better results than with traditional phonics instruction (Ouellette, Sénéchal, & Haley, 2013). Practitioners often express concern that encouraging invented spelling leads to children internalizing wrong spellings for words. However, studies show that while invented spellers do indeed initially make slightly more spelling errors, they quickly converge on conventional spelling, and the benefits of the practice clearly outweigh its disadvantages (Richgels, 2001).

2.4. Play

Many of the aforementioned advantages of writing practice can be summarized in a short statement: it is well-suited for expressive play. The term "playful" applies to a great range of behaviors in both humans and animals (Huizinga, 1949; Rubin, Fein, & Vandenberg, 1983;

Sutton-Smith, 2009). These behaviors have much in common: they tend to be (1) voluntary, (2) without any immediate practical purpose, (3) internally structured, (4) temporally and spatially separated from the flow of ordinary life, and (5) involving an element of tension (Huizinga, 1949). Huizinga (1949) says that “play is stepping out of common reality”. The evolutionary function of such stepping out is not entirely understood (Sutton-Smith, 2009). Sutton-Smith (2009) proposes a fascinating hypothesis that play creates a state of potentiality and serves a function analogous to genetic variability in natural selection. In this view, play is an exploration of possibilities related both to the internal and the external world, without committing to them. As such, play is associated with discovery - both on the scale of civilization and on the scale of an individual. Play is an important component of learning. Researchers particularly stress the importance of play in early childhood (Hirsh-Pasek et al., 2009). Unfortunately, many kindergarten and preschool environments around the world today, including those in the U.S., are play-deprived (Hirsh-Pasek et al., 2009; Resnick, 2017).

There is a widespread perception that learning can be made more playful by means of gamification: introduction of video-game-like virtual rewards. Such researchers as Chiong and Shuler (2010) identified such rewards as an important factor in the engagement of children with educational mobile apps. However, if we view play as exploration, then gamification doesn’t necessarily support play (Resnick, 2017). There is a large body of research indicating that external rewards are detrimental to exploration and creativity (for an overview, see A. S. Lillard (2016), pp. 152–172). I avoid them in the current approach. However, Gee’s (2007) analysis of video games highlights other aspects of them which are highly relevant to learning designs. One such aspect is learning in the process of doing, albeit in a simplified and forgiving context - similar to the notion of scaffolding. Complementary to that is the balance of exploration and instruction. The third aspect is right-on-time instruction - delivering instruction exactly when it is needed. Striving to achieve this principle led me to work on built-in, automated scaffolding for the present work. The fourth is the notion of projective identity - forging the child’s positive attitude to literacy by allowing him/her to imagine her/himself as an author.

Connections between play and expression are particularly prominent in the learning paradigm called constructionism. Building on Piaget’s ideas that children construct their knowledge, Papert (1980) proposed a methodology of learning in which mental construction is facilitated by literal construction: by building artifacts. Papert found computers a perfect medium for this methodology: because of their incredible flexibility, they can match any interests of children. A student of Papert, Resnick (2007) defines play as “constantly experimenting, taking risks, trying new things, testing the boundaries”. The outcomes of play are projects, driven by children's passions and intended to be shared with peers - what Resnick defines as “four P’s of creative learning”. The philosophy of four P’s has been implemented in remarkable digital learning systems, such as Scratch, used by millions of children around the world. However, it hasn’t seen much application to literacy learning thus far.

2.5. The Interplay Between Agency and Structure

Constructionist learning environments are notable for providing children with a strong sense of agency. At the same time, they typically are not highly structured. Indeed, structure and agency are often viewed in opposition (Brennan, 2013). However, Brennan (2013) argues that structure can enable agency. This type of structure can take multiple forms. It can manifest itself as rules of a game to which players voluntarily subscribe in order for the game to take place. It can also manifest itself as a resource that supports the learner, like training wheels on a bike.

The latter form is related to the notion of scaffolding (Wood et al., 1976), drawn from Vygotsky's theory of socially grounded learning and development (Vygotsky, 1978). Vygotsky's theory of learning emphasizes the social, not individual, nature of this process. He posits that children learn primarily within the Zone of Proximal Development (ZPD): a range of skills that they cannot yet do independently, but can do with the support of more knowledgeable others. There are beautiful applications of Vygotskian principles to literacy learning, such as Vivian Paley's storytelling/story acting curriculum (Nicolopoulou et. al., 2009). In constructionism, methods for facilitating social learning have been studied extensively (Resnick, 2017). Additionally, Vygotskian ideas in constructionism can be seen in microworlds (Tsur and Rusk, 2018) - simplified and restricted interest-based environments, reducing the difficulty of play while allowing for creative, child-driven expression.

Scaffolding itself is a custom application of Vygotskian ideas to (typically) well-structured learned domains (Wood & Wood, 1996). The metaphor comes from the notion of physical scaffolding, which can support a structure during its construction, and is later removed. Scaffolding, and related concepts, are based on a few key principles: (1) the scaffolder providing support in the context of the learner working on a problem; (2) the scaffolder simplifying the task for a learner when necessary; (3) the scaffolding fading when the learner's skills develop (ibid.)

What functions do scaffolders need to fulfill in order to be efficient at their task? Wood et al. analyzed this question in their 1976 paper (Wood et al., 1976), which introduced the notion of scaffolding. Their work looked at interactions between a child and an adult facilitator in the context of a task with a custom-designed physical material. Relevant to the context of this work, they performed their analysis with the purpose of identifying whether and how these functions can be automated. They highlight the following functions of a scaffolder:

- Recruitment: drawing children's interest and attention to the learning task;
- Reduction in degrees of freedom: reducing the scale of the task so that it is manageable for the learner;
- Direction maintenance: helping learners to stay on task and preventing them from digressing to other aims;

- Marking crucial features;
- Frustration management;
- Modeling.

We can see that scaffolder functions are both cognitive and relational. It is currently difficult to see viable alternatives to a human for the relational aspects of the scaffolder's job. However, some of the scaffolder's functions can be delegated to materials and technology. For instance, the Tools of the Mind curriculum (Bodrova & Leong, 2006) emphasizes the use of external materials (e.g. little marker tokens) by the children as tools to augment their thinking and to help develop it. To some extent, these tools can also be used to facilitate cognitive and emotional regulation. There is also a flourishing field of automated "intelligent tutors" that scaffold children's learning within their ZPD. One of the longest-running research projects in this area is the ACT architecture (Adaptive Control of Thought; Anderson et al. (1995); Anderson & Gluck (2001)). It served as a basis for cognitive tutors in such areas as LISP programming, algebra and geometry. ACT is particularly interesting because of its origins in cognitive science research, and corresponding insights into the human learning processes. ACT-based tutors maintain an internal model of the learners' knowledge and use it to interpret learners' actions - much like a teacher in Piagetian framework would do. Some of the key principles of ACT-based tutors are very relevant in context of the present work: (1) providing instruction in a problem-solving context; (2) immediately responding to the learner's errors; (3) providing the child with simpler tasks when s/he is struggling and "fading" when the child becomes competent; (4) minimizing working memory load (Anderson et al., 1995). Other notable automatic tutors strive to support not only the domain knowledge, but also skills needed for self-regulated learning (e.g. Jones & Castellano, 2018).

There is a variety of tutoring systems designed to scaffold students' reading and writing skills (Jacovina & McNamara, 2017). Interestingly, some of these systems are aimed at empowering humans within the system: e.g. by providing data for the teachers, or facilitating peer-to-peer feedback. While most of these systems are designed for older students, some tutoring systems are applicable for our age range. For instance, Gordon and Breazeal (2015) developed a system that accurately estimates the current reading level of the child, and the words s/he is having difficulties with, in order to scaffold reading. In a study very relevant to this work, Kegel & Bus (2012) examine the effect of a built-in, automated tutoring component in a digital early literacy game. The tutoring system they studied had three levels of responses: (1) repeating the instruction upon a mistake, (2) providing more hints upon the second mistake, (3) revealing the correct answer upon the third mistake. Kegel and Bus find that such tutoring was crucial in five-year-old children benefiting from the game. The effect of the automatic tutor was particularly positive for children with low regulatory skills. The authors conclude that built-in tutoring is essential for early literacy software.

The downside of most existing tutoring systems is that they are tailored to closed-ended tasks, leaving the child with little agency. The approach explored in this work is different in its attempt to

combine scaffolding of the student with allowing him/her to be in charge of what they build. However, a current limitation of my scaffolding systems is that they only allow for a very slight adaptation to the learner's skill level. Improving adaptability of such scaffolding remains an important subject for future work.

2.6. Previous Expressive Media for Early Literacy Learning

Although multiple sources point to potential advantages of expressive media for early literacy learning, not many existing designs follow this paradigm, or come close to it. A good example of non-digital technology is Montessori's Moveable Alphabet (P. P. Lillard, 1972). The Alphabet is simply a collection of letters (with different colors for vowels and consonants) that children can arrange into whatever words they like, either on their own or with the help of a teacher. To help the child, Montessori teachers use various scaffolding methods. They can pronounce the word slowly while emphasizing its sounds, direct the child's attention to the shape that lips take while pronouncing each sound, remind them of another word that contains the sound, suggest a known word that sounds similar, or invoke a mnemonics that connect the sound to the shape of the letter (e.g. point at round lips while pronouncing the sound o). This scaffolding is essential for Moveable Alphabet to work as a vehicle of learning.

One of the earliest digital examples of an expressive medium for literacy learning is Moore's Talking Typewriter (Moore, 1966). As the name suggests, it was a typewriter that could pronounce letters and words typed on it (although, in contrast with SpeechBlocks, it only handled real words). To enable this capability, a sophisticated (for its time) computer system supported the typewriter's operation. The machine was not standalone, but a part of a responsive environment, designed to allow a child: (a) be intrinsically motivated, (b) explore freely, (c) receive immediate feedback, (d) determine the pace of events her/himself, (e) discover various relations and be facilitated in that by the structure of the environment. This environment provided meaningful context for children's literacy activities (e.g. a "class newspaper"), as well as indirect scaffolding by adult facilitators who controlled the equipment. These facilitators could, for instance, constrain the keyboard so that the child could only press the keys needed to type a specific word. They made decisions about increasing or decreasing the level of structure in the environment based on the responses of the child. Moore considered his technology to be aimed at "exceptional" children: either far behind or far ahead of the normal developmental curve. A fascinating question is whether the same pedagogy is well suited to normally developing children as well.

An early block-based digital expressive literacy medium is Falbel's Talking Blocks (Falbel, 1985). Falbel, a student of Seymour Papert, developed this software as a reaction to instructionist computer-based early literacy systems. In his design, children can assemble words on the screen by dragging blocks out of their slots along the screen's edges. Interestingly, they can control the system's prosody by varying intervals between the blocks and their vertical position, adding an additional layer of expression. Another interesting feature is the use of phonemes, not letters, as

blocks. My own exploration of this area is described in section 3.2.2, with evaluation appearing in section 6.6. Unlike mine, Falbel's designs used fixed graphemes to denote phonemes (e.g. AW to denote *[ɑ]*), leading to unconventional spellings like TOKING BLAWKS for *talking blocks*. Most words were in fact impossible to spell conventionally in the system. This design choice highlights Falbel's focus on phonological, rather than orthographic, development. The system was intended for adolescents with delays in literacy development.

Most early literacy apps provide children with rigid tasks that have predetermined right and wrong answers. For instance, two highly popular apps, *Alpha Writer*[4] (which, according to its designers, is grounded in Montessori approach), and *Endless Reader*[5] (a winner of the *App Store Best 2013 Award*), are both based on children dragging letters into the right slots to form a word from a closed vocabulary. However, a few existing apps allow for interesting forms of exploration and tinkering. For example, in *Sesame Street Alphabet Kitchen*, a child can use a cookie-shaped letter to fill in a gap in a word. The system then pronounces the word whether or not it is real. If the word is non-existent, the system makes a funny and encouraging remark, e.g. "BOD - don't know this one, must be some sort of French pastry".

One preceding early literacy app, *Word Wizard*[6], gives children the same power of open-ended tinkering with words as SpeechBlocks I. In this app, children can build any words by arranging blocks on a grid, and the system can pronounce both real and nonsensical words. I was unaware of the Word Wizard at the beginning of my work, but followed a parallel trajectory. Thus, although my early work and this design differ in technical details, such as the method of word formation, they have many conceptual similarities. However, Word Wizard is intended to be used by a child-teacher dyad, whereas I strived to allow the child to play with the app autonomously, possibly at home. As a result, my late work diverged from Word Wizard by introducing various scaffolding mechanisms. Additionally, to my knowledge, there are no studies examining how children interact with Word Wizard, making the present research the first one that looks at how children use open-ended, self-expressive early literacy apps.

In the larger landscape of literacy technology, there are several expressive designs oriented towards older children. Such systems as *CBC4Kids Story Builder* (Antle, 2003), *Mobile Stories* (Fails, Druin, & Guha, 2010), *StoryTime* (Kuhn, Quintana, & Soloway, 2009), and *Arthur's Comic Creator* are designed for storytelling, sometimes in the form of a comic. *Scribblenauts* is an adventure game that uses words to give the player a remarkable degree of control over the game's environment. Typing nouns spawns objects, typing adjectives changes properties of these objects, and typing verbs directs the main character. However, all of these systems assume that their users are already fully capable of writing. Therefore, they are not well suited for helping children to master the basics of literacy.

---

[4] https://montessorium.com/app-guide/alpha-writer-app-guide. Retrieved Oct. 30th, 2020.
[5] https://www.originatorkids.com/?p=40. Retrieved Oct. 30th, 2020.
[6] https://lescapadou.com/LEscapadou_-_Fun_and_Educational_applications_for_iPad_and_IPhone/Word_Wizard_-_Talking_Educational_App_for_iPhone_and_iPad.html. Retrieved Oct. 30th, 2020.

In parallel to the present work, several other related projects were pursued within the same research group. There is mutual influence between the present work and these projects. These projects were oriented towards children who already have some basic literacy skills (typically 6-10 years old) and can be seen as natural next steps after SpeechBlocks.

The first of these media is PictureBlocks (Makini, 2018). This app allows the child to build a collage of small images (sprites) by utilizing his/her literacy skills: each sprite is obtained by typing the corresponding word. A network of semantic associations between words is provided to encourage exploration of vocabulary and provide the child with ideas for building scenes. To enrich the expressive capabilities of the medium, PictureBlocks allows children to record small snippets of their voice and associate them with sprites. The design of the app also includes an explicit social component, allowing children to share their creations with each other via network. The pedagogical intent of the original PictureBlocks was to support vocabulary acquisition. However, I found that multiple elements of PictureBlocks can be repurposed to provide a rich and meaningful context for scaffolded spelling activities. In such capacity, these elements are incorporated into the present work.

The second of these media is StoryBlocks, co-developed by Anneli Hershman, Juliana Nazare and Mark Exposito in collaboration with Sesame Workshop. The intent of StoryBlocks is to facilitate children's socio-emotional development, to encourage their interest in storytelling and to scaffold their narrative skills with the help of a human literacy coach. The coach provides guidance via asynchronous remote connection with the child's app. The coach system was first prototyped with SpeechBlocks, in the studies described in Chapter 5. StoryBlocks allow the child to build a comic strip using several characters, a set of speech bubbles and props. The first version of StoryBlocks is built around the framework of Conflict Stories: the app provides the beginning of a story and then asks the child to resolve the story's conflict. In this respect, the app is an adult-driven expressive medium: the child expresses him/herself, but within a framework that has been set up by adults. In its second iteration, the design of StoryBlocks moved towards the child-driven paradigm. Pre-composed frames were generally removed (although the coach could send the child a frame as a prompt), and elements of PictureBlocks were introduced into the system, making it much more open-ended.

There is a scarcity of information in current literature about the usage of digital expressive media for early literacy learning. Patterns of interaction of normally developing beginner learners with the technology are not documented. Advances in technology, such as mobile devices and progress in such areas as speech and text recognition, are not reflected. Only very rudimentary scaffolding capabilities are considered. Efficacy of the approach for learning foundational literacy skills is not assessed. There is no information about how individual differences of normally developing children affect their experience with the media (for whom such an approach works and for whom it does not). The proposed thesis aims at closing these gaps.

Chapter 3. Design

This chapter describes the design of the two generations of SpeechBlocks, and looks at the considerations that shaped it and the design exploration and play-testing that informed it. The early design, SpeechBlocks I, included features for tinkering with words, but no scaffolding, and offered only limited expressive capabilities. Taking this design as a starting point, I experimented with features for enhancing expression, scaffolding elements, phoneme-based blocks and features for external input. This experimentation culminated in a more sophisticated design, SpeechBlocks II.

3.1. Basic Design: SpeechBlocks I

SpeechBlocks I is an app designed for Android smartphones, which were chosen over tablets due to their availability to large swaths of the population, including families with relatively low socioeconomic status. The interface of SpeechBlocks consists of a single screen divided into three areas (Fig. 3.1). The canvas, which normally occupies almost the entire screen, serves for tinkering with letters and words. On the sides of the canvas, there are retractable word and letter drawers. The word drawer contains a selection of words that can be used as prompts and as material for remixing. It also stores child-generated words, which can be dragged to the word drawer to be saved in it. The letter drawer holds individual letters. The drawers are made retractable to save screen space, which is particularly limited on smartphones.

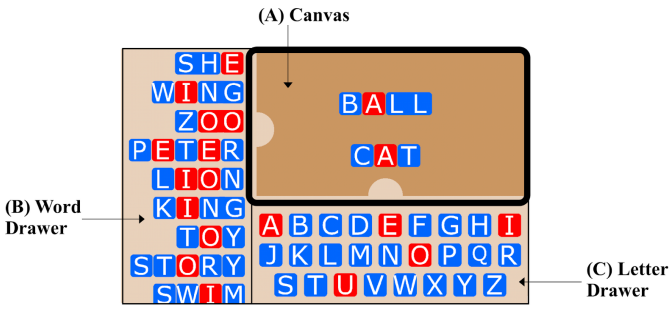


Fig. 3.1.[7] *SpeechBlocks I*

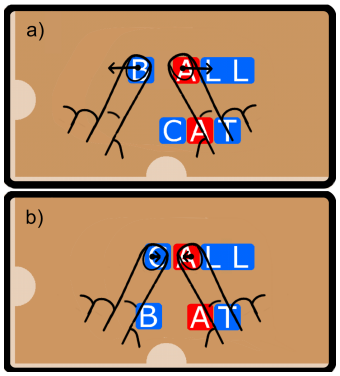


Fig. 3.2. SpeechBlocks I interactions

The blocks in the canvas are connected to each other like LEGO bricks. The user interacts with block combinations through tapping and dragging them with their fingers. Placing two fingers on one block combination and moving them in opposite directions will cause the combination to split into two (Fig. 3.2, a). Putting two combinations end-to-end and “pressing” them against each other will cause them to merge into one (Fig. 3.2, b). Every time a split or merge happens, the speech

[7] Fig. 3.1 And Fig. 3.2 are reused from Sysoev et. al. (2017). https://dl.acm.org/doi/10.1145/3078072.3079720

synthesizer immediately pronounces the outcome of the manipulation. Block combinations are also pronounced when tapped.

This simple mechanics allows the learner to explore a wide range of language regularities in a uniform way. S/he can decompose a word into constituent sounds by pulling it apart, and by merging strings s/he can witness how individual sounds blend together. In the process of play, the learner is exposed to the sounds that word chunks of different sizes make. When a learner strips off a part of a word and replaces it with something else, s/he observes an analogical transfer of pronunciation. However, play-testing experience showed that this method is not ideal for letter-by-letter word construction. There were many cases of letters accidentally snapping to random strings floating around on the canvas, often creating an obstacle for dragging the letter to the target word.

To create a safe space for experimentation, I excluded the notion of "right" and "wrong" from this design: If a child creates nonsense words, they are pronounced as regular words. I also decided to avoid game-like rewards because of the research findings on detrimental effects of rewards for creativity and exploration (A. S. Lillard, 2016).

I used red and blue colors to code consonants and vowels, similarly to Montessori's Moveable Alphabet. Deciding whether to use uppercase or lowercase letters on the blocks was non-trivial. The advantage of lowercase letters is that they dominate words often found in various texts surrounding the child. Therefore, it might be preferable to facilitate learning and reliance on knowledge of these characters. On the other hand, uppercase letters are more visually distinct, while some lowercase letters are mirror images of each other (e.g. "b", "p", "d" and "q"), which is often an issue for early readers. For this reason, several widespread literacy curricula, such as *Handwriting Without Tears*[8], introduce uppercase letters first. Furthermore, uppercase letters can be much more easily positioned on square blocks. Because of these two reasons, I decided to use uppercase.

Designing a proper method for deleting the words took several design iterations. I tried approaches such as flicking the words out of the screen and dragging them back onto the letter keyboard (where they break into letters that fall onto their respective slots). However, I observed that these design choices are associated with high probabilities of children accidentally deleting their words. In the final version of SpeechBlocks I, words are deleted by long-pressing on them, then tapping the cross button that appears.

[8] https://www.hwtears.com/hwt/why-it-works/teaching-order

3.2. Design Explorations: Between SpeechBlocks I and SpeechBlocks II

While initial studies suggested SpeechBlocks I was promising, multiple limitations of the medium became apparent. On the one hand, a vast gap was discovered separating initial tinkering with words from purposeful assembly of specific words. For the assembly of specific words, beginner learners needed either adult assistance (which, as we can see in the section 5.6, was problematic) or supporting materials (which were not open-ended). On the other hand, the medium had limited possibilities for expression, which may have contributed to quick loss of engagement in home studies (section 5.4.2). Design explorations were conducted in an attempt to address both of these issues. I employed play-testing of multiple prototypes to expedite exploration of the vast space of possibilities. Each prototype was tested with 2-5 children, ages 4-7, either at a preschool, in an afterschool program, or at a children's museum, all located in the Boston area. Due to the limited number of participants, these efforts were not very rigorous, and it is possible that some potentially useful design choices were prematurely discarded in this process. Nevertheless, it allowed me to hone in on a set of justifiable choices for the design of SpeechBlocks II.

3.2.1. Design Exploration on Expanding Expressive Capabilities

Early SpeechBlocks only allowed the player to have a few short words on the canvas at a time before s/he ran out of space, making advanced constructs, such as phrases and sentences, difficult to form . These constructs were also impossible to save, unless built as a single word. The search for expanded expressive capabilities went in three directions: (1) simplifying sentence construction, (2) making words trigger some action within SpeechBlocks (specifically, making animated characters speak), and (3) enabling imagery (following the footsteps of another research work, PictureBlocks).

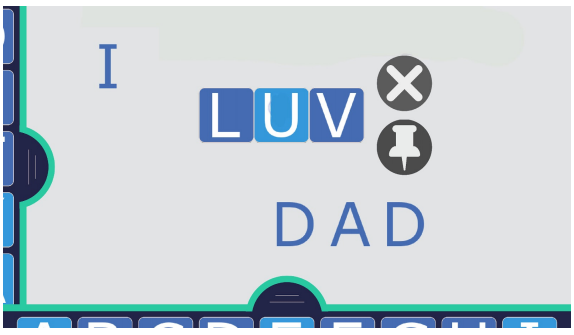


Fig. 3.3. Pinning words

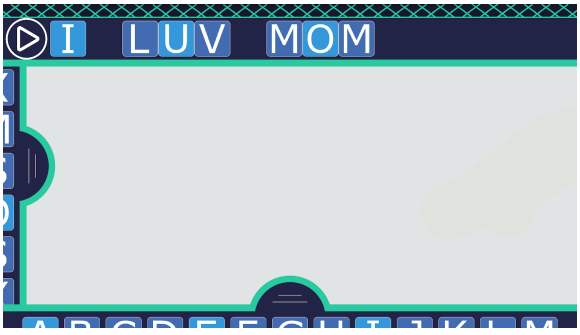


Fig. 3.4. Sentence line

During the first SpeechBlocks study, we noted that children as young as 4 years old sometimes attempted to create sentences, usually by building the entire sentence as one large word. This method led to an inability to complete any but the simplest sentences, as they didn't fit on the app's canvas. Additionally, the synthesizer often failed to pronounce such a construct correctly. To

capitalize on the natural tendency to build sentences,we tried two features. One was to “pin” the completed words to the canvas (Fig. 3.3). Pinned words didn’t interact with words-in-construction and therefore, didn’t clutter the construction space. When the canvas was tapped, the words were read in the order from left to right and from top to bottom. In theory, a flexible 2D-layout of pinned words imitates real-world print sources, such as cards and newspapers, potentially becoming particularly valuable if imagery is added to the app. However, play-testing showed that having both pinned and unpinned words confused children. The second method was to create a dedicated place for building sentences - a sentence line - where completed words could be arranged and rearranged (Fig. 3.4). This approach was somewhat more intuitive for children; however, they still preferred to build sentences as one long word.

Another experiment we tried was introducing animated characters that would read aloud any words that children provided (Fig. 3.5). We assumed children might be compelled to create conversation and interactions between the characters, similar to the PlayWrite system that was developed in the 1990s by Resnick and his colleagues. However, during the playtesting, children continued to interact with SpeechBlocks as usual and largely ignored the characters. A social version of this system was also developed, which allowed the player to send words to another child’s phone. However, I surmised that this capability would mainly become meaningful when children can send sentences to each other. Since the interface for sentence formation wasn’t figured out yet, its further development was postponed.

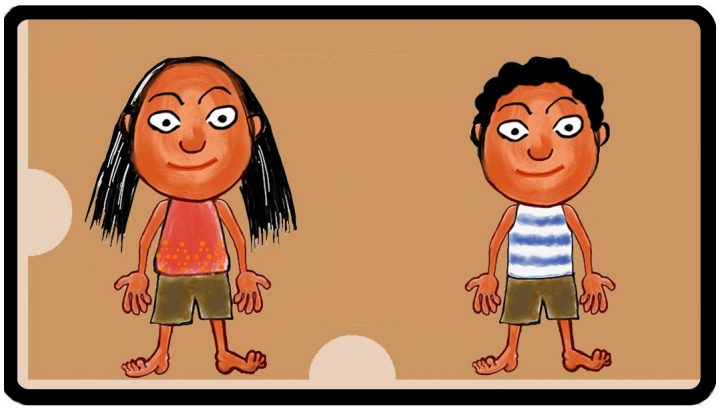


Fig. 3.5. SpeechBlocks with animated characters

In the meantime, Makini’s (2018) PictureBlocks were shown to be quite engaging for children. PictureBlocks (Fig. 3.6) allows the children to build compositions out of sprites which appear when they spell associated words. The app, which was originally conceived of as an extension of SpeechBlocks, was quite compatible with it design-wise. There were additional points of appeal in the PictureBlocks design. First, children in the target age range seem naturally attracted to imagery. Second, the activity of scene building allows the player to achieve meaningful expression through only building a couple of nouns, then lets her/him iterate on his/her creation. This is in contrast to sentence building, which requires pre-planned creation of many words, including adjectives, verbs and auxiliaries. Furthermore, scene building provides many expressive capacities through

non-verbal means. These properties lower the barrier of entry to rich expressive play. For these reasons, I decided to enrich the expressive capabilities of SpeechBlocks by utilizing elements of PictureBlocks design (Fig. 3.7).

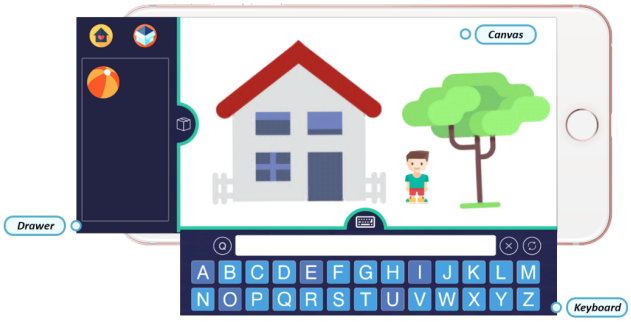


Fig. 3.6. PictureBlocks.

To do so, I imported the library of sprites from PictureBlocks, which contains 1,711 words. Sprites for these words originate from the FlatIcon[9] website, which allows free use of its content for non-commercial purposes. The sprite collection was curated by Makini using custom-designed tools and procedures described in her thesis (Makini, 2018). I added a few more sprites (derived from free sources on the Internet) to the collection. Sprites are arranged on the canvas. Versions of SpeechBlocks that include pictures use the canvas for imagery rather than word construction. Sprites on the canvas can be moved by dragging, and scaled and rotated by pinching. Similarly to PictureBlocks, touching a sprite brings it to the top.

Fig. 3.7. Picture composition in SpeechBlocks

I made two modifications to the design used in PictureBlocks. First, I allowed stacking of sprites. For instance, moving the car on Fig. 3.7 will cause the cat and the hat to move as well, and it would not cause the car to appear over the cat and the hat and obscure them, as it was the case in PictureBlocks. This modification is informed by play-testing sessions, during which I often

[9] https://www.flaticon.com/

observed children doing stacking for different purposes: giving the cat a ride in the car, putting a knight in a castle, seating a person on a sofa, etc. Second, I enabled placing words on the picture canvas along with sprites. The intent of this was to allow children to build imitations of physical print materials, such as books, comic pages and newspapers. Bissex (1980) and Strickland & Morrow (1989) describe a variety of cases of children doing so with a paper medium. However, I did not observe such cases during studies with SpeechBlocks. Words were used for simpler purposes, such as labeling the sprites (see section 6.2.1). Words on the canvas can be scaled via pinching (to allow for uniformity of the interface), but not rotated.

Some elements of PictureBlocks, such as backgrounds, sound recordings and social sharing, were omitted from current SpeechBlocks design because of the time constraints, but could be introduced to enrich children's play.

3.2.2. Design Exploration of Blocks

Early studies with SpeechBlocks led to a question of whether letters should be used as blocks. It seemed reasonable to require a building block to function in the same way regardless of how it is combined with others. Otherwise, the block's function would likely become confusing for children. Since SpeechBlocks was designed to facilitate development of phonological awareness, I wanted children to think in terms of sounds. It therefore made sense for manipulables - blocks - to be associated with sounds. However, in the context of English, letter blocks change their pronunciation, and frequently so, in the process of word construction.

An example from play-testing illustrates this and the resulting confusion. A girl tried to build a word BEAUTIFUL in order to form the sentence MY MOM IS BEAUTIFUL. She correctly identified the letter B for initial sound *[b]* and asked what should go next. I sounded out the word for her, and she correctly identified that the next sound was *[j;u].* Since this sound matches the name of the letter U exactly, she found the letter and attached it to the growing word. The synthesizer said *[b;u]*, and the child was confused because she had expected to hear *[b;j;u]*. Despite this confusion, she decided to continue. The next syllable in the word sounds like the name of the letter T, so she added T. The word became BUT, and so the system pronounced *[b;ʌ;t]*: Note how different this word is phonetically from the beginning of *[b;j;u;t;ɪ;f;ə;l]*. At this point, the child decided that she must have gotten something wrong, so she started to disassemble the word. However, because she didn't know the orthography of BEAUTIFUL, she was unsure how to proceed. I suggested that she add EA between B and U. She did so, and the synthesizer pronounced BEAUT as *[b;oʊ;t]* (analogously to CHATEAU), which sounded even further away from what she wanted, making her even more confused. After I encouraged her to keep sounding out the word, she added E to represent *[ɪ]*, and the synthesizer pronounced *[b;j;u;t].* Now the pronunciation matched the beginning of the desired word, but there was no sound *[ɪ]* that she intended to add. The child considered this development and t added another E, so that the desired *[ɪ]* sound could finally appear. Eventually, she built BEAUTEEFOL, which sounded approximately correct. However, her experience was frustrating and required a lot of intervention on my part due

to the orthographic complexity of the word. If we simply want children to learn how to hear sounds in words better, these orthographic complications are irrelevant.

An alternative to this is to have blocks that directly represent phonemes and whose sounds stay constant regardless of combination with other blocks. However, this raises two questions: (1) how to represent each phoneme visually, so that children can easily look for the block they need, and (2) whether using these representations instead of (or in addition to) the familiar letters essentially doubles the learning needed (first making children learn the symbols for phonemes, and then having them learn how they map to letters). For the latter reason, it was immediately clear that conventional symbols for phonemes (e.g. those from IPA phonetic alphabet) would be impractical. Several designs were play-tested in order to find an alternative that might be easier and more intuitive for children to pick up.

The first design attempt was *Inverse SpeechBlocks*. In SpeechBlocks, each block is associated with a fixed letter and context-dependent sound. In Inverse SpeechBlocks, each block is associated with a fixed sound and context-dependent spelling. Therefore, letters on a block may change as it snaps to other blocks. For instance, when spelling QUEEN left-to-right out of sounds, the spellings for the prefixes *[k]*, *[k;w]*, *[k;w;i]* and *[k;w;i;n]* would be K, KW, QUI and QUEEN respectively. The transitions between spellings are animated to make them more transparent to the player. Machine learning algorithms, similar to those described in section 4.2, were used to account for context dependencies. Inverse SpeechBlocks was only tried with adults: it was soon realized that without any explicit representation of phonemes, it was too hard to keep track of the identity of different blocks.

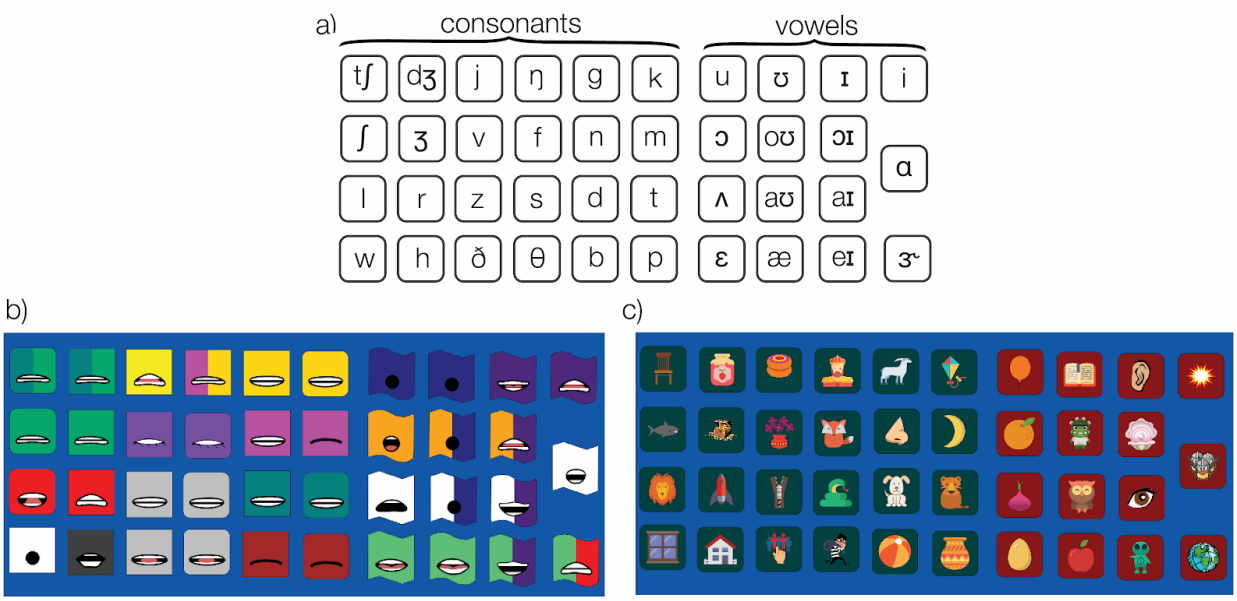


Fig. 3.8. Early phoneme keyboards: (a) layout, (b) viseme representation, and (c) rebus representation

The second design (Fig. 3.8, b) represented phonemes by using a combination of visemes (shapes that the mouth forms while pronouncing these phonemes), block shapes (to distinguish vowels and consonants) and colors (to distinguish classes of related phonemes). Phonemes were arranged on the keyboard based on their perceptual similarity. The third design (Fig. 3.8, c) utilized the rebus principle: each phoneme was represented by an icon of a word starting with this phoneme.

During play-testing of the second and third designs, the children most often either used phoneme blocks chaotically, or treated them as "letters in disguise". For instance, a girl tried to build her name, EVELYN[10]. She said: "It starts with E!", and located a block that made the sound *[i]*, corresponding to the letter's name. I asked her: "Does your name sound like ee-veline (*[i;v;ɛ;l;ɪ;n]*)?", and she said: "No, it's Evelyn (*[ɛ;v;ɛ;l;ɪ;n]*)". I responded: "Then you need sound *[ɛ]*, because these are sounds, not letters". The girl objected: "No, my name starts with E (*[i]*)". After a short discussion about the difference between letters and sounds, I allowed her to proceed. She managed to build *[i;v;i;l;ɪ;n]* and said: "The app reads it wrong!" I said: "This is because your name actually starts with the sound *[ɛ].* Look!" I corrected the name and asked her: "Does it sound right?" She said: "Yes. But my name starts with E".

These observations indicated that the child treated a familiar word logographically, recalling its exact spelling, but not tying it to the pronunciation. However, with less familiar words, children often resorted to phonetic knowledge, and thus used blocks as intended. Logographic and phonetic knowledge were sometimes used in the process of building a single word. Additionally, it was notable that children who used the blocks more purposefully appeared to already possess good phonological awareness, which I surmised was linked to how the blocks were designed. Indeed, to utilize the rebus principle, the child must be capable of recognizing the initial sound of the words with ease, so that this task won't overwhelm his/her mental resources. The viseme principle involves awareness of mouth shapes while talking and linking them to sounds of speech, which is not trivial for children.

I attempted to redesign the blocks to reduce demands on preexisting phonological knowledge. The fourth design used the onomatopoeic principle: each phoneme was represented by an animation that actually produced the sound of the phoneme. For example, *s* was represented by a snake that hissed: *ssssss*. Such a design provides a natural and intuitive association with the phoneme, not requiring it to be isolated from any words. Two versions of this design were tried: letter-based and letter-less. In the first case, the snake was shaped as letters *S* and *C* (Fig. 3.9, a and b): two graphemes typically associated with the phoneme *[s]* (as in SALT and CITY). Although the symbols look different, the underlying snake-based design reveals the common sound. This is somewhat similar to the Color Story Reading design (J. K. Jones, 1968), with the difference that a common animated character, rather than simply the color, is used to signify the phoneme. In the second case, the snake did not resemble any letter (Fig. 3.9, c). An advantage of the first version is that it utilizes children's pre-existing knowledge of letters and is more likely to align with many

[10] This and other children names provided in the dissertation are fictional, to protect privacy of the children

children's ideas of writing activities, which often have an elevated status for children because of writing's association with adults. The second version is beneficial because it avoids potential confusion between letters and phonemes and unambiguously uses a single iconic representation for each sound. Play-testing showed that children are capable of using both versions effectively and can quickly remember associations between the blocks and phonemes. However, the letter-less version was deemed too radical for the upcoming large study, and the letter-based version was chosen for SpeechBlocks II. Since the spelling of a phoneme depends on the context, these blocks behave similarly to the Inverse SpeechBlocks design: their spelling changes based on the context. During these transitions, the creature remains the same, but changes its form. Currently it is done via cross-fading; a more sophisticated hypothetical design could employ animations for morphing between creature forms.

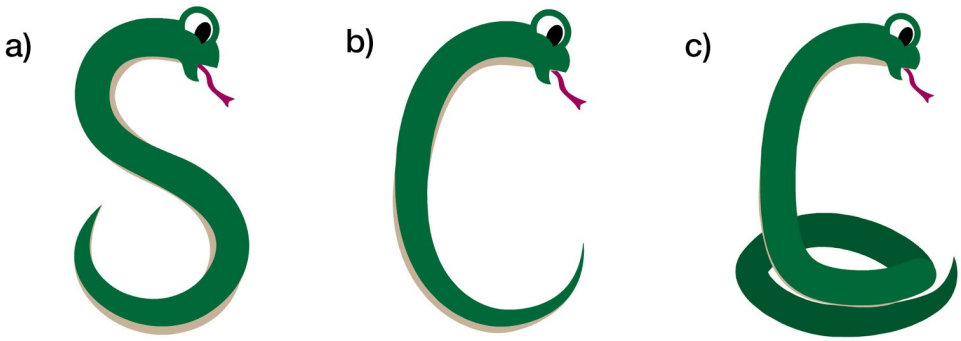


Fig. 3.9. Onomatopoeic representations of phoneme [s] as letters S and C and letter-less

The decision to use graphemes on the phoneme blocks implies that most onomatopoeic mnemonics come in several shapes, like the S and C versions of the mnemonic for *[s]*. To explain their shared identity to children in a more vivid and memorable way, my colleague James Gray suggested presenting the mnemonics as *sound creatures.* Each sound creature was presented as a virtual character who likes to make a particular sound and assume different poses (corresponding to letter shapes). Most creatures were drawn as animals, to make them equally appealing to all genders and races. Each creature was given a name that starts with the corresponding phoneme, adding a small bit of the rebus principle into the design as well. We also hoped that children might relate to the creatures socially, facilitating their better memorization.

For proper functioning of the onomatopoeic mnemonics, it is necessary that the animations are clearly visible to children. A large screen is highly desirable for that purpose, so the designs employing this feature migrated from smartphones to tablets.

3.2.3. Design Exploration of Scaffolding Procedures

During development of the first SpeechBlocks, I hypothesized that children would gradually acquire phonological knowledge by tinkering with nonsense words until they eventually could apply that knowledge to build real words. The experience of the first pilot, described in Chapter 5,

showed that this is not feasible. Building nonsense words quickly started to exhaust itself as the primary activity, and it became apparent that we needed to scaffold the quickest transition to real word building. Such scaffolding was performed largely by the adult facilitators. For the reasons discussed in the introduction, I wanted to incorporate similar procedures into the app itself.

Providing human-like scaffolding can be divided into two subtasks, which are nearly orthogonal: (1) recognizing what the child would like to spell, and (2) guiding the child through the process of spelling. This section focuses on the guidance mechanism. In order to be able to develop the guidance mechanism independently of the input subsystems, I performed play-testing with two input mechanisms: (a) a word bank, where children could select words they would like to spell via icons, and (b) a wizard-of-oz setup, where a researcher listened to children's requests and inputted them into the system. Once a satisfactory guidance mechanism was developed, various input subsystems were connected to it and evaluated.

To facilitate learning of phonology and phoneme-to-grapheme correspondence, the guidance system itself needed to know the phonemes in the target word and their correspondence to graphemes. I developed and implemented mechanisms to support these internal functions, as described in the sections 4.1 and 4.2.

The first version of the guidance system attempted to introduce as few changes as possible into the mechanics of non-guided SpeechBlocks. As the child progressed through building of a word, the scaffolding system pronounced the word with an emphasis on the next sound, and an invisible virtual companion dragged the necessary blocks from the keyboard onto the canvas. If a multi-letter grapheme was needed, several letter blocks were automatically combined. Play-testing of this version indicated, however, that the children were startled and confused by blocks moving on their own.

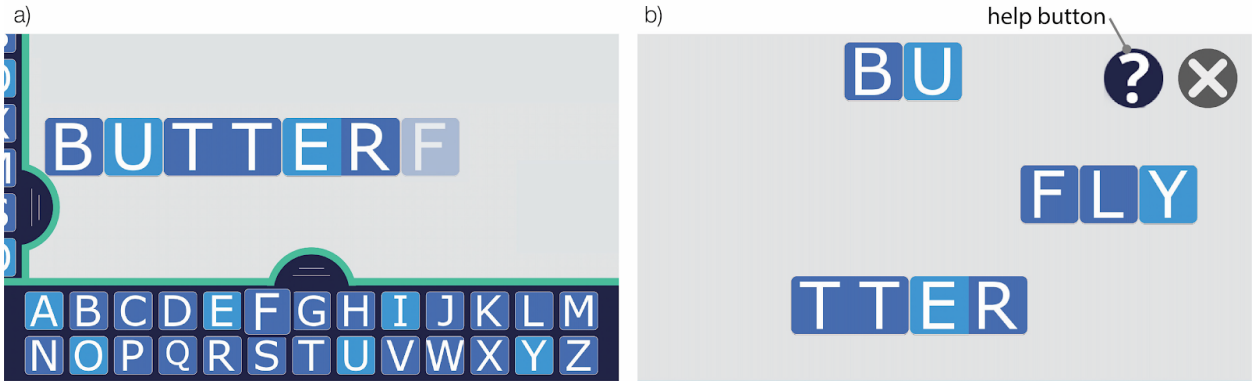


Fig. 3.10. Early guidance routines: (a) sequential, with visual cues, and (b) with scrambled word chunks

The second version of the scaffolding routine relied on visual cues instead of moving blocks (Fig. 3.10, a). It showed the next grapheme to be added as a semi-transparent, "ghostly" block while pronouncing the word with an emphasis on the corresponding phoneme. The corresponding

letters were highlighted on the keyboard. Play-testing showed that this approach was very intuitive and easy for children to use. It was therefore incorporated into the main branch of SpeechBlocks and used in the two home studies, described in the Chapter 5. In this setting, the input for the guidance system was provided by a remote literacy expert, the Family Learning Coach, who sent "suggested words" for the child. Unfortunately, this early scaffolding design allowed for a mode of usage that was not particularly "minds-on": the child could just mechanically drag the highlighted letters to their spots without paying attention to the phonemes that the system pronounced. To avoid this, I decided to exclude visual clues in the next iterations of the scaffolding routine.

The third iteration of the scaffolding routine attempted to provide less scripted, more exploratory interaction, inspired by the feedback provided by Mitchel Resnick. This version scrambled chunks of the target word on the canvas and allowed the child to combine them (Fig. 3.10, b). All irrelevant elements of the interface were temporarily hidden during the process of word building. If the child wished, s/he could press the "help" button, which sounded the word out while highlighting the corresponding chunks (it would say *"bu-tter-fly"* in the current example). If wrong chunks were combined, the system corrected the child by pushing them back apart with a spring-like *'boing'* sound. Two correction modes were tried. In the first mode, the system immediately corrected any mistake, saving the child the potential frustration of going astray for a long period of time. In the second mode, the system gave the child more opportunity to explore and to figure out mistakes on his/her own, correcting any mistakes only after all blocks had been put together. I also experimented with differently sized chunks (syllables, onsets and rimes, and grapheme-phoneme pairs), in order to target different developmental levels. Research shows that as the child's phonological awareness develops, the child is able to recognize progressively smaller parts of the word: first syllables, then onsets and rimes[11], and finally individual phonemes (Wolf, 2008, pp. 145–155).

During play-testing, I saw that children didn't make use of larger chunks (such as syllables). Instead, they pulled them apart into individual graphemes, which they then used to build words. Children naturally exhibited left-to-right, sound-by-sound dynamics of scaffolded word construction. As I mentioned earlier, the mechanics of SpeechBlocks I was not ideal for such dynamics. Furthermore, it didn't allow the scaffolding mechanism to easily correct frequently occurring mistakes such as building the words in a reverse order. Yet another issue with this mechanics was the difficulty of pre-building certain parts of words for the children, such as filling in the vowels in advance and letting children add the consonants. As we will see later, this is a desirable feature considering the pathway of children's phonological knowledge development.

Play-testing showed that immediate correction was preferable to the delayed one - at least for children at an early stage of literacy acquisition. Children were largely unable to discover a mistake on their own, even when the system produced odd-sounding pronunciations. Instead, they treated

[11] The relevance of the onset-rime structure to children's developing phonological awareness have recently been challenged (Geudens & Sandra, 2003). It has been suggested that CV+C structure reflects children's perception more accurately than C+VC (onset-rime). In any case, there are sub-syllabic units that children appear to hear before they are able to hear individual phonemes.

the fact of blocks clicking together as an implicit confirmation that they were on the right path, and they were confused when the system undid their efforts in the end. The observed need for immediate mistake correction was consistent with the recommendation of Anderson et. al. (Anderson et al., 1995) for cognitive tutors design.

Before the introduction of the fourth scaffolding system iteration (Fig. 3.11), the early put-together and pull-apart mechanics was replaced with the *word box*: an area where graphemes could be dropped, simplifying block-by-block word construction. After being dropped into the box, the graphemes automatically slid all the way to the left, and could be rearranged by dragging. In the scaffolded mode, a limited set of blocks was placed next to the word box. It included all phoneme-grapheme pairs present in the word, plus a few distractors. If an incorrect block was dragged into the box, it was immediately thrown out. Similarly to designs (1) and (2), the system provided verbal cues about the next phoneme, although now it provided no direct visual cues. However, adoption of the onomatopoeic mnemonics allowed for introduction of extra cues as references to the mnemonic (e.g. "next goes *[s]* - like Sally the snake hissing"). Similarly to what Kegel and Bus (2012) recommended, the system increased the level of support gradually: at first only saying the desired sound, and after an incorrect attempt employing the reference to the mnemonic. Additionally, the reference could be invoked by pressing on the "help" button. Playtesting showed that this design was straightforward and convenient for children to use.

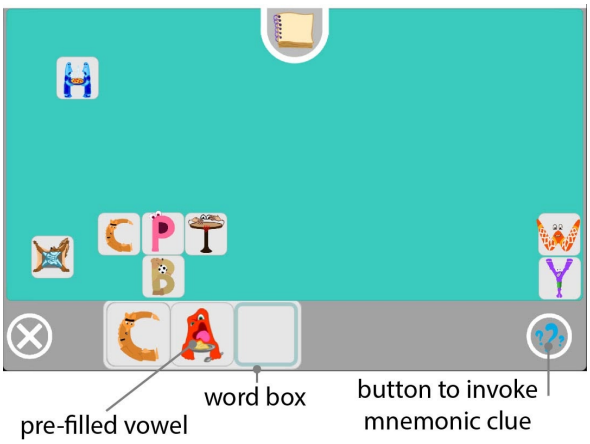


Fig. 3.11. Late guidance routine design

This design was further refined based on observations of children's behavior during the first two days of the SpeechBlocks II classroom study (Chapter 6). These observations confirmed some of the trends that I expected based on literature and previous studies. First, children sometimes dragged the correct block into an incorrect position in the box (e.g. at the beginning of the word, as when attempting to build the word in reverse order). Second, they sometimes dragged an out-of-order block (e.g. the last one) onto a correct position in the word. Third, children had significant trouble recognizing subtle differences between vowels such as *[ʌ]* and *[ɑ]*. Furthermore, I already knew that the difference between the pronunciation used in the child's dialect and the one used by the system could cause problems. For instance, in one of the home studies, a parent

complained that the system pronounced ELEPHANT “incorrectly”. In the parent’s dialect, the word was pronounced as *[ɛ;l;ə;f;ɪ;n;t]*, while the system pronounced it as *[ɛ;l;ə;f;ə;n;t]*. Such dialect differences can cause discrepancies between the child’s choice of next block and what is expected by the scaffolding system, even if the child’s choice is valid within her dialect. These differences are mainly concerned with vowels.

Based on these observations, I made the following modifications. First, if the expected block was dropped anywhere in the box, it drifted to its corrected position. Second, if an out-of-order block was placed in the correct position, it stayed there. Visual slots were added to the word box to highlight target positions. Third, vowel slots were pre-populated, leaving the child with only the consonants to fill. I viewed this practice as a natural step towards invented spelling, where it is not uncommon for children to also use only consonants at first.

An important element of scaffolding methodology is to adapt to the child’s level of skill (Wood & Wood, 1996). All scaffolding designs mentioned above allow it in principle. For instance, in the 4th scaffolding design, difficulty can be varied by providing larger or smaller sets of keys on the scaffolded keyboard or by pre-filling various sets of slots. Active learning techniques, similar to the ones used by Gordon & Breazeal (2015), can facilitate automatic adjustment to the child’s skill level. However, I avoided using such systems with the current version of SpeechBlocks II in order to reduce the risk of introducing too many novel elements at once. Instead, I tuned the system to a common difficulty level that was manageable for most children. Automatic adaptation to the child’s skill level remains the subject of future work.

Fig. 3.12. Invented spelling interpreter

In addition, I experimented with a design where the system recognizes the child’s intent as they build the world: an invented spelling interpreter. The positive impact of invented spelling in children’s literacy development has been discussed in section 2.3. Invented spelling also aligns well with the design philosophy of SpeechBlocks that welcomes nonwords. Unfortunately, many classical examples of invented spelling were not pronounced by SpeechBlocks in the way that children intended. In fact, it is impossible for the speech synthesizer to “correctly” interpret invented spelling, because there is a significant ambiguity in what it may mean. For instance, depending on the situation, the string KT may mean CAT, COAT, KITE or CARROT, or simply be KT if it was built

in the process of construction of a larger real word. The invented spelling interpreter instead makes guesses using an algorithm described in the section 4.3, and displays them to the child as icons, so that the child can choose one (Fig. 3.12). When a guess is selected, and the distance between the input string and the guessed word is low, the input simply morphs into the target word. The intent of this action is not to “correct” the child, but to avoid confusion for others who may attempt to read the word. If the distance to the target word is high, the guidance mechanism described above is invoked to help the child complete the word. I viewed invented spelling interpretation as a potentially powerful way to “boost” the child’s expressive capacities while letting him/her spell at her/his own level.

3.2.4. Design Exploration of Environmental Grounding

While objects in the virtual world of the app can mostly be observed by the player alone, objects in the real world can be seen by anyone. The public nature of physical writing is what gives meaning to many early writing activities, such as the creation of signs (Bissex, 1980; Strickland & Morrow, 1989). Furthermore, children are often naturally interested in environmental text and ask adults to read it to them. These factors motivated attempts to create a bridge between expressive media and the child’s environment.

The first design of this nature was called SpeechStickers (Fig. 3.13). It introduced custom-designed stickers, resembling blocks in SpeechBlocks, to allow children to build words outside the app. The app could read these words via text recognition. Limiting the recognized text to a particular font allowed me to use a very fast real-time algorithm with very low resource requirements, robust to orientation of the text (Torgashov, 2014). However, this algorithm relied on contour detection, which was sensitive to glares and camera shaking. Unfortunately, play-testing revealed that these problems occurred very frequently when the device was in the hands of children.

Fig. 3.13. SpeechStickers

While these issues could perhaps be mitigated with time, the studies with SpeechBlocks revealed the need for scaffolding of word construction. Because performing such scaffolding in the

physical realm was non-trivial, the SpeechStickers project was postponed. The subsequent designs served as an input to the scaffolding system, described in section 3.2.3. The intent was for children to “pick” words of interest from their environment and bring them into their in-app play via reproducing them in SpeechBlocks.

The second design used web-based Google Text Recognition API (Fig. 3.14). Players took pictures of interesting words in the environment. Then, a processing screen showing a robotic avatar of the smartphone reading the words appeared for a few seconds, while the server was processing the request. Finally, a screen with recognized words appeared. Children could hear the words by tapping on them, and then select a word to spell. This approach was applicable to a wide range of fonts and styles of the environmental text, including some handwritten text. It was also more robust to blurs and glares. However, play-testing revealed a few major problems with this approach. First, the large latency between taking the picture and receiving the recognition result didn’t allow children to smoothly explore the environment in search of interesting words. Second, the expected words often didn’t appear at all, either because of the problems with blurs and glares (e.g. children often shook the camera and smudged the image when pressing on the recognition button, or took pictures of texts on very shiny surfaces, such as plastic packages), or because children took pictures of highly unusual fonts (e.g. the logo on a bag of Doritos chips). The laborious usage procedure and multiple failures of recognition discouraged children from using the system.

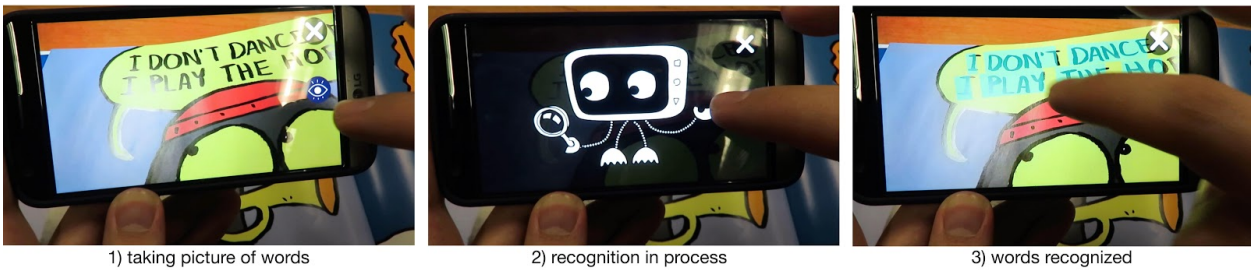


Fig. 3.14. Text recognition using web-based API[12]

I surmised that these problems could be mitigated if a real-time text recognition library was used, removing the waiting time and allowing children to see which words the system could and could not process in advance. Fortunately, two such libraries, one from Google[13] and one from Russian company ABBYY[14], had recently become available for Android devices. The ABBYY library was more robust on typed text and benefitted from integrating recognition results from multiple frames, while the Google library was better at recognizing handwritten text. Anticipating that handwritten text would be of significant importance in the classroom context, I opted for the Google library.

Initial play-testing of this design was discouraging: it appeared that children didn’t take much interest in environmental text and preferred to simply play with plain SpeechBlocks. I deduced that

[12] Book page is from *Waddle! Waddle!* by James Proimos
[13] com.google.android.gms.vision
[14] https://abbyy.technology/en:products:rtrsdk:start Retrieved on April 30th, 2020.

this preference might have been caused by the difficulty of exploring environmental text that children could not read. As a result, they were confronted by a wall of text without any clue as to which words might be of interest to them. In order to find an interesting word, they had to meticulously tap through scanned words. To correct for this, small icons were placed over the imageable words.

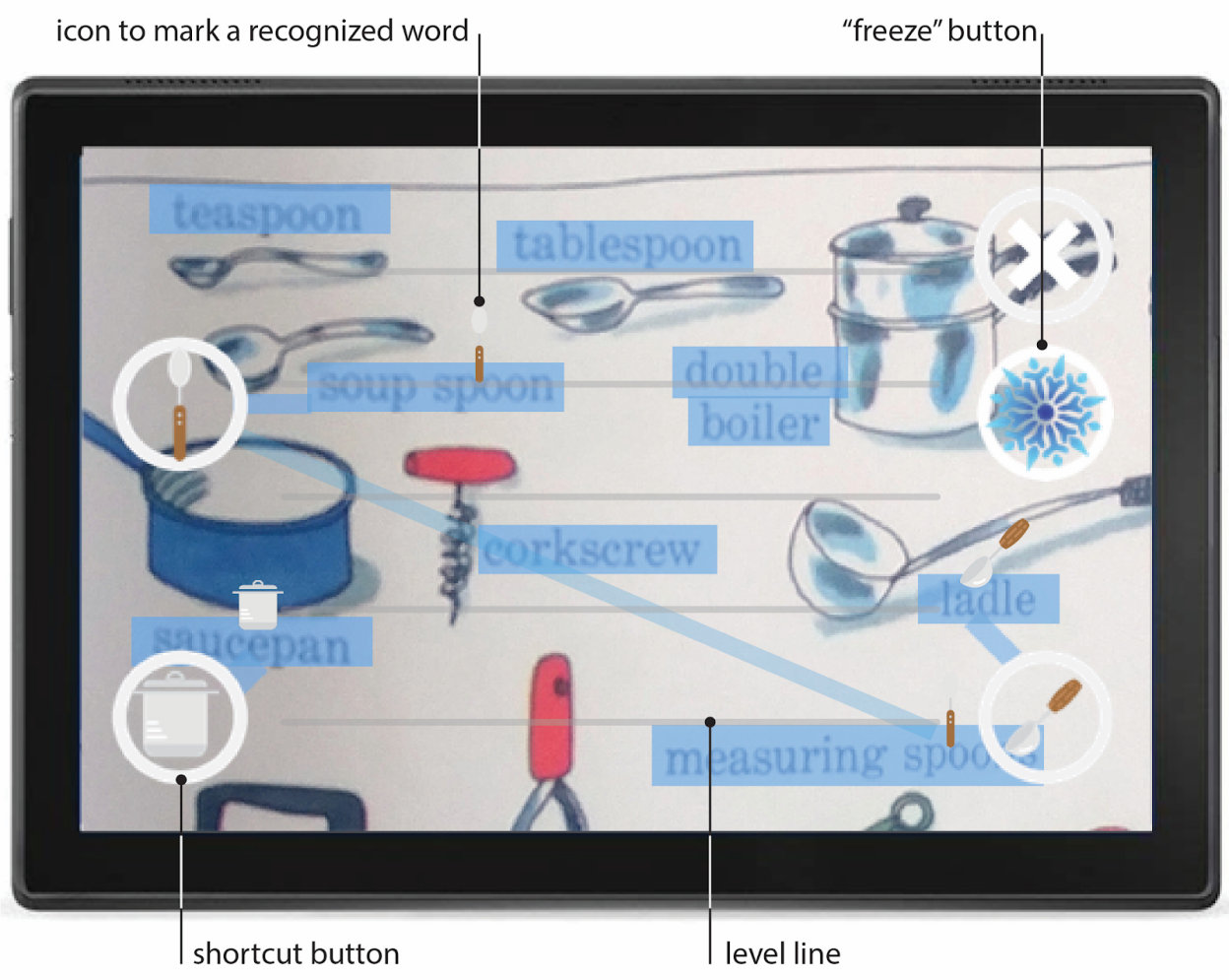


Fig. 3.15. Late text recognition interface[15]

As mentioned earlier, SpeechBlocks II was transferred from smartphones onto tablets in order to accomodate the screen demands of the onomatopoeic mnemonics. This change introduced a challenge for text recognition interface. Because they are relatively heavy and bulky, tablets required children to use both hands to hold them. In this situation, pressing on the words to hear them became very difficult. To combat this issue, I placed a "freeze" button on a side of the screen where it was easily reachable. This button allowed the child to freeze the picture, put the tablet down and explore the recognized words, combining the advantages of the real-time recognition and static-picture-taking approaches. However, as the section 6.2.3 details, this feature was often

[15] Book page is from *Best Word Book Ever* by Richard Scarry

abused by some children. I further modified this interface (Fig. 3.15) by adding shortcut buttons to the sides of the screen, based on classroom observations from the first few days of the Speech Blocks II study. Each of these shortcuts was linked by a thick line to a recognized imageable word, allowing children to select the word without having to freeze the screen first. I also found that children often held the camera sideways relative to the text, creating problems for the text recognition routine. To alleviate this problem, I introduced a set of level lines on the screen, as a reference for children to align the text with. With these modifications, the interface was somewhat successful; the details of its usage by children are presented in the section 6.5.7.

In addition to text recognition, modern computer vision techniques may allow us to use object recognition as an input for the scaffolding system. Object recognition might be more natural to use than text recognition for children at this age, as many of them cannot read yet. Section 6.5.7 provides some evidence in support of this assumption. While object recognition was not implemented in SpeechBlocks, I performed a brief experiment with the Google Cloud Vision API (Fall 2019 version), which showed that existing object recognition systems may already meet the system's demands. I tested the API on a series of photographs, as well as on images from children's books. The latter were included, because I occasionally observed children trying to scan images from these books via the text recognition interface during the SpeechBlocks II study (see section 6.5.7). For instance, one of the images depicted a cartoon character, "Rubble", from *Paw Patrol*. When I uploaded this image, the vision classifier yielded a series of roughly correct, but not very useful labels: *toy, cartoon, action figure, figurine, animated cartoon, animation, fictional character*. However, when I looked at the *Web Entities* section of the recognition results, I was amazed to find *Rubble* listed. In case of other images, different sections of the output were useful at different times, and sometimes no useful results were retrieved at all. However, it is possible that an object recognition interface can be made workable if it could retrieve multiple candidate results and prioritize them in an intelligent way.

The above-described systems aim to bridge the virtual world of SpeechBlocks with the physical world. However, so far this bridge only goes in one direction: into the app. To complete it, some means of output are necessary, and I considered them as well. At this moment, they remain a design speculation. A pilot with SpeechBlocks I (described in Chapter 5) showed that a simple sheet of paper and a pen could be a very satisfying "output mechanism" for children: they enjoyed using the app as a reference to copy words down on paper. A variation of this approach, not yet tested with children but appearing promising, is placing a thin sheet over the screen (so that the words come through) and tracing the letters. A special interface could be made in SpeechBlocks to facilitate this process by fixing the word in question in the center of the screen and making the image highly contrastive. There is also value in printing, since it permits reproduction of scenes that children make in SpeechBlocks II. I envision that printing can be performed not from SpeechBlocks itself, but through a complimentary interface for teachers, parents or coaches.

3.3. Advanced Design: SpeechBlocks II

The design explorations described above culminated in a highly modified SpeechBlocks design, called SpeechBlocks II (Fig. 3.16). Due to the high screen space demand of onomatopoeic mnemonics, SpeechBlocks II is designed to run on an Android tablet instead of a phone. SpeechBlocks II has two main screens: the keyboard screen and the canvas The keyboard screen has a keyboard with blocks and a word box for arranging the blocks. Each word in the word box has a handle for dragging (to distinguish dragging a word from dragging individual blocks). When a child taps on the handle, the word is pronounced: first phoneme-by-phoneme, with corresponding sound creature animations playing; then as a whole. Next to the word box, there is a slider holding invocation buttons for some of the scaffolding modes (namely, Word Bank, text recognition and speech recognition). When an imageable word is constructed, the corresponding sprite appears next to the word box. Both sprites and words can be dragged to the canvas where they can be assembled into compositions, like in PictureBlocks. The canvas has multiple pages that can be flipped: this way, children can save their creations and use fresh pages for new ones. Words and sprites can be deleted by dragging them onto the broom icon on the canvas.

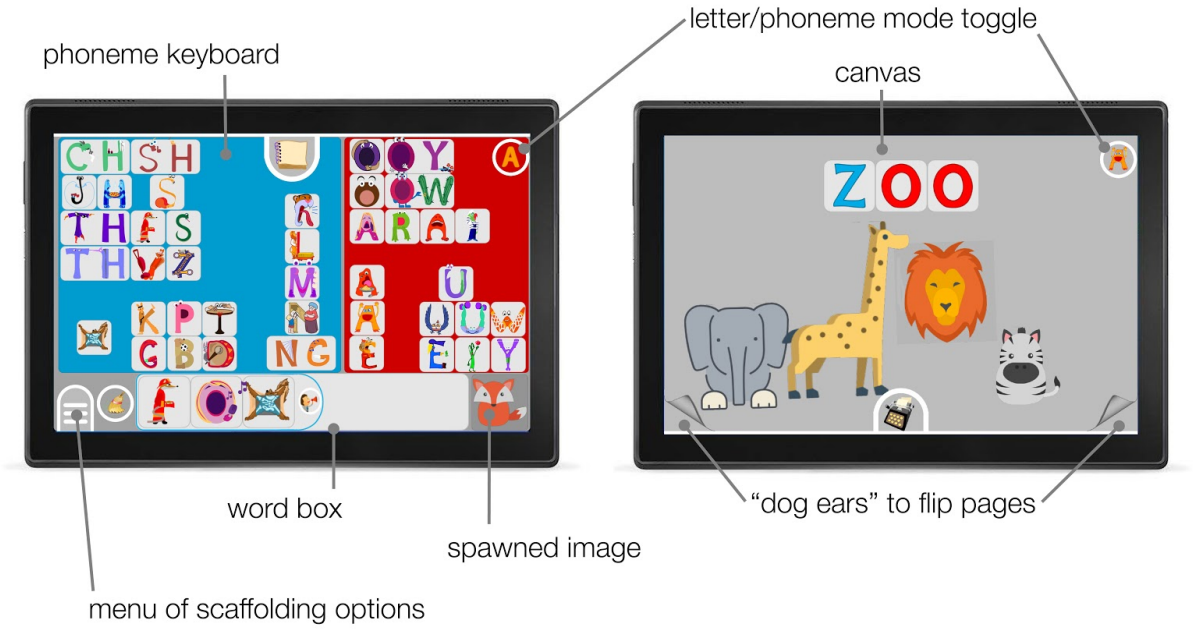


Fig. 3.16. SpeechBlocks II

SpeechBlocks II was designed to let children experiment with both letter and phoneme blocks. Phonemes were represented via the onomatopoeic sound creatures. Because I decided to use graphemes in the design of the sound creatures, I needed to cover most phoneme-grapheme combinations that occur in English. Fortunately, I was able to reuse some of the animations to represent graphemes with double letters and silent letters, simply by placing the letters in question next to the animation made for a shorter grapheme. Barring the graphemes that could be covered in such a way, I identified 82 frequent phoneme-grapheme combinations, using the dictionary of

phoneme-grapheme alignments that was derived via an algorithm described in the section 4.1. I designed the animations for 80 of these combinations and animated about a third of these designs myself, while the rest was covered by student volunteers and by a professional animator. The onomatopoeic analogies for the sounds of phonemes are partially my own, partially derived from four previous phonics programs: Dekodiphukan (Baratta-Lorton, 1985), Lively Letters[16], Reading Genie[17] and Leapfrog (Smith, 2003). 73 of the mnemonics were ready in time to be incorporated into the version of SpeechBlocks II that was tested with children. The catalogue of all sound creatures, along with credit attribution, is given in the Appendix B.

The semantics of the letter and phoneme modes differed at the beginning and end of SpeechBlocks II study. Originally, switching to phoneme mode brought the player to an environment where blocks stood for phonemes. The keys for the 42 phonemes of American English, plus the combination X/*k;s*, were arranged on a special phoneme keyboard, the layout of which (Fig. 3.17) was developed using the neurological data regarding perceptual similarity of phonemes (Mesgarani et al., 2008, 2014). As previously mentioned, phoneme blocks keep their sound, but can change their spellings (via procedurally animated transitions) when arranged in the word box. In the letter mode, the sound creatures were not continually displayed. However, they briefly appeared as the word constructed in the word box was sounded out, to highlight the underlying phonemes. Towards the end of the study, the semantics of the phoneme mode was changed to simply reveal the sound creatures while the blocks and the keyboard still operated in letter mode. The creatures associated with each letter were selected based on the most common phoneme that the letter stands for.

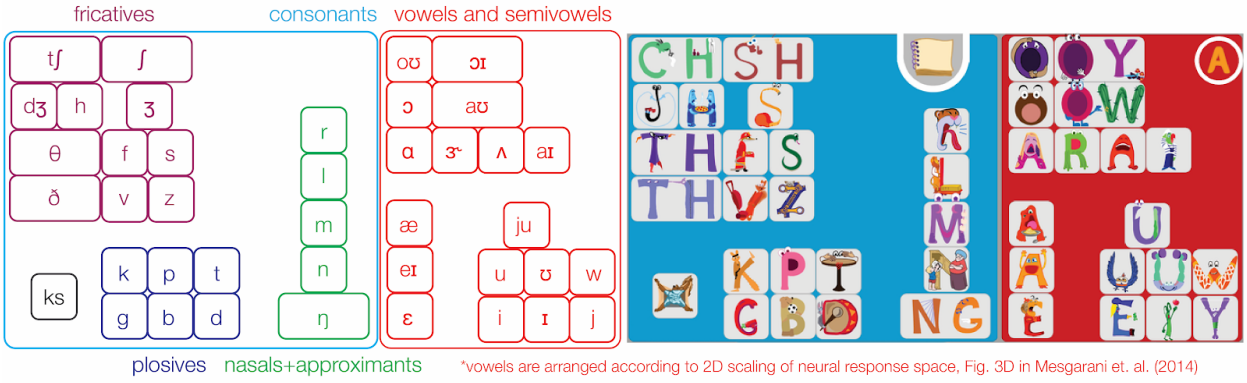


Fig. 3.17. Phoneme keyboard layout

Unlike the letter and phoneme modes, blocks change neither spelling nor pronunciation, regardless of context, in the scaffolding mode. Instead, blocks stay the same as in the target word. Initially, the scaffolding mode always displayed the sound creatures, so that reference to the mnemonics could be employed as a scaffolding hint, as described in the section 3.2.3. In the last few days of the study, it became possible to switch to the letter view.

[16] https://www.readingwithtlc.com/lively-letters/
[17] http://wp.auburn.edu/rdggenie/

To let children become more familiar with the sound creatures, I introduced information pages for each letter (Fig. 3.18, a) and each phoneme (Fig. 3.18, b). Letter pages show various sound creatures associated with the letter. Phoneme pages show various forms of the sound creature. Pages also include sample words using this letter/phoneme.

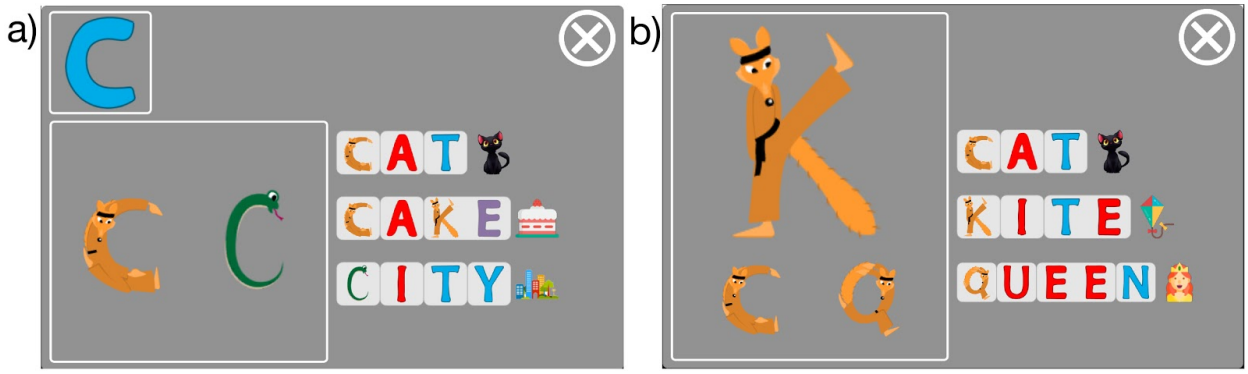


Fig. 3.18. Letter and phoneme pages

There are six ways to invoke the scaffolding routine in the system:

1. By choosing to spell a sample word on a letter / phoneme page (Fig. 3.18);
2. By choosing a word from a word bank (Fig. 3.19);
3. By choosing a semantic association with an image on the canvas (Fig. 3.20);
4. By using invented spelling (Fig. 3.12);
5. By using text recognition (Fig. 3.15);
6. By using speech recognition (Fig. 3.21).

Fig. 3.19. Word bank

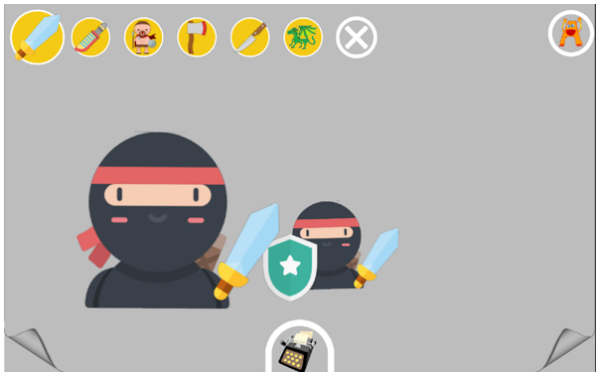


Fig. 3.20. Semantic associations

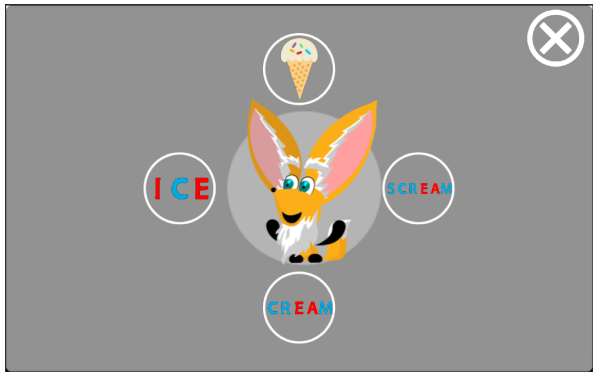


Fig. 3.21. Speech recognition

The text recognition interface was covered in section 3.2.4, while the invented spelling interpreter was covered in section 3.2.3. The word bank has a straightforward design. The top bar allows the child to select a category of words. Icons for words from this category appear on the bottom. Tapping on each icon causes the word to be pronounced, and a button for spelling the word appears over it. Tapping on this button brings the player into the scaffolded mode. I chose the following categories to be represented: *Names, Characters, Animals, Vehicles, Family, Food, House, Street, Fantasy* and *Jobs*. I chose these categories based on the types of words that children were interested in during previous studies and because these categories complement each other. For instance, workers from the *Jobs* category are a useful addition to a landscape created with items from the *Street* category; people from the *Family* category can be placed in an interior created with items from the *Home* category, and can be served items from the *Food* category. The *Names* category contained the names of the children in the classroom where the study was conducted. Because of logistic and privacy concerns, we didn't collect pictures of the children participating in the study, and instead I wrote their names on the buttons. Children typically searched for a name they wanted by tapping on the buttons in order.

The semantic associations system is similar to the one used in PictureBlocks (Makini, 2018). The network of semantic associations was simply imported from that project as a text file. Makini (2018) describes the machine learning algorithm that she originally used to create the network. The interface was also similar to the one used in her work, with a few modifications aimed at (1) making the interface simpler and more transparent in order to adapt it to the younger target population, and (2) transforming it into an input for the direct guidance system. The associations are invoked when the player taps on a sprite on the canvas. The button for the current word is shown larger than the buttons for the associated words. Tapping on any of the associated word buttons makes the corresponding word current: the button enlarges while staying in place, and a new set of associations, relevant to the new current word, surrounds it. This way, the association network can be traversed for an unlimited number of hops. Tapping on an association button also causes the button for spelling the word to appear underneath it; pressing the spelling button brings the system into the scaffolded mode.

Speech recognition allowed me to introduce true open-endedness to the scaffolding system. The speech recognition system is visually represented by an attentive character whom children in the studies called “Mr. Fox”. Mr Fox’s functions are twofold. First, he shows the child the current state of the recognition system - sleeping, listening, processing input, recognition success, and recognition failure (Fig. 3.22). Knowing this state allows the child to properly interact with the system at each moment. Second, Mr. Fox allowed me to introduce a backstory explaining to the child why performance of the speech recognition system is far from perfect. Previous research shows that such a backstory makes children more tolerant of speech recognition flaws, consequently increasing their overall success in using the system (Kory-Westlund & Breazeal, 2019). In this case, the backstory is that Mr. Fox is old and can’t hear very well. To demonstrate his old age, I drew the character with a beard. The giant ears of the character symbolize his function as a listener. I chose Mr. Fox to be an animal rather than a humanlike creature in order to achieve universal appeal to children of all races.

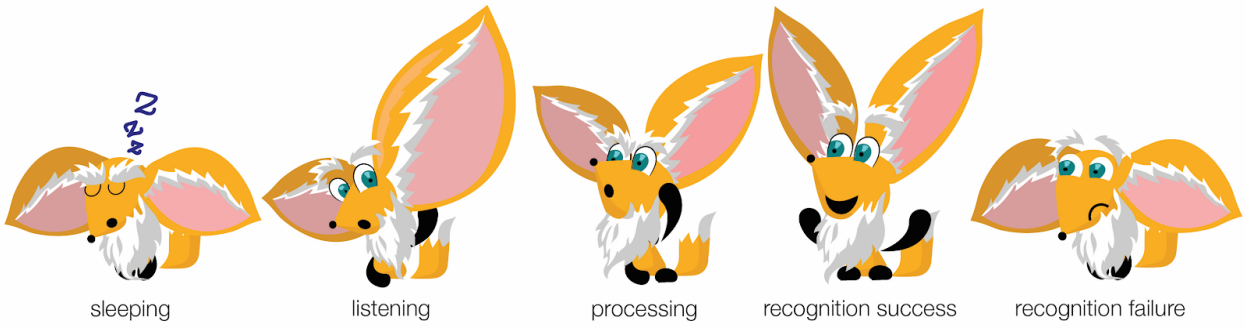


Fig. 3.22. Mr. Fox’s states

Interactions with the speech recognition originally proceeded as follows. The child pressed and held on Mr. Fox to record a snippet of voice. The snippet was then sent to the cloud for recognition; meanwhile Mr. Fox showed a thinking animation. Finally, the results of speech recognition arrived. To counter low performance of speech recognition on children’s voices, and to account for the children’s tendency to speak to the system in sentences (e.g. “Mr. Fox, please give me a GORILLA”), multiple candidate results were displayed around Mr. Fox. Figure 3.21 shows such results for the request “Ice cream”: it includes the words ICECREAM, ICE, CREAM and SCREAM. As usual, by tapping on a candidate result, the child can hear the word and invoke a button for spelling it.

After a few days of using the speech recognition interface in the classroom, I observed that children had difficulty pressing on Mr. Fox while speaking. Sometimes they pressed as intended, but sometimes they simply tapped on the character and then spoke. Both modes of interaction needed to be accounted for. To do this, I introduced voice detection that recognized intervals when children spoke. Voice detection was still triggered by touching Mr. Fox, helping the system to distinguish phrases directed to the character from the conversation between children and background noise. Introducing this modification made children’s interaction with the system significantly easier and contributed to its frequent use.

Chapter 4. Under the Hood: Models and Algorithms

This chapter examines the various models and algorithms needed for SpeechBlocks to function. Because of SpeechBlocks' focus on phonological awareness, the system needs to know how spelling and pronunciation of different words align, and how they can jointly be split into aligned "building blocks." I refer to these building blocks as "atoms". Section 4.1 explains how atomizations are inferred for words known to the app. Section 4.2 describes how the app infers pronunciation or spelling (depending on whether letter or phoneme mode is used) and atomization for unknown words. Sections 4.3 - 4.5 describe the algorithms supporting various modes of scaffolding. Section 4.3 illustrates how children's invented spellings are interpreted before the system shows the child the list of guesses. Section 4.4 describes how the text recognition system keeps track of the same words in multiple video frames. Section 4.5 explains how the speech recognition system post-processes the ambiguous speech recognition output in order to show the child a list of candidate interpretations of what s/he has just said.

4.1. Segmenting Words into Atoms with Aligned Pronunciation and Spelling for Within-Vocabulary Words

SpeechBlocks aims to help children learn to distinguish the sound structure of words. However, the words in the app are spelled with letters. This is the case even when using phoneme blocks, as I made the design decision to put the corresponding graphemes onto the phoneme blocks. Therefore, it is crucial for the app to "know" which letters in a particular word correspond to which sounds. For instance, the word PHONE has five letters, but only three sounds: *[f]*, *[oʊ], [n].* The app needs to "know" that PH corresponds to *[f]*, O corresponds to *[oʊ]* and NE corresponds to *[n]*[18]. This capability is necessary for several reasons.

First, when SpeechBlocks guides the child through construction of PHONE, it would be undesirable for the app to do it letter-by-letter (e.g. "Find P. Now find H..."). Such guidance would tell the child little about the sound structure of the word, as its letters do not directly correspond to its phonemes. Instead, the direct guidance system used in SpeechBlocks provides the child with the blocks corresponding to sounds (PH/*f*, O/*oʊ* and NE/*n*) and guides the child to first look for the block corresponding to *[f]*, then *[oʊ]*, then *[n]*. For similar reasons, "knowledge" of the

[18] Alternatively, if we allow the "building blocks" of the word to be non-continuous, we can say that PH corresponds to *[f]*, O_E corresponds to *[oʊ]*, and N corresponds to *[n]*. The algorithms described in this section deal with continuous blocks, because they map more naturally onto the design of SpeechBlocks. However, a modification is possible to account for non-continuity.

letter-to-sound mapping within each word is also necessary for other forms of scaffolding, such as invented spelling interpretation (for details of its workings, see section 4.3).

Second, both in SpeechBlocks I and SpeechBlocks II, blocks visually reflect underlying sound structures. In SpeechBlocks I, letters T and H are visually merged into a single block in the word BATH, but are separate in the word BOATHOUSE. In SpeechBlocks II, the child also sees the underlying "sound creatures" if s/he switches to the phoneme mode, or when the word is sounded out upon being tapped.

Third, in SpeechBlocks I, "knowledge" of letter-to-sound mapping is also necessary to determine how the words should split. For instance, if the child pulls apart P and O in the word PHONE, where should the split line be? The split PH-ONE would be reflective of the underlying sound structure, while P-HONE would not.

"Knowledge" of letter-sound mappings within each word is also necessary in other scenarios, which could be useful in later versions of SpeechBlocks, or in other learning-oriented software. Such systems as Voice Dream Reader[19], which helps dyslexic people read texts, can benefit from sound-by-sound reading mode. Armed with this "knowledge", software can reason about word similarities, rhymes, or word play (e.g. how to remix words into other words while preserving letter-sound associations). All these capabilities can be beneficial for software aimed at development of phonological awareness.

In this document, I will refer to the "building blocks" of words used by the above-mentioned algorithms as "atoms". Each atom consists of a letter or a sequence of letters matched with a phoneme or a sequence of phonemes. For instance, the initial atom in PHONE consists of letters PH and phoneme *[f]*, while the last atom in FOX consists of letter X and phonemes *[k]* and *[s]*. I denote these atoms as PH/*f* and X/*k;s*. I refer to a breakdown of a word into atoms as *atomization* and denote an atomization by a sequence of atom codes separated by hyphens: e.g. PH/*f*-O/*oʊ*-NE/*n.*

While breaking the word PHONE down into atoms may seem trivial for a speaker of English, it is not trivial for a machine. Moreover, in the case of some words, such breakdown might be non-trivial even for a human. For example, should the word RUSSIA be broken down as R/*r*-U/*ə*-SS/*ʃ*-IA/*ə* or R/*r*-U/*ə*-SSI/*ʃ*-A/*ə*? Should we treat TION in the word NATION as a single "building block": N/*n*- A/*e*-TION/*ʃən*, or should we split it: N/*n*-A/*e*-TI/*ʃ*-O/*ə*-N/*n*? If we split it, should we attach I to TI or to IO? Although the name "atom" implies that they cannot be subdivided any further, we are interested in non-divisibility from a practical, not theoretical, perspective: namely, is it better to present a certain letter-sound pattern to a learner as a "building block" in its own right, or to subdivide it? In fact, from a certain standpoint, it may even make sense to treat common morphemes, like ING, as atoms in SpeechBlocks.

[19] http://www.voicedream.com/

I found no existing dictionary that defines atomizations for a sufficiently large set of English words [20]. The closest to my need was the CMU pronouncing dictionary (Weide, 1998), which establishes a relationship between words and their pronunciations (e.g. PHONE - *[f oʊ n]*), but provides no information on how phonemes and letters within the words are related. I therefore decided to create an algorithm that takes the CMU dictionary as an input and then outputs these relationships for every word in the dictionary. Its output forms a new dictionary that can then be imported into SpeechBlocks to support the internal work of the app.

Because the problem of dividing words into atoms is not well-defined, I was looking for a simple, theoretically justified computational criterion that is grounded in a certain view of what is optimal for the learner. In section 4.1.3, I propose an information-theoretic definition of such a criterion. Using this criterion, an optimization method can derive atomizations of words from vocabulary of any length in an unsupervised manner.

4.1.2. Related Work

The task at hand is closely related to two previously studied problems. The first of these problems examines how the sequences of letters and phonemes belonging to a word can be aligned with each other by inserting empty symbols into the sequences. This formulation of the problem doesn't explicitly involve segmentation. This problem is motivated by the sliding-window design of grapheme-to-phoneme transduction modules within speech synthesizers. In this design, the synthesizer looks at one letter and its neighbourhood at each moment and emits either a phoneme or an empty symbol. Properly aligning phonemes and letters is important to provide good training data for such design. NETtalk (Sejnowski & Rosenberg, 1987), an early experiment in grapheme-to-phoneme transduction, introduces a manually aligned database of approximately twenty thousands words. Black et al. (1998) introduce a semi-automatic approach to do the alignment task, based on a table of allowable matches. Luk & Damper (1992) propose an algorithm for this task based on Dynamic Time Warping, while Damper et al. (2004) propose an algorithm inspired by Expectation-Maximization. This approach is conceptually similar to one of the algorithms explored in this work.

The second related problem explores explicit segmentation of words into graphemes. Lawrence & Kaye (1986) developed an algorithm driven by a large (more than 500) set of expert-defined rules, derived from Walker's Rhyming Dictionary of the English Language and Collins English Dictionary. Ling & Wang (1997) describe an unsupervised learning algorithm driven by a set of four heuristics, each of which targets incremental improvement of performance of a speech synthesizer. Baldwin & Tanaka (2000) compared this approach to another heuristic, based on TF-IDF metrics. Lukeš & Litsas (2015) created an engine specifically designed for phonics guidance. For this engine, they developed a set of grapheme segmentations based on the combination of several heuristics and manual checks. Unfortunately, they only provide

[20] There were several dictionaries that do it for relatively small sets of words, up to a few thousands. They are described in Related Work.

segmentations of about 5000 words, and since their system is not fully automatic, this list is not very easy to extend further.

The present work differs from all the studies listed above in several aspects. First, previous studies have examined how elements of orthography align with *individual* phonemes. In a few cases where a sequence of two or more phonemes maps to a single letter (like *[k;s]* in FOX), these works have manually introduced new pseudo-phonemes representing these phoneme combinations. Contrastively, the present work accounts for the possibility that atoms of written language may include many phonemes, and allows for automatic discovery of such atoms. Second, while the previous studies operate either with expert knowledge or with a set of heuristics, the present work proposes a simple measure grounded in information theory for unsupervised inference of atomizations of words.

In this aspect, my approach is related to various works on unsupervised language acquisition from the field of computational linguistics. These works attempt to model how children acquire the understanding of structure and semantics of language from the raw sensory data available to them. One of the key questions studied in these works is whether some aspects of language can be learned from experience, as opposed to being "hardwired" into our brains. For example, Futrell et al. (2017) present a model that learns phonotactic rules from positive examples only. De Marcken (1996) describes a model that learns to segment raw audio stream into words and word-like units on several levels of hierarchy. Creutz & Lagus (2002) have created two models for unsupervised discovery of morphemes. Roy & Pentland (2002) present a system that learns to identify sensorial referents of words.

Many works on unsupervised language acquisition are based on the Minimal Description Length (MDL) principle. MDL has been used in a variety of works on unsupervised learning, including a model of a general-purpose artificial intelligence agent (Hutter, 2003). MDL is closely related to the Bayesian Occam Razor principle (Tenenbaum et al., 2011) that is widely used in Bayesian models of cognition. Both of these principles provide a natural balance between complexity of the learned model and its fit to data, and therefore control overfitting. De Marcken (1996) explains that MDL is an approximation of the structural risk minimization principle (Vapnik, 1982) that seeks for optimal balance between bias and variance of the learned model. As I show later, the model considered in this section can be thought of as an MDL model.

However, works in unsupervised language acquisition deal primarily with acquisition of *oral* language. Written language may have attracted less interest from researchers in the field because it is most often acquired through instruction. However, research on invented spelling, as well as learning from subtitled televised programs (e.g. Kothari & Bandyopadhyay, 2014), suggests that some children can decipher written language through exposure. This work might serve as a basis for a computational model of such a process. However, in its current form, my model makes assumptions that are not very plausible from the standpoint of human learning. Therefore, adaptations are necessary for it to serve truly as a cognitive model.

4.1.3. Minimum Entropy Approach

Since the goal is to build a set of word atomizations that can aid children in learning to read and write, the criteria for selecting some atomizations over others should be based on learnability of atomic letters-to-sounds patterns in the set. I assume that the set of atoms is most easily learned when it is small and consistent: instances of the same atoms appear over and over again. One possibility is to simply minimize the number of atoms over the vocabulary. However, such an approach doesn't take into account the frequency of atoms' occurrences. I propose a slightly different approach based on minimizing the entropy (Shannon, 1948), which in this case can be thought of as a measure of disorder of the set of atoms.

Speaking specifically, I require each of the atoms to include a non-empty sequence of letters and a non-empty sequence of phonemes. It can be argued that some atoms, such as long vowels, are better represented with disjoint sequences of letters (e.g. A_E/*eɪ* in LATE). For the sake of simplicity, I don't consider such cases in the present work, and I require the letters and phonemes sequences to be consecutive. Each word is composed from a set of atom instances. Then, given either a vocabulary or a corpus of text, one can obtain a probability distribution over occurences of atoms. From this probability distribution, one can calculate the entropy:

$$E = \sum_{a \in A} - p(a) \cdot ln(p(a))$$

where A is the set of atoms. One can change A and associated probabilities by moving the boundaries between atom instances within the words. The goal is to find A that minimizes the entropy.

Observe that minimizing the entropy also minimizes the length of the Shannon-coded description of a vocabulary (or corpus) containing parallel orthographic and phonetic representations. This happens under the constraint that one models the vocabulary/corpus as a sequence of independent atom instances (and therefore, does not model for the transition probabilities between atoms). From the learning standpoint, this constraint makes sense. Our hypothetical learner is likely to first acquire the set of atoms, and only then start to learn the relationships between them. Because of this property, the present method can be placed in the family of MDL approaches. If we agree with the hypothesis that MDL is related to how our cognition structures our experience, then it gives us an additional argument as to why minimizing entropy is relevant to learnability of the atoms set.

4.1.4. Minimizing the Entropy

I used a subset of the CMU pronouncing dictionary as input data to compute atom frequencies for entropy minimization. For computational experiments, I selected only the first 5000 most frequent English words[21], but used a subset of the 50000 most common words for the final library

[21] https://www.wordfrequency.info/intro.asp - based on the Corpus of Contemporary American English (Davies & Gardner, 2013)

to be used in SpeechBlocks. I placed a limit on the size of atoms: they could contain no more than 3 phonemes and no more than 4 letters. While the primary reason for this constraint is computational, it also makes sense from the viewpoint of learnability of individual atoms because it might be hard for a child to memorize a complex combination of letters and sounds.

Entropy is a function of atomization of the entire vocabulary. Its optimization is therefore a combinatorial problem that doesn't seem to be easily decomposable. I have tried several methods to optimize the entropy.

a. Heuristic Method: an EM-like Algorithm

My first approach was to use a heuristic method inspired by the Expectation-Maximization family of algorithms, driven by the following intuition. Imagine that we have a generative model spawning sequences of atom instances that form words. If we knew the parameters of the model then, for a given word, we could have inferred the most likely sequence of atom instances that the model spawned to produce this word. On the other hand, if we knew correct splits of all the words into the sequences of atom instances, we could have inferred the parameters of the model. We arrive at a chicken-and-egg kind of problem. However, this is exactly the kind of problem that Expectation-Maximization algorithms are designed for. We start by assuming that every possible atomization of a word is equally likely. Then, we compute the pseudo-counts of atoms: each atom instance in each atomization contributes the value equal to the likelihood of the atomization to the pseudo-count of that atom. We use these pseudo-counts to estimate various probabilities in the generative model. Finally, we use the generative model to update the likelihoods of the atomizations; now the cycle can repeat. We iterate until convergence, and then take the atomizations with maximum likelihoods as the output.

I tried two generative models in conjunction with this approach. My first model simply emitted a string of atoms according to their unigram probabilities until the stop marker was emitted. However, I found that applying this model led to over-grouping. The model tends to prefer atomizations with a small number of elements, since their likelihoods involve a small number of multiplications of numbers less than one. Therefore, the model treated many word parts and even entire small words (like EAR) as atoms, which led to high resulting entropy. I replaced this model with a more complicated one that treated probabilities of atoms themselves as composite values: a combination of transition probabilities between phonemes within the atom and an emission probability from the set of phonemes to the grapheme. This model was free of the aforementioned problem and worked well in practice.

Since this approach is computationally fast and produces reasonable quality of alignments, I ended up using its output in SpeechBlocks (see section 4.1.8). However, from a theoretical perspective, I was interested in evaluating the entropy minimization principle in general. Since the EM-like approach doesn't minimize the entropy directly, I couldn't guarantee that its final result would be anywhere near the optimal value. Therefore, I implemented several direct entropy minimization approaches for evaluation purposes.

b. Gradient Descent.

This method represents the first attempt to minimize the entropy directly. I attempted to substitute the discrete problem of placing boundaries within the words with a continuous one. To do so, let's suppose that every possible atomization of every word in the vocabulary manifests itself simultaneously with competing atomizations, with a certain "probability". We put the word "probability" in quotes, because here we treat "probabilities" as parameters of the system, rather than actual probability estimates computed from frequencies. Adjusting "probabilities" of atomizations changes frequencies of different atoms. By driving down probabilities of "bad" atomizations and raising probabilities of "good" ones, an optimization algorithm can minimize the entropy. The set of atomizations with maximum "probabilities" is then selected as an output.

For example, there are eight possible ways to atomize the word *box (baks)*: B/*b*-O/*a*-X/*ks,* BO/*b*-X/*aks,* BO/*ba*-X/*ks,* BO/*bak*-X/*s,* B/*b*-OX/*aks,* B/*ba*-OX/*ks,* B/*bak*-OX/*s,* BOX/*baks.* In the current approach, each of them is associated with a parameter $\lambda_{w,i}$, where *w* is the index of the word and *i* is the index of the atomization. The values $\lambda_{w,i}$ are used to define "probabilities" of atomization *i* using the softmax function:

$$p_{w,i} = e^{\lambda_{w,i}} / \sum_j e^{\lambda_{w,i}}$$

Using the chain rule, the derivative of the entropy with respect to $\lambda_{w,i}$ can be computed. Any algorithm for optimizing the value of a differentiable function can then be applied to solve the problem. In particular, I tried gradient descent, gradient descent with momentum, and a variation of gradient descent where at each step we move in the direction of gradient until the minimum along that direction is reached. I found that all of them achieve similar results, and that their entropies are close, though not as optimal, as the results of the expectation-maximization method. They perform worse than expectation-maximization even when random restarts were used. Since expectation-maximization is an approximate method, while the current one optimizes the target function directly, the only explanation for this problem is that gradient descent gets stuck in a local minimum or on a plateau. Therefore, I found it necessary to use non-local optimization methods: genetic algorithms and simulated annealing.

c. Simulated Annealing.

In this approach, I consider a state to be a set of atomizations of every word in the vocabulary. Movement from a state to a neighbouring state occurs using three transition functions:

a. **Split.** Select an atom with at least two letters and at least two phonemes from the set of current atoms. Split both the letter and the phoneme lists within the atom at some positions, such that the lists resulting from the split contain at least one item. Form two new

atoms from the splitted lists. In all words that contain instances of the old atom, replace it with the two new ones.

b. Merge. Select two atoms that are neighbours in at least one word. Make a new atom by merging the phoneme and letter lists of the old two. In all words that contain consecutive instances of the old two atoms, replace them with the new one.

c. Rebalance. Select two atoms that are neighbours in at least one word and that have either at least three letters or at least three phonemes together. Adjust the boundary between the atoms so that some letters or some phonemes move from one atom to another. In all words that contain consecutive instances of the old two atoms, replace them with the new two.

Note that transition functions modify atomizations of many words at once in a similar manner, rather than modifying one word at a time. This process ensures that the algorithm makes progress in any reasonable time. Otherwise, too many random changes are needed for the algorithm to move from one local optimum to another. The synchronous change should not present a problem, since we are looking for consistency in the set of atomizations.

In practice, straightforward implementation of simulated annealing tended to become stuck in the unfruitful regions of the search space for any cooling rate that I tried. However, simulated annealing with reset (jumping back to the current optimal solution if no improvement occurred for a certain time) worked well.

d. Genetic Algorithm.

Genetic algorithms appear to be particularly well-suited for this problem because of relatively weak interdependence between different parts of the solution. Indeed, there seem to be systems of atomizations for groups of words that work well together and have little dependence on systems of atomizations for other groups of words - ones with very different letters-to-sounds patterns. Therefore, the search for an optimal solution can benefit from combining the “good ideas” that emerge in different branches of the search. Genetic algorithms have the capacity to do just this kind of combination.

In my implementation, I consider the genotype to be a set of atomizations of every word in the vocabulary. I define the mutation procedure in exactly the same way as the neighbour state transition procedure in the simulated annealing approach. The crossover procedure can be defined simply as randomly taking atomizations of some words from one genotype and atomizations of the remainder - from another. This procedure works, but it leads to a quite slow convergence, because it is likely to break systems of atomizations that work well together. I found that these systems can be better preserved if I add a post-processing step to the naive crossover described above. Namely, for each word in the vocabulary, I look again at the two corresponding atomizations in the parents’ genomes and choose the one that works better (results in lower overall entropy) with the

other atomizations in the proposal genome. Finally, to define a genetic algorithm, one needs a selection scheme. I found that I cannot directly define a mapping from the entropy of a genotype to its selection likelihood in a non-artificial way. Instead, I used the tournament selection (Miller et al., 1995): I randomly drew a set of genotypes from the population, and selected the best one of them for the crossover.

In my experiments, working on its own, the genetic algorithm converges quickly and produces the result with the lowest entropy in the timeframe of my experiment (1 day per run of the algorithm). However, I noticed that if I use the output of the genetic algorithm as a starting point for simulated annealing, then the latter is able to further refine the result. This suggests an interesting direction for further research: try to use small simulated annealing runs as a mutation function within the genetic algorithm.

4.1.5. Analysis of the Output

Table 4.1 shows the evolution of word atomizations during the run of the genetic algorithm. The words displayed are a random sample of all words that have different atomizations at all three levels of entropy shown in the table. These samples were not cherry-picked. Notice how the system starts with random alignments between letters and sounds: for instance, it aligns B with *[m]*, R with *[b]*, A with *[r]* in EMBRACE; GH with *[t]* and T with *[ɪ]* in LIGHTNING, NN with *[ʌ]* in RUNNING, etc. Also notice how it arrives at quite intuitive alignments in the rightmost column. The evaluation section of this paper provides quantitative evidence that minimizing entropy corresponds to subjective improvement of atomization quality.

Table 4.1. Evolution of word segmentations during a run of genetic algorithm

word	E = 6.802	E = 4.452	E = 4.135
BUREAU	BU/*bj*-RE/*ʊr*-AU/*oʊ*	B/*b*-URE/*jʊr*-AU/*roʊ*	B/*b*-U/*jʊ*-R/*r*-AU/*oʊ*
EMBRACE	EM/*ɛ*-B/*m*-R/*b*-A/*r*-CE/*eɪs*	E/*ɛ*-M/*m*-B/*b*-RAC/*reɪ*-E/*s*	E/*ɛ*-M/*m*-B/*b*-R/*r*-A/*eɪ*-CE/*s*
ENTITLE	EN/*ɛ*-TI/*nt*-T/*aɪt*-LE/*ʌl*	E/*ɛ*-N/*n*-T/*t*-I/*aɪ*-T/*t*-L/*ʌ*-E/*l*	E/*ɛ*-N/*n*-T/*t*-I/*aɪ*-T/*t*-LE/*ʌl*
LIGHTNING	LI/*laɪ*-GH/*t*-T/*ɪ*-ING/*ŋ*	L/*l*-IG/*aɪ*-HT/*t*-IN/*ɪ*-G/*ŋ*	L/*l*-IGH/*aɪ*-T/*t*-I/*ɪ*-NG/*ŋ*
PHYSICAL	PH/*fɪ*-Y/*zɪ*-SI/*k*-C/*ʌ*-AL/*l*	PH/*f*-Y/*ɪ*-S/*z*-I/*ɪ*-C/*k*-A/*ʌ*-L/*l*	PHY/*fɪ*-S/*z*-I/*ɪ*-C/*k*-A/*ʌ*-L/*l*
PRICE	PR/*pr*-ICE/*aɪs*	P/*p*-R/*r*-ICE/*aɪs*	P/*p*-R/*r*-I/*aɪ*-CE/*s*
RUNNER	RU/*r*-NN/*ʌ*-E/*n*-R/*ɝ*	R/*r*-UN/*ʌ*-N/*n*-ER/*ɝ*	R/*r*-U/*ʌ*-NN/*n*-ER/*ɝ*
SCHEME	SC/*s*-HE/*k*-M/*i*-E/*m*	S/*s*-C/*k*-HEM/*i*-E/*m*	S/*s*-CH/*k*-E/*i*-ME/*m*
TEACHING	TEA/*t*-CH/*itʃ*-ING/*ɪŋ*	TE/*t*-A/*i*-CH/*tʃ*-IN/*ɪ*-G/*ŋ*	T/*t*-EA/*i*-CH/*tʃ*-I/*ɪ*-NG/*ŋ*
UNCERTAIN	UN/*ʌ*-CE/*ns*-R/*ɝ*-T/*t*-AI/*ʌ*-N/*n*	U/*ʌ*-NC/*n*-E/*s*-R/*ɝ*-T/*t*-AI/*ʌ*-N/*n*	U/*ʌ*-N/*n*-C/*s*-ER/*ɝ*-T/*t*-AI/*ʌ*-N/*n*

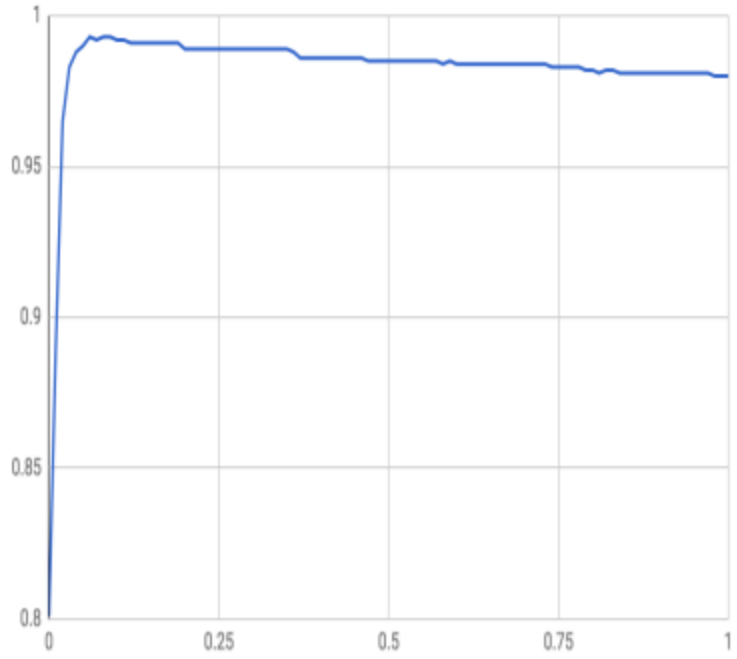


Fig. 4.1. Proportion of single-phoneme atoms over time

Table 4.2. Emergence times of atoms. Blue cells are consonants, red - vowels

#	lettrs	phs	em. time
1	y	i	1.87E-03
2	y	aɪ	2.05E-03
3	j	dʒ	2.14E-03
4	g	g	2.24E-03
5	h	h	2.24E-03
6	f	f	2.38E-03
7	s	s	2.51E-03
8	b	b	2.56E-03
9	m	m	2.62E-03
10	k	k	2.62E-03
11	d	d	2.63E-03
12	n	ŋ	2.64E-03
13	p	p	2.81E-03
14	t	t	3.04E-03
15	e	ɪ	3.08E-03
16	g	dʒ	3.33E-03
17	c	k	3.37E-03
18	a	æ	3.46E-03
19	a	ʌ	3.50E-03
20	u	ʌ	3.51E-03

Although the algorithm has the capacity to group phonemes together, notice that most of the atoms in the rightmost column of the table contain only one phoneme. Indeed, by the end of the run of the whole word set, a vast majority (98%) of discovered atoms contained only a single phoneme. Figure 4.1 shows how this fraction evolves over time (where time is computed as the number of the current generation divided by the total count of generations). This trajectory is remarkably consistent over five runs of the genetic algorithm with different initializations - the difference in respective proportions at a certain iteration is around 0.001. One can see that the algorithm rapidly splits words into single-phoneme atoms, and then very gradually starts to group some phonemes together; a manifestation of this process in the word PHYSICALLY can be seen in the above table. Notice how the atoms PH/*f* and Y/*ɪ* were separate at the entropy level 4.452, but were grouped together later. On the other hand, morphemic units with relatively predictable spelling, such as -ING and -TION, do not end up being atoms, but instead are split into smaller units.

It is notable that some atoms emerge much earlier in the learning process than others. I considered five different runs of the genetic algorithm with different initializations. For each atom instance in every word, I looked at the time when it first emerged in the process of optimization (where the time was defined in the same way as above). The emergence time of an atom was computed as an average emergence time of its instances. Table 4.2 presents the top 20 atoms

that emerged early in these five runs; note that most of them belong to single-letter consonants. There is an interesting parallel between this outcome and the emergent literacy learning of children. Indeed, it has been observed that early in the development of English invented spelling, children typically represent words with consonants only, and later start to introduce vowels (Richgels, 2001). One possible explanation for this phenomenon is that vowels are not as distinct as consonants in the sound stream. However, this hypothesis is not aligned with the fact that in a *transparent* orthography (Italian), where letters and sounds map mostly one-to-one, early spellers make much fewer mistakes writing vowels than consonants (Cossu et al., 1995). Instead, it seems plausible that children acquire vowels later for the same reason that the model does: because spelling patterns of vowels in English are much more diverse and ambiguous.

Similarly, an interesting pattern occurs with emergence times of atom instances in initial, medial and final positions of their respective words. Once again, emergence times were gathered from five runs of the genetic algorithm with different initializations. As seen in Table 4.3, initial atoms emerge first, followed by final ones, and lastly, medial ones. These differences were all statistically significant, with $p < 0.001$ (using Welch t-test). This ordering might be related to the fact that during development of invented spelling in English, children often start by denoting words with letters representing their initial sounds, followed by letters representing their final sounds, and only introduce letters representing medial sounds after setting these boundaries.

Table 4.3. Average emergence time of atom instances in initial, medial and final positions in their words.

initial	medial	final
2.46E-03	8.3E-03	5.7E-03

The most notable counterintuitive behavior of our model is the occasional splitting of letter pairs to place them within two different atoms, as in words ST**EE**R *(*S/*s*-T/*te*-E/*ɪ*-R/*r)*, D**EE**R *(*DE/*d*-E/*ɪ*-R/*r),* A**CC**URATE *(*A/*æ*-C/*k*-CUR/*jɝ*-A/*ʌ*-TE/*t)*, A**LL**EY *(*A/*æ*-L/*l*-LEY/*i)*. This splitting occurs in about 11% of words that contain double letters, and cannot be explained by the stochasticity of genetic optimization alone, because about 70% of the same double-letter splits occur in five different runs of the genetic algorithm. To understand this behavior, consider a word of the type XYY with corresponding phonetic transcription *[ab].* Presume that the pattern Y/*b* is much more frequent than the pattern YY/*b* (which is often the case), whereas the patterns X/*a* and XY/*a* occur with comparable frequencies. Thus, it is preferable for the system to drive down the number of occurrences of the rare pattern YY/*b*, because reducing frequencies of rare patterns reduces the entropy. To observe this effect, consider the example ST**EE**R *(*S/*s*-T/*te*-E/*ɪ*-R/*r).* Note that the CMU pronouncing dictionary transcribes this word with short phoneme *[ɪ]*[22], which is rarely associated with EE (as opposed to the long phoneme *[i]*). On the other hand, association TE/*t* occurs relatively frequently (with silent E at the end of words). This results in the system choosing the counterintuitive behavior. It is possible that such a situation could be avoided if the system was aware of relationships between different phonemes and graphemes, e.g. if it took into account that phoneme *[i]* is close to phoneme *[ɪ]* and grapheme E is close to grapheme EE.

[22] This is possibly a mistake in the CMU dictionary

4.1.6. Evaluation of the Entropy-Based Atomization Approach

a. Experimental Setup

An ideal way to evaluate the above models would be to see if incorporating its output in literacy tools helps learners. Unfortunately, this kind of evaluation is extremely difficult to conduct, which makes it impractical - at least at the current stage of development. Instead, I applied a sanity check to my model — I tested whether decreasing entropy does in fact lead to subjective improvement of atomization quality.

To perform this test, I designed a web interface that presented atomizations atom-by-atom. For each atom, first the phonemes were pronounced, and then the grapheme was highlighted. The participants were then presented with three options: to disagree with the choice made by the model, to agree completely, or to agree with a qualification that the participant would split the word differently. I chose three levels of entropy from a single run of genetic algorithm optimization. Level 1 occurred early in the process (E=6.802), level 2 was located in the middle of the run (E=4.452) and level 3 happened in the final stages of the process (E=4.135). Set A of 50 words for which atomizations changed between level 1 and level 2, and set B of 50 words for which atomizations changed between level 2 and level 3, were randomly selected, resulting in 200 atomizations to be annotated. Each participant received a random selection of 30 atomizations from this list. The target set for each participant was selected to avoid having two different atomizations of the same word in one selection. This was done to avoid the bias caused by previously made annotation decisions. To estimate inter-annotator agreement, three participants received the same set of atomizations for annotation.

This experiment design resulted from several iterations with a few adult volunteers. Early on, I noticed a curious phenomenon resembling the Stroop effect (Stroop, 1935) that interfered with the annotation. Namely, when a grapheme didn't match the sound, the effect of seeing the letters was so strong that the participants often ignored the actual sound produced by the system. To resolve this issue, I decided to (a) present atoms one-by-one, giving the participants more time to pay attention to letter-to-sound matches, and (b) separate presentations of the grapheme and the sounds in time, giving the participants an opportunity to attend to each one. These modifications helped the participants notice the discrepancies, but they were still quite confused upon encountering them, even suspecting a bug in the annotation system. To address this problem, I added a thorough explanation in the beginning of the experiment including a few demo examples and a dry run on a sample word. Nevertheless, even after these modifications, a small amount of confusion with mismatched letters and sounds remained because the mismatches continued to be very unnatural for the participants. This effect might be related to the profound changes in the brain caused by acquisition of automaticity in reading and writing (Wolf, 2008).

I had options to use either a speech synthesizer or a recorded human voice for pronouncing phonemes and their combinations. Unfortunately, no synthesizer available to us was able to produce high-quality sounds of individual consonants, and using a synthesizer for some sounds and human voice for others was likely to introduce a bias in the annotation process. Therefore, I decided to use a recorded voice for all phoneme combinations appearing in my sample. I limited my sample to 50 words for each condition, as it would be difficult to record all combinations of phonemes from an unbounded sample.

Because multiple atomizations were annotated by each participant, and multiple participants could have annotated the same atomization, we couldn't treat each annotation as an independent random variable while estimating statistical significance. I resorted to a linear mixed-effects model to account for dependency between different annotations. I converted each response to a numerical value: "no" - to negative 1; "yes with a qualification" - to 0; and "yes" - to positive 1. I associated the annotation of each atomization by each participant with the average of these responses among all atoms in the atomization. I then looked at the effect of entropy level, controlling for the effect of word and the effect of annotator.

b. Results

A total of 17 participants took part in the study. Table 4.4 shows the word-average of responses calculated in the above fashion. We see that as entropy decreases, participants are more likely to agree with the alignments produced by the model, and this trend has a very strong statistical significance. As seen in the table, the effect size is also quite large. One might wonder why the average response for the entropy level 4.452 is much lower in word set 2 than in word set 1 (0.473 vs. 0.7). This is the result of selection bias: remember that set 2 consists of words whose atomizations changed late in the optimization run. As the analysis of output shows, such words are likely to have difficult letter-to-sound patterns.

*Table 4.4. Average (by word) response in each condition. *** = p<0.001*

Set A	
E = 6.802	***E = 4.452***
0.06	0.7***
Set B	
E = 4.452	***E = 4.135***
0.473	0.792***

The agreement between annotators was only slight (Fleiss' kappa 0.156). These differing preferences could be due to the annotators having been taught to read in different ways. For example, some participants commented that they would prefer to split words at the syllable boundaries, while other participants didn't mind subdividing syllables into smaller units. In this respect, it would be interesting to see whether the agreement level might change if the annotation were conducted by people who taught themselves to read, or by literacy professionals (e.g. speech-language pathologists). Other sources of disagreement may be the dialect differences; some participants noted that they would pronounce the word differently from the system, causing them to mark certain letter-to-sound matches as wrong regardless of the alignment. Finally, as

noted above, there was a significant confusion among the participants regarding what to do with letter-to-sound mismatches. This confusion might have contributed to the low level of agreement.

4.1.7. Discussion

This section introduces a simple, theoretically grounded measure of quality for a set of word atomizations along with algorithms that optimize it. The measure and the algorithms were used to derive atomizations that are vital for several aspects of SpeechBlocks functioning. Aside from SpeechBlocks, the atomizations can be applied in a variety of other technologies designed to support learning to read and write. The output of the automatic procedures corresponds well to human intuition, except for a limited number of cases. The present method is language-independent: it can be applied to any alphabetical language, given that phonetic transcriptions are provided. However, both the current model and its evaluation have a number of limitations.

One problem with the current model is that it is unable to relate similar phonemes and graphemes. For example, it doesn't take into account that phoneme *[i]* is close to phoneme *[ɪ]* and grapheme E is close to grapheme EE. This limitation could be the cause of the counterintuitive behaviors described above. It can, however, be overcome by introducing a set of pseudo-atoms which act as modifiers of the neighbouring atom instances. For example, there can be such pseudo-atoms as 'nasalize the last phoneme in the previous atom instance,' or 'duplicate the last letter in the previous atom instance.' Introduction of the modifier pseudo-atoms will allow the system to operate on a higher level of abstraction. This idea is similar to De Marcken's (1996) word perturbations.

From a cognitive standpoint, the present method exhibits curious parallels with how people might learn to read and write without being formally taught. However, from a cognitive perspective, this approach makes a set of unrealistic assumptions. Most importantly, it assumes that the learner already knows the set of phonemes of the language and phonetic transcriptions of each word. In reality, recognizing and manipulating individual phonemes is the skill of phonological awareness - a difficult one, and the very skill that SpeechBlocks are intended to support. This skill has been shown to develop *in parallel* with literacy acquisition rather than preceding it (Wolf, 2008). Furthermore, phonemes are abstract categories over the set of actual sounds of speech. Our phonetic categorization considers some characteristics of the sounds as important distinguishing features of phonemes while disregarding other characteristics. By paying attention to different characteristics of the sounds, it is possible to come up with a different set of phonemes. That's indeed what children do (Read, 1971). For example, they may spell TRAIN as CHRAN because they find the initial sound of this word closer to *tʃ* than to *t.* A realistic system for acquiring letters-to-sounds patterns should work with the actual sound stream, not with its symbolic representation. A further unrealistic assumption is that the learner knows that words are supposed to be read from left to right. It is also unclear (although it doesn't appear completely implausible) whether learners assume that every letter of a word is related to a certain sound. From the

standpoint of human cognition, another questionable aspect of the current model is that it performs a global optimization over the entire vocabulary. A more realistic model may need to include separate processes running in short- and long-term memory, as in Roy & Pentland's (2002) model of learning words' referents. Despite its shortcomings, the current model may serve as a basis and a baseline for more sophisticated cognitive models that could address these limitations. Models simulating development of children's invented spelling might be of particular interest.

The model produces word segmentations at only one level of fidelity. It would be interesting to consider a model that could produce a hierarchy of segmentations, similar to what De Marcken's (1996) model does. For example, such a model would be able to group ING/ɪŋ into a high-level block that is further divisible into I/ɪ and NG/ŋ. Such a model might account for different segmentation preferences of different annotators. Its outputs also may be related to the developmental trajectory of literacy learners. Ehri (2005) describes that as learners gain more and more expertise in reading, they memorize the pronunciation of more and more complicated letter patterns (called sight chunks) until eventually, they expand the sight chunks repertoire to entire words.

In my experiments, I didn't take into account word frequencies. Considering them might not only make the model's learning more similar to human learning, but also improve the quality of atomizations. Because it introduces only a minimal change to the model, considering word frequencies is a promising short-term direction for future work.

The annotators in my experiment were taught to read and write in their childhood, rather than having learned written language independently. Nor were they literacy experts who know the theoretical aspects of literacy acquisition. It would be interesting to re-run the experiment with these two categories of people and see if there would be any difference in results.

Finally, while I have shown the directional association between decreasing entropy of the model and subjective quality of atomizations, I only examined a few atomization sets. Therefore, I haven't demonstrated that *any* decrease in entropy necessarily leads to an increase in subjective quality. In particular, I haven't compared the output of my model with results of different existing methods. Conducting such a comparison is an important direction for further research.

4.1.8. Application to SpeechBlocks

In SpeechBlocks II and the late versions of SpeechBlocks I, I ended up using the output of the EM-like algorithm, since it was available early on and was deemed to be of sufficient quality. To exclude artifacts introduced by the limitations of the algorithm, I performed some manual post-processing. I printed out the set of discovered atoms, each with 50 examples of words in which it occurs. I then manually scrutinized unusual atoms and corrected the segmentations if necessary, e.g. to exclude the above-described phenomenon of subdividing the pairs of double letters.

4.2. Inferring Pronunciation/Spelling and Atomization for Out-of-Vocabulary Words

For within-vocabulary words, SpeechBlocks retrieves pronunciations and atomizations from a large dictionary, whose derivation was described in the previous section. However, regardless of the size of built-in vocabulary, it is always possible to encounter a novel word via text or speech recognition, or to encounter one built by the child in the open-ended mode. SpeechBlocks I relied on the speech synthesizer to infer pronunciation of such words on its own. However, this encapsulates the pronunciation information within the synthesizer. As described earlier, various systems within SpeechBlocks need to 'know' both pronunciation and atomization of each word. Furthermore, when the system works in the phoneme mode, it needs to make inferences in the opposite direction - to infer spelling from pronunciation. This section describes the machine learning models that are used to make these inferences.

Inferring pronunciation of an out-of-vocabulary word is a much-studied problem called grapheme-to-phoneme transduction. Good examples of modern grapheme-to-phoneme transducers can be found in (Rao et al., 2015) and (Toshniwal & Livescu, 2016). These systems use a sequence-to-sequence approach with layered LSTMs, or with neural attention. Inferring spelling of an unknown word from its pronunciation is called phoneme-to-grapheme transduction. This type of transduction is much less common, but methods used for grapheme-to-phoneme transduction can still be applied to it. In this work, I perform transduction jointly with atomization. An alternative approach could use an existing transducer to derive pronunciation or spelling, and then position atom boundaries within pronunciation and spelling so that the entropy of the current atomization plus the entire atomized vocabulary is minimized. I went for the joint approach, because I was curious whether I could improve upon the accuracy of existing transducers by utilizing alignments derived in the previous section.

For benchmarking purposes, I used the Phonetisaurus dataset (Novak et al., 2011), which is a particular version of the CMU Pronouncing Dictionary commonly used for benchmarking of grapheme-to-phoneme transducers. However, for the inverse (phoneme-to-grapheme) system used in SpeechBlocks, I used a subset of the 50K most common words from that dataset. This is because, qualitatively, I saw that the system trained on a larger dataset starts to generate overly elaborate spellings, such as BAUGH for *[b;ɔ]*. Therefore, I had to artificially "dumb it down" to make it more usable for children.

For the grapheme-to-phoneme transducer, the data was encoded in the following way. Each letter of the input word is represented as a one-hot vector. Similar representation is used for output phonemes (and their combinations, like *[k s]*). Output phonemes are aligned with the initial letters of their grapheme. For all other slots, the one-hot indicator for "empty" phoneme is raised. Therefore,

the output information can be used to decode both pronunciation of the word and its alignment with graphemes.

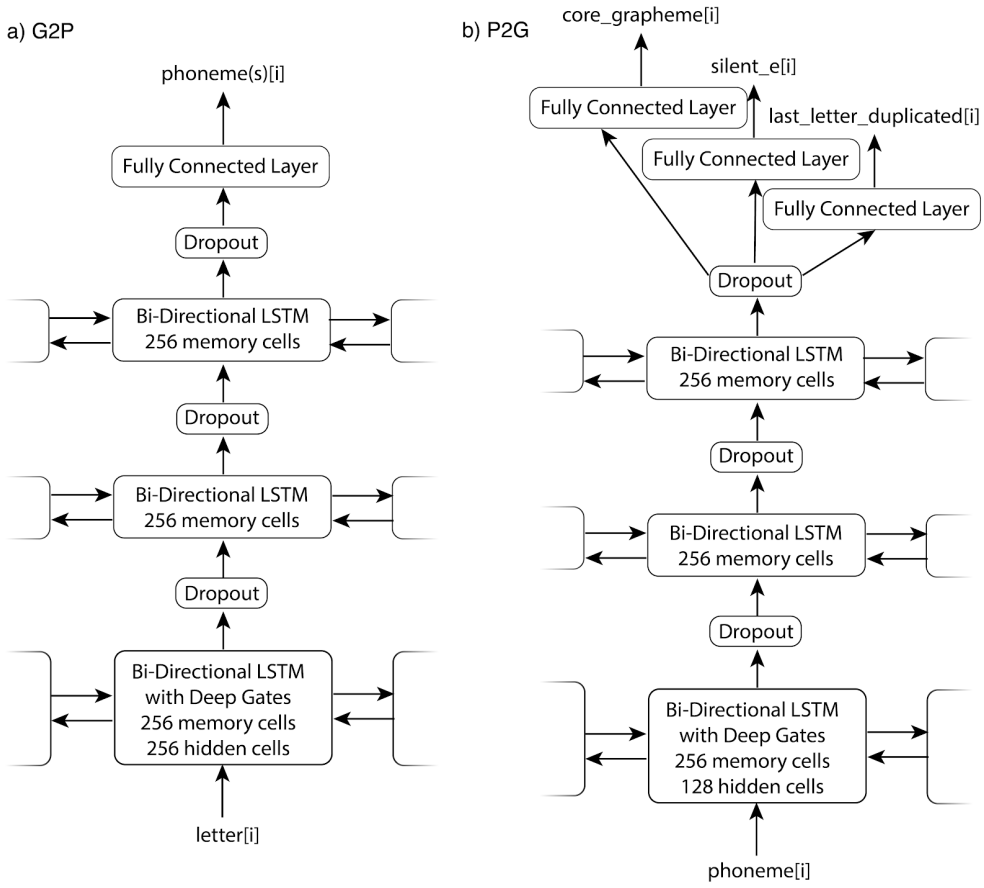


Fig. 4.2. Best architectures for
(a) grapheme-to-phoneme and (b) phoneme-to-grapheme transducers

For the phoneme-to-grapheme transducer, a slightly more sophisticated encoding of the output graphemes is used. There are two common phenomena occuring at the end of many graphemes: addition of a silent *e* duplication of the final letter. It is beneficial to abstract these features away to reduce variability of the outputs. Therefore, each slot of the phoneme-to-grapheme transducer is associated with three outputs: one-hot vector for the core grapheme, and two boolean flags controlling duplication and silent *e* modifiers. A similar strategy is used to signify alignment.

I tried several machine learning approaches - decision trees, logistic regression and recurrent neural networks - and found the latter to offer by far the best performance. The architectures that showed the best performance for grapheme-to-phoneme and phoneme-to-grapheme transducers are presented on Fig. 4.2 (a) and (b) respectively. They are fairly conventional and are composed of LSTM (Hochreiter & Schmidhuber, 1997) and dropout (Hinton et al., 2012) layers. One unusual feature is the “deep gates” at the first LSTM layer. This feature adds an additional hidden layer to the networks controlling LSTM gates; the layer is shared between all gates. The intuition behind it is

to allow LSTM to make more sophisticated decisions about keeping, forgetting, inputting and outputting bits of data. In this setup, the deep gate architecture performed better than adding an additional layer to the stack or adding an embedding to the input. However, deep gates have proven to be useful only on the first layer.

The networks were implemented in Tensorflow (Abadi et al., 2016). Training was performed in batches of 128 words. Adam optimizer (Kingma & Ba, 2014) was used, with the initial learning rate of *0.001*. The learning rate was decreased whenever no progress of the training WER was observed for more than *P* epochs (with initial P being selected as 40). In this case, the learning rate was divided by 2 and P was increased by 2. Early stopping was used: the iteration with the lowest WER on development set was selected as the final one. Gradient clipping was applied, with clipping value of 1.

To compare my approach with existing ones, I removed the atom boundaries from the output of the grapheme-to-phoneme transducer and looked at its accuracy for pronunciations only. The word error rate for this task ended up being 29.23%. Unfortunately, this is quite a bit higher than the rates from sequence-to-sequence approaches on the same dataset: 25.8% by (Rao et al., 2015), 21.69% by (Toshniwal & Livescu, 2016) and 20.24% by an ensemble model of the same authors. Therefore, my hope that knowledge of alignments would help to improve over these results didn't materialize. Nevertheless, I found the performance to be good enough to be useful within the current version of the app. Interestingly, the phoneme-to-grapheme transducer trained using similar techniques showed still higher word error rate: 44%. This might reflect an interesting feature of the English language: its spelling rules might be much less regular than its reading rules.

4.3. Interpreting Invented Spelling

One of the approaches to scaffolding used by SpeechBlocks II is interpreting children's invented spelling. Based on the word assembled by the child, the interpreter comes up with a list of guesses of what the child may be attempting to spell, and displays these guesses to the player. If the player selects one of these guesses, the app either completes the word (if the difference between the invented spelling and target word is relatively small), or guides the child through the process of completing it. In this section, we look at the algorithm allowing the app to produce such a list of guesses.

My approach is based on using an indexation scheme to quickly retrieve a list of candidates, and then applying a modification of the Wagner-Fischer algorithm (which is normally used for computing the Levenstein distance) to rank the candidates according to their similarity to the child's word. The custom modification of the algorithm is necessary, because there is a lot of specificity in how children tend to form invented spellings. In fact, the desired similarity function is not even a metric, because it doesn't conform to the symmetry requirement. Indeed, while plausible interpretations of the string *KT* are *CAT* and *KITE*, *KT* is not a plausible interpretation of either *CAT* or *KITE*. Speaking more generally, it is unlikely that the intended word is shorter than its invented spelling - so a short

source string might have a high similarity to a long word, but the same won't be true when their places are switched. Even though the target function is not a metric, the Wagner-Fischer is still applicable to this task.

I built the invented spelling interpreter to work both with letter and phoneme modes of SpeechBlocks II. To accomplish this, I treated both the source (the child's word) and the target (the word it is compared to) not as sequences of letters, but as sequences of atoms (as defined in section 4.1; roughly speaking - grapheme-phoneme pairs). In the letter mode, this added an additional benefit of the algorithm being able to utilize pronunciation information derived by the grapheme-to-phoneme prediction network. However, it is possible to re-formulate the present algorithm to work on sequences of letters.

Ideally, one would build an invented spelling interpretation algorithm by starting from a large dataset of invented spellings together with their intended targets. Unfortunately, I was unable to identify any such dataset openly available. At the time, I also didn't have access to invented spelling data from SpeechBlocks play, which was extracted later. So, instead of the data-driven method, I used the literature on the development of invented spelling to inform the algorithm. Invented spelling reflects development of children's early phonological knowledge (Read, 1971) and goes through several stages. Early invented spellers typically represent just the initial letter or sound of a word (Ouellette et al., 2013). Sometimes this initial symbol is followed by a string of random symbols. Next, learners typically progress by adding the final letter or sound (ibid.) Then, the medial sounds are added (ibid.) Vowels typically start to be used later than consonants (Richgels, 2001). In the early stages, letter names are commonly used to represent letter sounds. For instance, *HR* may mean *CHAIR*: here, *H* stands for the *[tʃ]* sound, since the name of the letter is pronounced as *[eɪ tʃ]*. Alternations between voiced and unvoiced consonants (e.g. *[t]* and *[d]*) and tense and lax vowels are common. Invented spelling can represent some phonological phenomena that conventional spelling glosses over. For instance, notice how initial sounds of *TRAIN* and *DRAGON* are perceived not as *[t]* and *[d]*, but as *[tʃ]* and *[dʒ]*. This phenomenon is called *affrication*: the pronunciation is affected by the nearby fricative *[r]*. Conventional spelling doesn't reflect this phenomenon, but child spelling can: children may spell these words as *CHRN* and *JRGIN*. As children's knowledge further progresses, they start to utilize orthographic patterns, such as silent letters, in their spelling (Richgels, 2001). Sometimes they over-generalize the usage of these patterns - for instance, by inserting silent letters where they don't belong (Bissex, 1980). Children may also remember what parts of the target word *look like*, without involving their phonological knowledge. We call this phenomenon logographic spelling. Sometimes, logographic and phonetic knowledge were combined in building a single word. Table 4.5 provides some notable examples of invented spelling together with descriptions of the corresponding patterns.

Table 4.5. Invented spelling phenomena

Invented Spelling	Intended Word(s)	Notable Phenomena	Source
KT	CAT	Using only initial and final sounds. Representing sound *[k]* with letter K - most common for this sound.	observation
SOD	SWORD	Using O to represent the prolonged *[ɔ]* sound made by WOR	observation
PL	APPLE	Representing only the initial and final consonant sounds	observation
LAD	LADY	Representing the last syllable (DY) as D, because D's letter name *[di]* is similar to the pronunciation of that syllable.	(Ouellette et al., 2013)
GRDN	GARDEN	Using letter names *[ʌr]* and *[di]* as the corresponding syllables	(Read, 1971)
TABIL	TABLE	Note that the word is pronounced as *[teɪbəl]*; the child uses I to capture the schwa sound, because the name of the letter *[aɪ]* sounds close to the desired vowel.	(Read, 1971)
RUDF	ARE YOU DEAF?	(1) Using letter names *[ʌr]* and *[ju]* as the corresponding words. (2) Using letter names *[di]* and *[ɛf]* to represent syllables in DEAF.	(Bissex, 1980)
FES	FISH	Representing phoneme *[ɪ]* with E, since the name of the letter is phoneme *[i]*, which is close to *[ɪ]*.	(Read, 1971)
FOTR	FATHER	(1) Representing sound *[ɑ]* with O - a letter name that sounds similar. (2)Representing the vowelized r sound *[ɝ]* simply as R.	(Read, 1971)
FEGR	FINGER	(1) Using E to represent sound *[i]*. The child does not represent the nasalization, which is a common pattern for children before five. (2) Representing the *[ɝ]* sound with letter R.	(Read, 1971)
JRAGIN	DRAGON	Perceiving the affricated *[d]* as *[dʒ]* and representing it correspondingly. Representing the schwa sound as I.	(Read, 1971)
SRKIS	CIRCUS	(1) Representing sound *[s]* with letter S - most common for this sound. (2) Again, representing the *[ɝ]* sound with letter R and representing the schwa sound as I.	(Read, 1971)
ALRVATA	ELEVATOR	(1) Representing *[ɛ]* as A (whose name sounds *[eɪ]*). (2) Alternating between *[ʌ]* and *[ɝ]* in both directions within the spelling *[ɛlʌveɪtɝ]*.	(Read, 1971)

These patterns were accounted for in computing the distance between the source and the target strings in the following way. Tokens or short sequences of tokens in the source string are assumed

to match tokens or short sequences of tokens in the target. Each match is associated with a certain price, equal to zero in case of perfect match and becoming higher in case of more distant matches. Every unmatched token in the source and target strings is also associated with a price. The price is set much higher for the unmatched tokens in the source, out of assumption that children are more likely to undergenerate than to overgenerate symbols in their invented spelling. The cost of unmatched consonants is higher than the cost of unmatched vowels, to accomodate for the observation that children are more likely to skip vowels in their invented spellings. Lack of match in the final, and particularly the initial position of the word, is also much more costly, accounting for the fact that children are very likely to incorporate initial and final letters/sounds in their invented spelling[23]. All the costs specified above were selected based on my best judgement. When a database of invented spelling and their interpretations becomes available, it would be possible to optimize these costs using machine learning techniques. We assume that the order of tokens in the child's invented string is the same as in the target word - otherwise the search for candidate matches becomes too loose. The total distance between the source and target strings is computed as the minimal sum of all matches and mismatches costs for all possible mappings of tokens from the source to the target string. The complexity of the algorithm is *O(N*M)*.

To see how the Wagner-Fischer algorithm can be adapted to this problem, we can first look at the original, unmodified version of the algorithm (Wagner and Fischer (1974), likely first proposed by Vintsyuk (1968)). The algorithm computes the minimal number of letter insertions, deletions and substitutions needed to convert one word into another. To do so, it utilizes a technique known as dynamic programming. Dynamic programming in general goes from small sub-problems of the target problem to the larger ones, storing the solution to each sub-problem in memory, and uses the already computed solutions to derive the new ones. In particular, the Wagner-Fischer algorithm keeps a table *D* where each cell with coordinates *i, j* contains the cost of converting the prefix of the first *j* characters of the source word into the first *i* characters of the target word. An example of this table is shown on Fig. 4.3. I will denote the length of the target string as *M*, the length of the source string as *N*; the cell at the intersection of *i-th* row and *j-th* column as *D[i, j]* (assuming that indexes start from 0), and the source at target prefixes as *S[0 : j]* and *T[0 : i]* respectively. The 0*th* row is filled with 0 ... n, because the shortest way to convert *S[0 : j]* into the empty string is to delete *j* letters. The 0*th* column is filled with 0 ... m, because the shortest way to convert the empty string into *T[0 : i]* is to insert *i* letters. Let us see how we can compute a value in a *D[i, j]* (with *i* and *j* greater than 0), assuming that the table already contains values for all the cells with lesser indices. We have only three choices of what can be done with the last letters in *S[0 : j]* and *T[0 : i]: S[j]* and *T[i]*. One choice is to delete *S[j]*. The price of this option is the cost of converting *S[0: j-1]* to *T[i]* plus the cost of one deletion. So, it is *D[i, j-1] + 1.* Another choice is to insert *T[i]*. Reasoning similarly, the price of this option is *D[i-1, j] + 1*. The final choice is to align *S[j]* and *T[i]*: if they match, the price is equal to *D[i-1, j-1]*; otherwise, we have to add 1 for substitution. Out of these three choices, we select the one resulting in the smallest cost. Using this formula, the algorithm sequentially fills the

[23] This is implemented in a bit more complicated way in the actual code. Noticing that children can spell APPLE as PL, I also account for the case when the initial vowel is skipped, but initial consonant matches. I omit this detail for simplicity of the explanation. Modification of the algorithm to accomodate for the case above is quite straightforward.

cells of the table from top to bottom and from left to right, until it arrives at the final solution in *D[M, N]*. The algorithm can then start from *D[M, N]* back-track through the table, looking for the neighbour of the current cell with the minimal cost, in order to reconstruct the optimal sequence of actions taken to perform the conversion. The complexity of the algorithm is *O(K*L*N*M) = O(N*M)* (since K and L are constants).

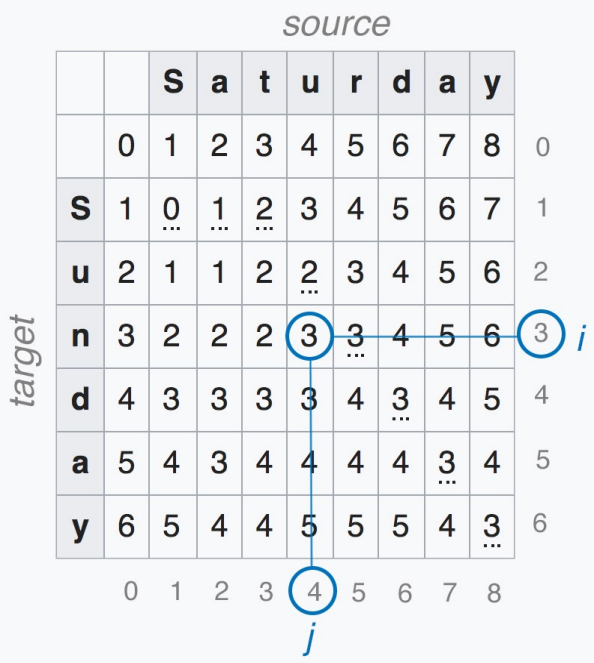


Fig. 4.3. The dynamic programming table of the original Wagner-Fischer algorithm. Image adapted from Wikipedia[24].

To accomodate for the specifics of the invented spelling interpretation, the algorithm needs to undergo the following modifications:

1. Introduce differential costs of various types of matches, e.g. to account for vowels being more likely to be skipped;
2. Introduce an option to substitute several tokens with several other tokens, e.g. to account for encoding multi-letter atoms logographically, such as PH/*f* - as P/*p*-H/*h*. Tokens in the substitution must be consecutive, and their number must not exceed *K* for the source and *L* for the target. In the actual implementation, we used K and L both equal to 2;
3. Account for the extra price of un-matched tokens in the initial and final position of the target string.

Table 4.6 provides the costs of various matches and non-matches used by the algorithm. These costs were manually selected to facilitate intuitive behavior of the algorithm. Once a good training

[24] https://en.wikipedia.org/wiki/Wagner-Fischer_algorithm

set of invented spellings becomes available, it would be possible to automatically fine-tune these costs to maximize performance on the dataset.

Table 4.6. Costs of matches and lack of matches

Type	Example	Cost
Perfect match	R/*r* → R/*r*	0
Phonetic only match	S/*s* → CE/*s*	0.25
Letter name as sound	H/*h* → CH/*tʃ*	0.25
Voiced/unvoiced alternation	T/*t* → D/*d*	0.5
Logographic only match	C/s → C/k	0.5
Affrication match	CH/*tʃ*-R/*r* → T/*t*-R/*r*	0.25
Skipping source token	-	10
Skipping target token	-	1
Extra for skipping initial in source	-	15
Extra for skipping final in source	-	5
Extra for skipping initial in target	-	4
Extra for skipping final in target	-	2
Extra for skipping consonant	-	1

The modified algorithm will again keep a dynamic programming table *D* of the size *M+1 x N+1*. The *0th* row is filled with costs of skipping *j* tokens from the source string. The *0th* column is filled with prices of inserting *i* unmatched tokens at the beginning of the target string. In these calculations, an extra cost is paid for skipping/inserting tokens in the initial position; we also pay different amounts for unmatched consonants and vowels. Now let's look at how *D[i, j]* (with *i* and *j* greater than 0) is computed. Again, we are looking at different options of what can be done with the tails of *S[0 : j]* and *T[0 : i]*. The first two options are to leave unmatched the last tokens either in the source or in the target; in this case, we take the values of *D[i-1, j]* and *D[i, j-1]* correspondingly and add the cost of the corresponding lack of match. If *i=M* and *j=N*, we pay the additional price for leaving the final token of the source and the target unmatched. We can also look at all the matches of the source suffixes *S[j-k : j]* to the target suffixes *T[i-l : i]*, where $k \in 1 .. K$ and $l \in 1 .. L$. The cost of each such option is *D[i-l, j-k]* plus the cost of the suffixes match. Of all these options, we select the one with the minimum cost. Similar to the original algorithm, we proceed left-to-right

and top-to-bottom to fill the table and retrieve the result from *D[M, N].* We can then back-track to find the alignment between the strings.

Table 4.7. A sample dynamic programming table of the invented spelling interpreter

i/j		0	1	2	3
			CH/*tʃ*	R/*r*	S/*s*
0		0	0+10+15+1 = 26	26+10+1 = 37	37+10+1 = 48
1	T/*t*	0+4+1+1 = 6	26+4+1+1 = 32	37+4+1+1 = 43	48+4+1+1 = 54
2	R/*r*	6+1+1 = 8	32+1+1 = 34	0 + 0.25 = 0.25	0.25+10+1=11.25
3	A/*eɪ*	8+1 = 9	34+1 = 35	0.25 + 1 = 1.25	1.25+10+1=12.25
4	CE/*s*	9+1+1 = 11	35+1+1 = 37	1.25+1+1=3.25	1.25+0.25 = 1.5

Table 4.8. Examples of how cells in table 4.7 were computed.

Cell	Formula	Explanation
D[0,1]	0+10+15+1 = 26	0 for the cost of D[0,0] + 10 for skipping a source token + 15 for skipping the first token + 1 for skipping a consonant
D[1,0]	0+4+1+1 = 6	0 for the cost of D[1,1] + 4 for not matching initial target token + 1 for not matching a target token + 1 for not matching a consonant
D[3,0]	8+1 = 9	8 for the cost of D[2,0] + 1 for not matching a target token. No extra penalty for not matching a vowel applies
D[2,2]	0 + 0.25 = 0.25	0 for the cost of D[0, 0] + 0.25 for the affrication match of CH/*tʃ*-R/*r* → T/*t*-R/*r*
D[3,2]	0.25 + 1 = 1.25	0.25 for the cost of D[2,2] + 1 for skipping a target token.
D[4,3]	1.25+0.25 = 1.5	1.25 for the cost of D[3,2] + 0.25 for matching only the sound in S/*s* → CE/*s*

Let's illustrate this algorithm on a specific example: matching invented spelling *CHRS* with the target *TRACE*. This target is not very likely, but it works well for illustrating how the algorithm handles various invented spelling phenomena. As we are using grapheme-phoneme pairs as tokens, we are actually matching strings CH/*tʃ*-R/*r*-S/*s* and T/*t*-R/*r*-A/*eɪ*-CE/*s*. In this example, we will be using the costs of matches/non-matches specified in table 4.6. The corresponding dynamic

programming table is shown in Table 4.7. Explanations of how a few cells in the table were computed are given in Table 4.8. Backtracking the optimal trajectory in the table, we discover that we need to match CHR in source with TR in target, skip A in target and match S in source with CE in target.

Running the Wagner-Fischer algorithm for every word in the vocabulary would be prohibitively slow. To expedite this process, I use an indexation scheme to quickly retrieve a set of candidate matches from the vocabulary for a more detailed analysis by the Wagner-Fischer algorithm. This indexation scheme assumes that either the initial token or the initial consonant in the target word is represented. For each word in the vocabulary, it considers all pairs consisting of initial token / initial consonant in the word and the arbitrary token following it. For each pair, it derives a set of signatures in the form *A-B*, where A is either a phoneme or letter of the first token, and B - that of the second. The indexation procedure then takes all signatures for all pairs and stores the word under each one. When a source word arrives, the retrieval procedure computes a set of signatures from its first and last tokens, and retrieves all words with matching signatures from memory.

In forming guesses of what an invented spelling might mean, we should take into account not only the goodness of fit between the source and the target words, but also children's likelihood of trying to use such a target word. For example, it is rather unlikely that they would attempt to spell such words as *population* and *attorney*. I assumed that a good proxy for how likely children are to use a word is its frequency in children's literature. To estimate this frequency, I used Facebook's Children Books Dataset (Hill et al., 2015). However, I found this dataset somewhat biased - e.g. it contains only a few instances of such words as *robot*, which I found quite frequently used by children. To compensate for this bias, I artificially assigned a count of at least 1000 to all imageable words. Names of children from the study classrooms also received this artificially high count. Furthermore, I assumed that it is more likely for children to target nouns than adjectives or verbs. The final cost of a word was computed using the formula $WF + POS - C * log(freq)$, where *WF* is the cost derived by Wagner-Fischer algorithm, POS is the cost of using a certain part of speech (zero for nouns, and a higher cost for adjectives and verbs), *freq* is the word's frequency count, derived as described above, and *C* is a scaling constant.

4.4. Tracking Text In Different Frames During Text Recognition

One of the inputs for the scaffolding procedure explored in SpeechBlocks II is text recognition. As it is described in section 3.2.4, the text recognition interface went through several iterations. Results of playtesting suggested the importance of supporting real-time, on-device text recognition. I examined several existing text recognition libraries capable of fulfilling this demand. However, a challenge emerged. Existing real-time OCR libraries operate at about 2 frames per second and tend to produce inconsistent results between frames. A word recognized in one frame may not be recognized in another, or may be recognized as a different word. To maintain smooth

user interaction, the markers for the recognized text (e.g. the blue boxes in the current design) need to be stable and to "stick" to their source irrespective of camera motion. It is also desirable to aggregate recognition results of the same word over multiple frames, to improve recognition quality. This section describes how the recognized words were tracked across multiple frames in order to facilitate these functions.

To perform tracking, I utilized the OpenCV computer vision library. The tracking procedure starts with identifying good features to track within the image, utilizing Shi & Tomasi (1993) algorithm. It tracks the motion of these features, utilizing Bouguet et. al.'s (2001) implementation of Lucas-Kanade optical flow algorithm. As this operation is repeated from frame to frame, the track of the original features is gradually getting lost. When the number of tracked features falls below a certain threshold, the procedure re-generates them. Thus, at every moment of time, the procedure has (1) a reference frame with a number of features in it, and (2) positions of a subset of these features on the current frame.

Using original features and their new positions, the tracking procedure computes the homography - the matrix describing the relation between the same points on a plane, viewed from two different angles - using Random Sampling Consensus method (RANSAC) (Fischler & Bolles, 1981). RANSAC allows the system to exclude outlier points while computing the transformation, and, as a result, to ignore motions happening in the background and not related to overall motion of the camera. To avoid accumulation of error, the procedure computes the homography between the current frame and the reference frame, rather than between the current frame and the previous frame. For each piece of text, the procedure keeps another homography, projecting it on the reference frame. Using the product of these two homographies, the procedure can compute the current position of each piece of text, which is displayed for the user.

The approach described above has its limitations. It only tracks the overall motion of the scene/camera and assumes that this motion is reducible to a combination of translation, rotation, scale and perspective transformation. Therefore, it doesn't work well when (a) the motion of the tracked word is not aligned with the overall motion of the scene (e.g. it is written on a moving car), and/or (b) the transformation of the scene is not linear (e.g. bending the book page). Nevertheless, I deemed the approach working sufficiently well for typical text recognition scenarios.

Text tracking also allows matching of the recognized words in different frames. However, I found that optical tracking alone is insufficient. Deviations from perfect linear transformation prevent the projected text from aligning perfectly with its actual new position. With only optical tracking, it becomes easy for the system to confuse word identities when they are closely spaced (e.g. in a book), and the camera performs quick motions (e.g. shaking in the hands of a child). Therefore, in addition to optical tracking, I also check the similarity between the matched bits of text. The matching procedure between two frames (called the source and the target) is as follows. First, it enlarges the bounding box of the source word by a certain offset, to increase the likelihood that it will overlap with the new box of the same word in the target frame. It then projects the source bounding box onto the target frame and retrieves all the words in the target frame whose bounding

boxes overlap with the projected box. Those words are the candidate matches. The procedure then computes the Levenstein distance between the candidates and the source word and narrows the list of candidates to those with minimal Levenstein distance. If there is more than one such candidate, it selects the one with its position closest to the projection of the original word. I was satisfied by this routine in practice.

4.5. Selecting Candidate Results for Speech Recognition

Another important input for the scaffolding procedure is speech recognition. However, speech recognition of children's voices is much less reliable than of adult voices. Additionally, children tend to converse with the speech recognition in sentences, rather than simply requesting the word that they want to build. To alleviate these issues, the speech recognition interface was designed to show the child multiple candidate words that may correspond to her/his request. Here, I describe the procedure used for this purpose.

Speech recognition on children's voices is relatively difficult. One reason is because children's voices are out-of-domain: speech recognizers are typically trained on adults' voices, and children's voices have different parameters. Another reason is that children's voices are simply harder to recognize, even for humans, because children's articulation is still developing. In our studies, this was confounded by additional issues. The system was used in a classroom environment, with its background noise. Much of the children's pronunciation was characteristic of minorities, further deviating from the recognizer's likely training data. Finally, children often didn't simply say the word they wanted to build, but communicated with the system in sentences, e.g. "I want to spell PIRATE", or "Please give me a ROCKET". Parsing these sentences would be difficult because of the great variety of possible request patterns, further amplified by speech recognition errors.

I found that the following strategy worked quite well in addressing all these problems. The app requested the Google Speech API to provide the top 20 candidate interpretations of the input speech segment. Each of the returned interpretations had an associated confidence score and could be an entire sentence. The app splits each interpretation into different words, censored swear words and weighed each word by the product of the associated confidence score and the word frequency score (derivation of which is described in section 4.3 about the invented spelling interpreter). Therefore, imageable words, and names and words that appear frequently in children's literature all received a boost. The app then outputs top 10 results to be presented to the player.

Chapter 5. SpeechBlocks I in Action

This section examines the experience derived from the earlier medium, SpeechBlocks I. Although the app had limited capabilities (in particular, its early variants didn't implement the scaffolding principle at all, and its later variants only implemented it in a minimal way), it showed the potential of expressive media for early literacy learning. In fact, it was after observing children's play with SpeechBlocks I that the notion of expressive play came to the forefront in my own thinking. Before, I had thought of SpeechBlocks primarily as a tool to explore patterns of letters and sounds. I found that SpeechBlocks facilitated children's sense of agency, self-efficacy and ownership of their work. Such behaviors as goal setting and planning were observed. Children exhibited many signs of engagement, although in absence of a supportive structure in the home condition, their engagement dropped quite quickly. Children enjoyed talking about what they made both with adults and with each other. I saw that children were remarkably social during their play with the app. Their interactions centered around their building process and showed potential to serve several functions for learning: inspiring each other, maintaining mutual engagement, and learning from each other. Play with SpeechBlocks took a variety of forms: (1) impulsive exploration, (2) word/phrase crafting, (3) narration, (4) remixing and rhyming, (5) communicative play, and (6) using SpeechBlocks as a reference to write words on paper. These forms will be analyzed below. In home studies, I saw more diversity of play types and more sophisticated play, which can be partly attributed to the environment and partly to the inclusion of older participants. The play of younger children was limited by their spelling skills, causing frustration and disengagement. I discovered the need for scaffolding of word building, and found that the facilitator's capacity to provide such scaffolding was limited by the need to divide attention among multiple children. This observation suggested the need for built-in scaffolding.

5.1. Studies Setup and Procedures

5.1.1. The First Pilot: SpeechBlocks I at a Preschool

This study marks the first time when SpeechBlocks was used by children. The goal of the study was simply to see what happened when children used the medium. We were interested as to whether children would be engaged; what types of words they would try to build with the app; what kind of supplementary materials and activities would be needed; and whether they would play on their own or with one another. We observed children playing both freely and in the context of structured activities. We also experimented with different supplementary materials to use alongside the app. Results from this study were published (Sysoev et al., 2017).

The study took place in a daycare center belonging to a university in the Greater Boston area. Its daily routine was structured around rotations between different activity stations, occuring every 15 minutes. We set up one of the stations to work with SpeechBlocks, consisting of a table around

which four children and a facilitator (one of the researchers) could sit. Another researcher sat nearby and took notes. On the table, there were phones and, in some sessions, supplementary materials. Children were directed by teachers to come to our station during their rotation, but they were free to leave early and play in the common area of the classroom if they didn't want to play with the app anymore. 16 children (12 girls and 4 boys) participated in the rotations. They were between 4 and 5 years old, had no speech or hearing disorders and were typically developing. They were mostly from the families of faculty and staff of the university, and thus were likely to have relatively rich literacy environments at home. However, their literacy skills varied greatly: some were already able to spell a few polysyllabic words, while others did not yet know all the letters of the alphabet. We ran sessions with children two days a week, for ten weeks.

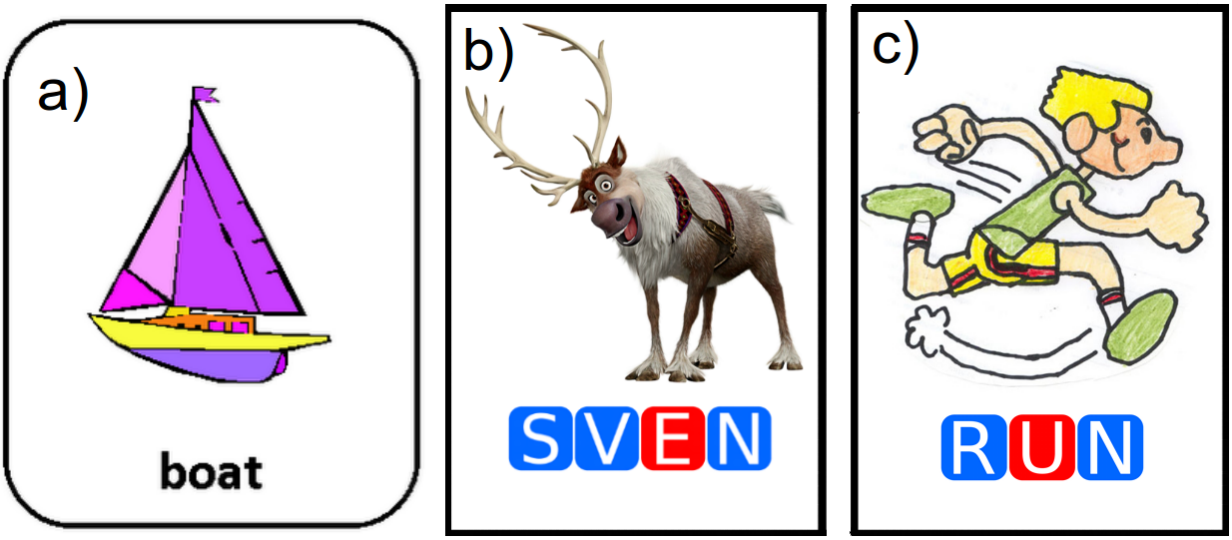


Fig. 5.1.[25] Supporting materials: articulation (a), character (b) and action (c) cards

Fig. 5.2. Supporting material: a story card

[25] Fig. 5.1 And Fig. 5.2 are reused from Sysoev et. al. (2017). https://dl.acm.org/doi/10.1145/3078072.3079720

During this early study, we tried to figure out a "recipe" for how SpeechBlocks should be used, and thus experimented not only with the app, but also with supplementary materials. The first day of play was set up as free play: we were interested in what children would choose to do on their own. Thus, we demonstrated the use ofSpeechBlocks, but suggested no activity. During the second day, we suggested some themes - e.g. animals, parties, food - which were, however, largely ignored by children. Starting from the third day, we experimented with supplementary materials aimed at helping children build real words. We began by bringing articulation cards (Fig. 5.1, a) - a tool normally used by Speech-Language Pathologists to practice common pronunciation patterns. In our case, they were used as a source of ideas for children and as an inspiration for making words. On the fifth and sixth day, we attempted to introduce a structured activity, similar to Mad Libs[26]. Children were read a small story with missing words, and were asked to fill in these words however they liked by spelling them in SpeechBlocks. On the seventh and eighth day, we reverted to free play. On the ninth day, we introduced a new type of card, inspired by children's apparent interest in cartoon characters that we observed on earlier days (Fig. 5.1, b). These cards also featured words written in "SpeechBlocks font", to help children locate relevant blocks within the app. To encourage children to create phrases and stories, we later introduced cards for verbs, which we called "action cards" (Fig. 5.1, c). Finally, we wrote down some of the phrases and stories that children produced orally, and created cards with them as well (Fig. 5.2).

A different type of supporting material was aimed at preserving children's creations. The earliest version of SpeechBlocks didn't have a means to save words. As a substitute, we used little notebooks in which children's words were written down by hand. Originally, we intended to do the notations ourselves, but we discovered that many children were both able to copy words from the screen and highly willing to do so - so we gave the notebooks to them.

There were two sources of data. First, SpeechBlocks recorded log files, which captured everything that happened within the digital realm. Second, we collected video recordings of the sessions and open-ended handwritten observer notes. Tools called PlayObservatory and PlayTrees (Soltangheis, 2017), detailed in the Appendix C, were used for analysis of this data.

5.1.2. SpeechBlocks I in Home Conditions

In two subsequent studies, children used SpeechBlocks in home environments. These studies were led by other students in the lab, my colleagues. They worked on a technology-supported learning network that included the child, the parent and a literacy expert, called a Family Learning Coach. The coach analyzed the digital traces of the child's interactions with SpeechBlocks and: (1) sent parents updates about the child's play to encourage a supportive learning environment at home, (2) suggested short contextualized activities and (3) sent the child suggested words that s/he might be interested in building. While the questions asked by my colleagues were about how the role of the coach should be constituted (first home study) and whether coaching has an impact

[26] http://madlibs.com

on children's engagement and parental awareness of their literacy progress (second home study), I saw these studies as an opportunity to gain insight into the performance of SpeechBlocks in the home setting, with children of different ages, and as a part of a learning network. Thus, my research questions piggybacked on the studies. Nevertheless, I took a significant part in their preparation. The results from the first study have been published (Hershman et al., 2018.; Nazare et al., 2018). Publications on the second study are forthcoming.

In these home studies, the participants played with SpeechBlocks on devices provided by the researchers. The devices had other apps removed, and access outside of SpeechBlocks restricted by a parental control app called Kids Place. In the first home study, there were 16 participants from 9 families. Their ages ranged from 4 to 10 years old, with the majority of participants being between 6 and 8. This age range was not optimal for my purposes, since many children of that age already have well-developed phonological awareness. However, it was selected because my colleagues and I were concerned that younger children would not be able to use SpeechBlocks I effectively without adults' scaffolding. The study lasted for 10 weeks. In the second home study, 64 participants, ages 5-8, were recruited and placed into two conditions: with or without the literacy coach. Participants used SpeechBlocks at home for at least 8 weeks (with some participants having the devices for a few weeks longer).

5.2. Types of SpeechBlocks Play

The open-ended nature of SpeechBlocks led children to use the app in a variety of ways. I distinguished five prominent play types:

- **Remixing and Rhyming.** This form of play focuses on transforming some words into others. Younger children primarily created nonsense words via this activity, by concatenating real words. The play of older children was at times more sophisticated. They rearranged words into other meaningful words, and created series of rhyming words.

- **Word Crafting.** This activity focuses on the creation of real words, or multi-word compounds. Younger children were able to create very few words on their own, and those they created were primarily names. They focused on copying words of interest from such sources as the above-described cards. At home, children created a variety of words and phrases related to their everyday experiences, events in their lives and personally meaningful subjects. Some children actively employed invented spelling. Furthermore, a group of children iterated on their spellings, trying to make the words sound correct.

- **Proto-Narrating.** This activity goes beyond simply making words and phrases by using them to tell a simple story. The story can be told verbally, or within SpeechBlocks. In the latter case, stories are typically very simple - sometimes just thematically related collections of words conveying certain images. They can describe imaginary situations or recount

events in children's lives. In the classroom study, children narrated verbally, while in the home studies, some children produced written proto-narratives.

- **Communicative Play.** This type of play was observed only in home conditions. Its goal was to make the app "speak" for the child. The words spoken typically communicated a message for someone: a greeting, an expression of the child's feelings, or an opinion. Children even used the app to give commands to their pets and to tease their siblings. They also arranged lines out of songs and made SpeechBlocks "sing" by tapping on words in succession.

- **Using SpeechBlocks as a Reference.** In this activity, the app itself acted as an auxiliary material, in two ways. First, children enjoyed copying words from it onto paper. Second, they used the app to sound out the words they didn't know how to read, but could reproduce in SpeechBlocks letter-by-letter.

- **Impulsive Exploration.** This is an unsophisticated play type which focused on "probing" the app via haphazard interactions, without much system to it. I first conceptualized this type of play while studying data for SpeechBlocks II, and applied it to SpeechBlocks I retrospectively. The reader can find the description of impulsive exploration with SpeechBlocks II in the section 6.2.3.

Let us now have a detailed look at these play types.

5.2.1. Remixing and Rhyming

Remixing words from the word drawer was usually one of the first activities children tried with SpeechBlocks. Most often, they concatenated several words to obtain results like CUPEAR, BALLCAT and ZOOBALLBALL. Such nonsense words had a strong comical effect. Children laughed and made exclamations like these:

- (laughing) "ZOOBALLBALL! What's that like?" (later) "ZOOBALLBALL – that's what I say!"

- (hearing a word made by peer's phone) "CUPEAR!" (laughing) "Can I see? Can I see how you spelled CUPEAR?

This initial comical effect, achieved at the cost of nearly no effort, likely helped children to feel at ease and comfortable playing with the app. Thus, it appears to be a good initial step for the players.

For young children in the first study, remixing was limited to nonsense word creation. For older children, some amount of remixing words into other meaningful words was observed. Some examples are BARACK -> BE ROCK; KAMALA -> COME MALL; RAINBOW -> BRAIN; JESUS ->

JESSE. These cases are reminiscent of wordplay in poetry or rap. There were also interesting cases when a part of the initial word was retained, and the remainder was built anew. The examples include FIREMAN -> FIREWOMAN; BALL CAT -> BAT; HOTPOT -> HOT POT -> HOTALL; BATMAN -> BATMOM -> BATDAD -> BATKID -> BATDOG -> BATMOBIL.

However, remixing that goes beyond mere concatenation is quite rare. In the home studies, only six out of 80 children engaged in remixing systematically (with five or more cases of wordplay), and there were just about 90 total cases - slightly more than one per child. In general, children rarely decomposed any words. For instance, in the classroom study, we found that children utilized the merge functionality 8 times more frequently than the split function, and the average ratio of merges to splits in the home studies was about 5 to 1. Furthermore, most of the splits were generated in the process of self-correction, when children were rebuilding words that did not sound as expected. Only a few children split words in order to either explore how the parts sounded, or in order to create building material for other words, and each child did that only a few times.

Although we do not know the exact reason why splits were less frequently used, we can see several possible causes. First, the limitations of the SB version used in the pilot may have made splitting less convenient than merging: A child's finger could accidentally touch a neighboring block, leading to a split happening in a wrong place. Second, children seemed to be interested in building long words, as evident from their exclamations, such as: "I am making such long words!" and "I made the longest word ever!". Since splits lead to shorter words, they might have been less interesting for children. Third, children may have perceived splits as undoing their work, rather than creating something new.

The limited amount of splitting activities in SpeechBlocks play might be a missed opportunity. Such activities can be helpful for children in understanding the morphological and phonetic structure of the remixed words. Currently, there is a debate among researchers on whether "analytical" or "synthetic" approach (decomposing words or composing them) works better for phonological awareness learning (Castles et al., 2018). It is possible that a combination of both is desirable. Breaking words apart introduces an "analytical" dimension into SpeechBlocks. In the future, designers might consider how to encourage it.

In addition to remixing, a few children experimented with rhyming. Some of this activity was prompted by the coaches, so I only counted instances of rhyming that children appeared to do on their own. Twenty three out of eighty children engaged in this activity, and eleven of them did so regularly (with at least 5 rhymes made). A couple of children leaned heavily into rhyming: one produced about 200 rhyming words. In total, I found about 450 children-initiated rhyming words. Rhyming was done differently by different children - in some cases, each rhyming word was spelled anew, whereas in other cases children replaced the first letter of a word. Examples of rhyming include DOVE-LOVE-HOVE-DOVE-PUOVE (visual rhymes), LOP-HOP-SHOP, BALL-TALL-CALL and CAT-BAT-FAT-MAT-SAT-HAT-RAT. Various subsets of the latter two sequences appeared in play of many children.

5.2.2. Word Crafting

Word crafting focuses on creating real words with SpeechBlocks. In the classroom study, the first real words that children attempted to create were usually names. Children didn't always get them right on the first try, but they persisted through attempts to build the names for many days, and their spelling gradually evolved towards conventional spelling of their name. For instance, for a child named Addia[27], the evolution of name spellings appeared like this: DD, ADD, ADID and DDAA (day 1); ADD, ADDI and ADDIA (day 4); AGIHD and ADDE (day 5); AWI (day 12); and finally 4 occurrences of ADDIA on days 11 to 14. In that study, upon completing their names, children often combined them with something else: e.g. BUZZALEX (where BUZZ was Pixar's character Buzz Lightyear). The child who spelled this proudly explained to us that his word meant "That I am Buzz".

Names were also a popular subject in the home studies. Children wrote their own first and last names, as well as names of their friends, parents, siblings and other relatives, and (likely) teachers (e.g. MRGRAY). In addition, children made a number of words denoting relationships, such as GRAMY (grammy), DADY (daddy), BROTR (brother), SITR (sister), MOMMYANDDAD.

Another subject that caused much excitement in the preschool study was cartoon characters. Children's pronounced interest in the characters was discovered accidentally. During the very first day, one child concatenated two words ZOO from the word drawer into ZOOZOO and, upon hearing it back, excitedly exclaimed: "This is Zazu from Lion King!". Another child responded, and the two had a conversation about Zazu. Subsequently, the duo built ZOOZOO during every session. It was this persistent interest in making Zazu that suggested the idea of character cards. The cards received an extremely warm welcome from all children except one, and dominated the play after their introduction. Usually children picked three to five cards at the beginning of each session, arranged them in a row in front of themselves, and went through spelling the related words in order. Action cards were also actively used after their introduction. We will look at how children used action and character cards together in the next play type, Proto-Narrating.

In contrast to the character and action cards, articulation cards didn't see much use: only 10 words were constructed using them. There are two likely reasons for that. First, the words on the articulation cards were generic and didn't cause any visible excitement, as opposed to the character cards, which speaks to the power of personally meaningful items in expressive play. Second, children had difficulties matching the lowercase letters on articulation cards with the uppercase letters (which were also written in a different font) on the blocks and were asking adults for help. The SpeechBlocks-like font on the custom-designed cards simplified this process significantly and allowed children to build words autonomously.

Just as children in the first study were excited to build cartoon characters, children in the home studies built a variety of words around entertainment, their interests and hobbies. Examples include

[27] This and other names are fictional, and the spelled words are adjusted accordingly

JKROWLIN (J. K. Rowling, author of the Harry Potter book series), BRNOMARS (Bruno Mars, a musician), ZAC EFRON (Zac Efron, an actor and singer; the parent commented: "She is all about *The Greatest Showman* right now), JOKER, BANE, VENOM and LOKI (comic villains), SUPERFAST and SUPERSTRONG (attributes of superheroes, located next to superheroes themselves in the play), HULU and ENTFLIX (Hulu and Netflix, online streaming platforms), SAVANNAH (name of a favorite TV show, according to the mother). A series of words by one child was likely inspired by the How to Train Your Dragon cartoon: DRAGONFIRE, FIREDRAGON, WATERDRAGON, DRAGONGOLD; another child spelled names of trolls from a popular child series. Games and sports were also present, such as SOCCER, FUBOL, FOOTBALL, TENNIS. One child, who apparently did cheerleading, made a number of words related to the sport: CHEER, ALLSTAR, COMPETE, BUZZER. Hobbies were represented too, by such words as ROCK COLLECSHUN.

In the first study, only 10 words were inspired by children's surroundings, day-to-day activities, events in their lives, etc. Examples include: LOV (love), MOM, POPCORN, BNGA (bang), and SINK (an item that could be observed in the classroom). There were several cases when children verbally stated that they were going to make a word related to their experience, but didn't proceed. In all likelihood, this reflected children's limited ability to build words without scaffolding and the limitations of the human-provided scaffolding for such a purpose. We will return to this question in section 5.6.

In contrast, children in home studies created a great variety of words and phrases pertaining to their surroundings and day-to-day experiences. One prominent topic was food, represented by such words as STEU (stew), SPGETE, PAPRRONE, PARONEY (pepperoni), MCDONLS, STARBUCKS, HOTPOT, and CINNAMON TOAST CRUNCH (spelled at 9am, during the child's breakfast time). Clothes-related words also appeared, such as DRAS (dress), BOOTS, CAPE, BELT, SCARF. Two holidays - Valentine's Day and Easter - occurred during the study, and made their way into children's play: EASTR, ESTREGG, VALNTIN, VALTINS. Children also replicated a variety of environmental texts. In the logs, we can see brands such as KFURIG (Keurig), PURELL and MARLDORO (Marlboro). Locations appear as well, such as QUEENPLACE, VIDANT MEDICAL CENTER, DADS WORK, THEBANK, ASHVILLE (place where their family is originally from, according to parents). One child spelled a series of inspirational messages, such as LOVEWELL, COMPASSION, LISTEN, ASKFORHELP, TODAYISANE ("today is a new...", incomplete). His mother reported that he was copying plates that hang on the walls of their house. Another parent reported that her child was spelling road signs on a road trip, and yet another child spelled STOPSIGN. Realities of adults' political lives also made their way into SpeechBlocks play, in such words as HILLARY and DONLD (Donald) TRUMP (which appeared during the 2016 election campaign). Some words stemmed from children's activities at school. For instance, one child's class had special activities related to the Olympics, and the theme of the Olympics made its way into the child's play.

One curious type of word occurring surprisingly often in the SpeechBlocks data is long and unusual words, such as ONAMANAPIA and ONOMATOPO (onomatopoeia), MULTIBULICATION, ESOPHAGUS, ELASMOSAURUS, ENTOMOGOGY, METEOROLOGI, ORNITOLOGY,

PALEONTOLOGIST, NINEDEY (ninety), SIXDEY (sixty), SUPERCALIFRAGIL (a reference to Mary Poppins), etc. Themes such as astronomy - PLUDO, PLOODO, ERTH, MGRE (Mercury), VENS (Venus) - and geography - ANARCTICA, CONNETICUT, MASSACHUETTES, BEIGING (Beijing), DENMARK - were prominently manifested. Many of these words are spelled with characteristic invented spelling patterns, suggesting that children worked on the words on their own. Other words are spelled correctly or with small mistakes, suggesting that children might have copied them from such sources as textbooks. The presence of these words likely reflects children occasionally challenging themselves to write complicated words, similarly to what (Bissex, 1980) reported. The drive behind such activity is likely the sense of self-efficacy associated with success in a challenging venture.

If 4-5 year-olds appeared mostly overwhelmed by the challenge of spelling arbitrary real words, many 6-10 year-olds found a way of tackling this problem. In their play, we see invented spelling flourishing. Nearly every child in the two home studies attempted it at least once, with the median amount of invented spelling attempts being 6, and the maximum amount being 133. In total, almost 1100 examples of invented spelling were found. Table 5.1 shows a few notable examples of invented spelling in SpeechBlocks. One can see that the patterns of invented spelling in SpeechBlocks align well with the patterns reported in literature.

Table 5.1. Some invented spellings observed in SpeechBlocks play

word	interpretation	comment
FON	phone	phonetic encoding
CUPCACK	cupcake	overgeneralization of CK; A's name used for its sound
MARCKERS	markers	overgeneralization of CK
GURMS	germs	UR pattern for *[ɜ˞]* used similarly to FUR, OCCUR, NURSE
KG; KEI	king (inferred from final result)	classic case of inv. spelling with only initial and final sounds; using E's name for its sound (adding I was possibly an attempt to make *[iŋ]*]
MOME; CANDE; BABE	mommy; candy; baby	using E's name for its sound
ORAGE; FEGER	orange; finger	omitted nasal
SNAC THAC	snake; thank [you]	C used for *[k]*; A's name used for its sound; omitting nasalization in the case of THAC
GOODNAT	goodnight	A used to represent *[aɪ]* sound, because the first sound of the diphthong is the *[a]* vowel that often corresponds to A

Table 5.1 (continued). Some invented spellings observed in SpeechBlocks play

word	interpretation	comment
CIN; KIN	kitten (inferred from final result)	a case of simplistic invented spelling: initial and final sounds + 1 vowel are represented. C used for *[k]*
WHEEOL	wheel	representing the ghostly vowel that seem to appear after *[i]* when speaking wheel
TRUKE; TOWNE; WATRE	truck; town; water	overusing the silent E at the end of the word
TWRL	twirl	omitting vowels, or letting R represent the vowel
TROOTH; PLOODO	truth; Pluto (inferred from astronomy theme)	using OO - *[u]* pattern similarly to BOOTH and TOOTH; alternation between T and D
PARONNEY	pepperoni	omitting the unstressed syllables
PAPRRONE	pepperoni	A for *[ɛ]* (because letter name starts with this sound); E's name for its sound; representing *[ɝ]* as R
OLVEIA; JAODEN; DITEY; RADEY; VENUISE; FAIVER	Olivia; Jordan; dirty; ready; Venus; fever	various experiments with representing vowels
SPGETE	spaghetti	omitting vowels in unstressed syllables; using E's name both for its sound and *[ɛ]*
COLLECSHUN	collection	using relatively frequent letter-to-sound patterns SH - *[ʃ]*, U - *[ʌ]*, N - *[n]*, instead of the rare one TION - *[ʃʌn]*.
COLLD	cold	overusing double letters

Some children repeatedly tinkered with some words in order to make them sound right. In total, about 350 such tinkering sequences were found. Twenty six children (out of 80) engaged in this activity systematically (five or more tinkering sequences), with one child producing 43 examples of tinkering. The longest tinkering sequence included 10 repeated attempts to make the same word. Below are some of these sequences of attempts:

DRTY, DITY, DITE, DITEY - in an attempt to spell DIRTY;

MISALFY, MISELFY, MISELF - in an attempt to spell either MYSELF or MY SELFIE;
BECUS, BECUSAE, BECASUE, BECASHE, BECASE and finally BECAUSE (it seems that here, the child might have tried to involve visual knowledge of how BECAUSE is supposed to look)
PRETTE, PRITTE, PRITTY - in order to spell PRETTY;
BROTTR, BROTHRE - in order to spell BROTHER;
SISTHR, SITR, SISTRE - in order to spell SISTER;
RADE, REDEY, REDE, RADEY - in order to spell READY;
JODEN, JOAD, JAODEN, JRAN, JOANDON, JORDEN - in order to spell JORDAN;
MOWN, MOOWN - in order to spell MOON;
DRIS, DRAS, DRASS - in order to spell DRESS;
TAWELE, TOUWLE - in order to spell TOWEL;
etc.

A note needs to be made as to how I identified invented spellings and their interpretations. I used both the appearance of the word and its context in its play session (e.g., was the child building furniture-related words that day?). If the child iterated on the spelling of a word, I could use more refined versions of the same word to infer the meaning of less refined ones. This introduced a certain selection bias: unfortunately, there are very few examples of simplistic invented spellings (e.g. spelling CARROT as KT), because it was difficult to distinguish such spellings from nonsense words, much less interpret them, unless some lucky contextual source provided help. I attributed misspelled words to the invented spelling category if it was plausible that the children came up with the spelling based on their developing phonological and orthographic knowledge - as opposed to, for example, visual memory of the word.

The reader might ask whether these spellings represent children's independent efforts. This is quite likely, because we observed that whenever parents guide their children, they use conventional spelling, and the same is true when children copy words from a print source. This means that 6-10 y.o. children tend to actively tinker with invented spellings. This is quite close to the originally envisioned role of SpeechBlocks as a tinkering tool. But the learning potential of such activities is limited by the fact that children only engage in them at the age when they already have pretty well-developed phonological awareness skills. Still, given the existing evidence of the positive effect of invented spelling on this skill, practicing invented spelling in SpeechBlocks gives children an opportunity to further develop and solidify phonological awareness.

5.2.3. Proto-Narrating

On some occasions, children didn't stop at building words in SpeechBlocks, but used these words to tell a simple story with them or about them. Almost all of these proto-narratives were extremely simple: a mere phrase, sentence or a collection of thematically related words that created a certain impression. Lacking plot, they were not stories in a strict sense of the word, but they did nevertheless convey a certain image. Proto-narratives were told both orally and in SpeechBlocks, and were both planned and serendipitous.

Four-to-five year-olds in the classroom study actively used action cards to describe what their cartoon characters of interest were doing. However, only once did this process result in a complete sentence. In the rest of the cases, children either built and deleted words one-by-one, or only built one word and produced the rest of the sentence orally. When sentence cards were introduced, about a third of the children copied sentences from them. Children also came up with imaginary situations involving words they made in the app. Sometimes that happened even with unplanned words. For instance, one child made BALLCATTOYZ, and the following monologue emerged:

- (Repeating after the speech synthesizer) “BALLCATTOYZ... Wait, I spelled toys?!” (Clearly surprised by this . Then, speaks to another child) “Oh, you know what? You know why I spelled it? Because a cat is playing with ball toys – that’s why I spelled it.”

However, almost all cases of verbal narration were observed when children were making real words, particularly when using character cards. The fact that real words were preferable for narration is perhaps not surprising.

In addition to talking about imaginary situations, participants also connected the words pronounced by SB to their life experiences. For instance, a child could say “I have a dog” after hearing someone’s phone pronouncing DOG. In many cases other children responded, resulting in several turns of conversation.

In the play of 6-10 year-olds in home conditions, we can see accounts of the child’s own life, for instance: LIKETO LEARN RUN PLAY EAT; EIGHT BIRT PARTY AMME (on the child’s eighth birthday) and FEELBAD, DOCTORSHOT, COLD, SICK, FLU, GURMS. While the latter one is simply a collection of words, it nevertheless conveys a poignant image. Other mini-narratives may describe an imaginary character: SHELIKES FANSYCLOE PRETTYHAIR NICE SHOES PLASES FASHEN.

5.2.4. Communicative Play

This type of play pertains to children making SpeechBlocks “speak” for them. It was only observed in the home studies, where the environment seemed to provide children more occasions for communicative use of the app.

Some of the communicative constructs that children made were merely descriptive, such as I SEE MY DOG. Some were relational, such as BESTFRIENDS, BECAUSE I LOVE YOU, LOVEYOMOM, LOVEYOU and IDONOTL. The last one is apparently a beginning of “I do not love you”, which seems to be used playfully, as it appears right after LOVEYOU. Children used SpeechBlocks to convey their emotions or opinions, such as IAMSOBROD (I am so proud), IWISHIWAS, IHOPE and IDKARE (I don’t care). There were conversational elements, such as DOYOUNOIF (do you know if), THANKYOU, SEEYOU, and a few colloquialisms, such as

COOLDUDE and PEACEOUT. Interestingly, SpeechBlocks was once used to talk to the coaches: HELLO SPEECHBLOCKS WORKERS. SpeechBlocks was used for holiday wishes, such as MERRY CHRISTMAS, HAPPYEASTR (Happy Easter) and HAPPY BIRTHDAY TOO YOU. One child used SpeechBlocks to give commands to his dog: CODYSIT, I SAID COME OVER, OPENTHEDOOR. SpeechBlocks was also apparently used to respond to requests: PICKED IT, HOLDING. In some cases, children's communications were whimsical, for instance SHINE BRIGHT, BYE BYE BIRDIE, FLYBUGFLY and WELCOME EARTH (welcome to Earth). Alas, children also used SpeechBlocks to tease their siblings, e.g. TAMBIGBUT (Tam's a big butt). This particular phrase was tapped dozens of times!

An interesting variation of this form of play, albeit less related to communication with another person, is making SpeechBlocks say something interesting or funny. For instance, children made the app "sing songs" by spelling a song line and repeatedly tapping on it. Examples include BEBEUGON ("baby, you gone" - likely referring to a song by Brian Adams), THIS IS MY FIGHT SONG TAKE BACK LIFE (mimicking *Fight Song* by Rachel Platten), WATCH ME WHIPNAENAE (imitation of *Watch Me* by Silento), OPPAGANANG (imitation of *Gungnam Style* by Psy). They also arranged funny-sounding words, such as TRALALA, BUBDEEBALL, JOYJOYJOY, FEEFEE, LOOPNOOP, COPFOPE.

5.2.5. Using SpeechBlocks as a Reference

In this form of play, SpeechBlocks itself acted as an auxiliary material, in two ways. First, children enjoyed copying words from it onto a sheet of paper. Second, they used the app to sound out words they didn't know how to read.

Most children in the classroom study copied their creations into the journals to preserve them. However, several children turned copying into an activity in its own right and dedicated entire sessions to it. Instead of only writing down the words that they had made, they copied the contents of the word drawer. As they weren't able to read yet, they used the speech synthesizer to find an interesting word to copy. The process of writing was also new to them, so they carefully redrew the shapes of the letters - an activity that seemed to tap into children's Zone of Proximal Development for knowledge of letter shapes. Often, the letters they wrote weren't arranged in a conventional way, but scattered throughout the page. That didn't prevent children from proudly saying: "Look, I wrote X!" Children's self-initiated engagement in it may have been inspired by the overall spirit of free play surrounding SpeechBlocks activities.

SpeechBlocks also turned out to be handy as an auxiliary tool for reading. For instance, in the classroom study, a child pointed to a character card and asked: "Who is it?" The facilitator suggested that she spell the word to figure it out. She copied the inscription on the card letter-by-letter, and when the synthesizer pronounced the word, she exclaimed with the joy of discovery: "This is TOTORO!" Another notable episode of this type was relayed to us by a child during the post-study interview in one of the home studies. He said that he saw a word written on

a car, and started to spell it in SpeechBlocks to figure out what it was. Upon building TAX, he realized that the word was TAXI, but was motivated enough to finish it. Out of everything that he created in SpeechBlocks, TAXI ended up being the most memorable, and the one he was the proudest of.

5.2.6. Impulsive Exploration

While the sections above describe focused, "minds-on" activities with SpeechBlocks, there was also a large amount of impulsive interactions with the app. Children performed random taps and swipes, cluttered the canvas with words from the word bank to see how many they could fit, built entirely random strings of letters to say "Look! I made such a long word!", etc. In one play-testing session (which occurred outside of the context of four studies), a child used the red (vowels) and blue (consonants) blocks on the app's canvas to play a "soccer match" between the red and the blue "teams". In a later study with SpeechBlocks II, such impulsive behavior seemed to be indicative of children finding the core SpeechBlocks activity of making words too difficult to be enjoyable. While I don't have sufficient data to support a similar claim for SpeechBlocks I, it appears plausible.

5.3. Agency, Self-Efficacy and Ownership of Work

One of the most valuable aspects of the children's experience with SpeechBlocks was their agency and their senses of self-efficacy and ownership of their work. In the examples of play described above, children's agency manifested in their ability to pursue a variety of self-chosen and personally meaningful activities. The picture of agency and self-efficacy is complemented by such phenomena as independent goal-setting, planning, challenging themselves, displays of their work, and desire to preserve it.

One manifestation of learner's agency was the ability to freely choose goals for oneself and execute them. During the classroom study, the evidence of such behaviors was frequently observed in children's self-directed speech. For example: "I'm going to make Simba. I need S, I, M... I'm writing about Lion King today." Such phrases were almost always unprompted: children naturally talked about their plans and goals. As they were making their plans, I could see noticable excitement and anticipation.

Setting up a goal and being able to accomplish it naturally resulted in a sense of self-efficacy for the children. In the classroom study, this sense was highlighted by students frequently drawing attention to their completed works, for example:

- "I spelled LALLA! That's what my name says."
- "I can spell ZIVVY with mine [phone]!"
- "Look! I spelled BUZZ!"
- "I made so many words!"

Another way in which the sense of self-efficacy manifested itself, in both classroom and home studies, was children challenging themselves to build words that were difficult for them. In the first study, these words were names, while in the home context, we saw tinkering with a variety of challenging polysyllabic words. These occurrences were described in section 5.2.2.

Upon completion of a challenging task, children naturally wanted to preserve the result. In the classroom study, once the journals were introduced, children started to ask us to write down most of the words that they made, and later started to do so on their own. They also inquired whether they could take the journals home, and whether they could keep them. One child even asked: “Can we keep them forever? Until the end of our lives?” At the end of the study, we gave the journals to children and also compiled small booklets out of the words and sentences they made. We also gave children similar booklets at the end of the home studies. In both cases, children were delighted to receive these gifts. In the last two studies, another indicator of this sense of ownership was the large number of words saved in the word drawers.

These observations raise a hope that interactions with SpeechBlocks could help children to establish a more empowering and more personal relationship with literacy.

5.4. Engagement

SpeechBlocks exhibited a capacity to engage children, but this engagement seems to be dependent on scaffolding. In all the studies, children exhibited significant initial interest in SpeechBlocks, which was associated with the freedom of play and fueled by the fun of building nonsense words. However, this activity started to exhaust itself after several days, with a corresponding drop in engagement. In the classroom study, we attempted to introduce structured activities to counter this, but they only exacerbated the situation. Engagement was restored after character, action, and sentence cards were introduced, and remained high afterwards. In the home studies, engagement continued to drop at quite a high rate. A plausible reason for this dynamic is the combination of “high floors” (in the absence of word-building scaffolding, doing sophisticated activities required a lot of skill and effort) and “low ceilings” (limited expressive capabilities of the app).

5.4.1. In a Classroom

SpeechBlocks received a very warm reception during the initial days of the classroom study. Children frequently laughed, produced enthusiastic exclamations, and exchanged delighted remarks. Teachers told us that after the sessions, several students continued talking about the words they had made in SpeechBlocks. On the second day, upon learning that they were going to play with SpeechBlocks again, one child exclaimed “yay!” This initial excitement can be almost

entirely attributed to the fun of making nonsense words by remixing existing words in the drawer, which was described in the section 5.2.1.

The high entertainment value of making nonsense words lasted for two to three sessions, and then gradually started to recede. Children started to laugh and exchange comments less frequently. They began to look away from the app more often, and the impulsive exploration mode of play intensified. While the nonsense-word-making began to exhaust itself, it didn't naturally give way to crafting (or attempting to craft) real words. This was likely because, for most children, such activity was far beyond the limit of their current skills. In light of this situation, we attempted to introduce new activities, such as themed word building, and Mad Libs, but they only exacerbated the dynamic. The disengagement reached its peak when we tried to organize the Mad-Libs-like activity. While a few children enjoyed the activity, several children explicitly said that they were bored, and three out of sixteen children left the station prematurely. Several children had specific, real words that they wanted to spell in order to insert them into the story, and tried to do so with the help of the researcher who led the activity. However, since the adult's attention was limited and often occupied by other children who addressed her, these children ended up waiting for most of the session, and looked somewhat frustrated. A few other children simply continued the nonsense word making. One of them asked: "Can we just play?", exhibiting a preference for free play. At the end of the activity, when it was time to read the Mad Libs story, the researcher had a hard time recruiting the children's attention.

Engagement quickly recovered when we introduced the character cards. Children now had an activity that was well aligned with their interests and required some degree of effort, but was within their capabilities. Their focused efforts were noticed by the teachers. Once a teacher who was passing by the station table noted the remarkable degree of focus of a child who was known to her as very distractible. When she saw the words that the child had made, she was very impressed, noting that she didn't expect the child to be able to craft such complicated words on her own. While anecdotal, this case illustrates the power of internally motivated learning combined with proper support structure.

Several other instances of anecdotal evidence illustrate children's engagement with the app. First, children tended to repeat words after the speech synthesizer, demonstrating their close attention to the app. Second, after the study had officially ended and we didn't plan to have SpeechBlocks sessions anymore (we came in for a wrap-up session), multiple children requested to play. Once they received the phones, they started self-initiated and highly engaged play with SpeechBlocks. Third, the researchers became known to one child as "the SpeechBlocks people", and she was excited to introduce us to her parents in this way upon meeting us on the street.

5.4.2. At Home

It is a bit more difficult to provide a detailed assessment of children's engagement with SpeechBlocks in homes, where children could not be observed. During the initial session, when

SpeechBlocks were introduced, children were just as excited as during the first session of the pilot study. During the post-sessions, the children whom we interviewed spoke positively of SpeechBlocks, and eagerly recalled their favorite moments and experiences with the app (although it should be noted that children were aware that the apps were made in our lab, which likely skewed their responses towards socially desirable). Another indicator of children's engagement was their persistent efforts to make the words sound right through repeated tinkering. However, several children found the app boring (according to their parents) and quickly stopped playing with it.

Quantitative data also shows that the pattern of engagement with SpeechBlocks was not entirely satisfactory. Let us look at the joint engagement dynamic for children in both studies (since the dynamic was very similar between them) during the first eight weeks of their play (because every child had the device for at least that long). We can see that while, during the first week, children accumulated more than an hour of play time on average, by the eighth week it dropped to about five minutes (Fig. 5.3, a). Moreover, we can see that starting from week 4, on every given week, more than half of the children didn't play at all. Further breakdown shows that this decrease is mostly associated not with the amount of play during the days when children were active (Fig. 5.3, b), but in declining play frequency (Fig. 5.3, c). By week 8, a median child played less than once in three weeks.

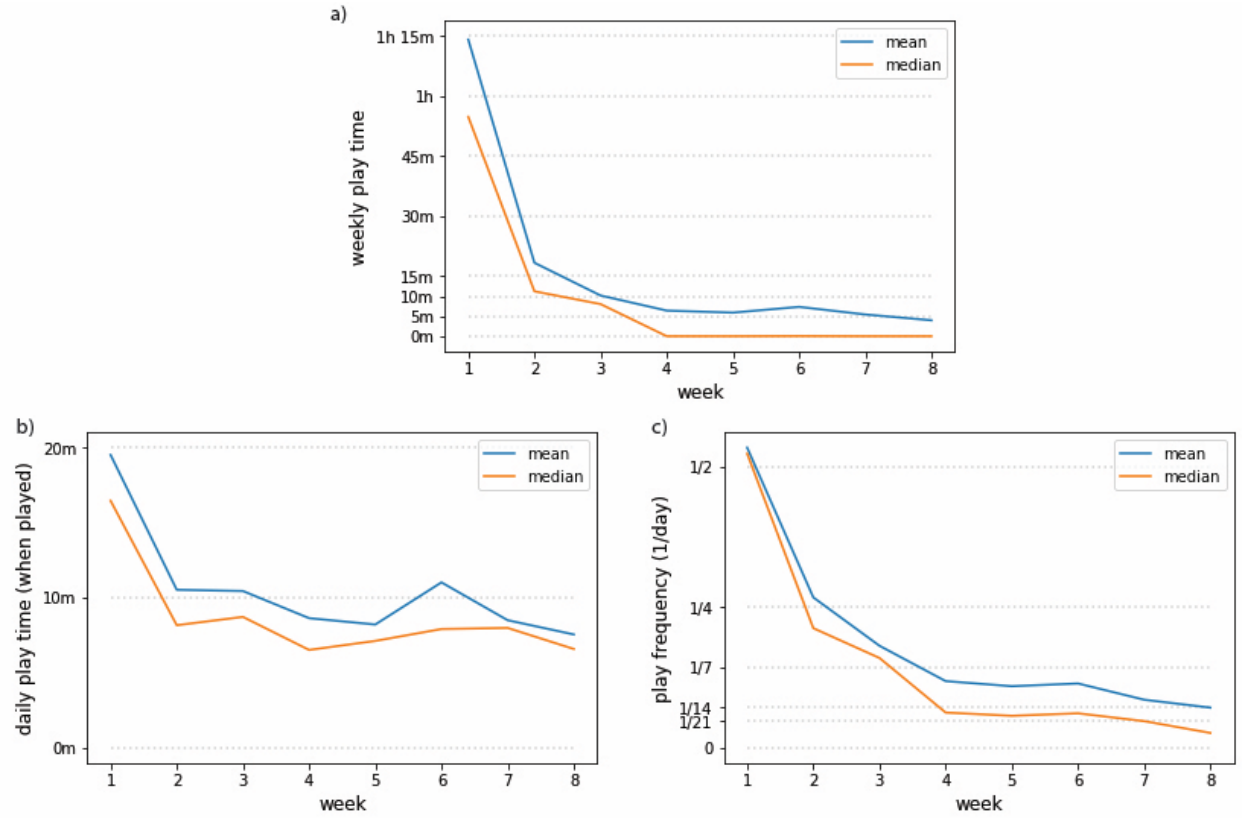


Fig. 5.3. Engagement measures in the home studies: (a) weekly play time, (b) daily play time (for those who played), (c) play frequency

This dynamic resulted in the median cumulative play time being about 2 hours. Fig. 5.4 shows how play time was distributed among children. We can see that there was a group of children, about a tenth of the total number, who were “power users” of SpeechBlocks. Their high cumulative play time is related to a much less steep decline in their play frequency: by the end of the study, they still played on average about once in five days, with a few still playing about every other day. These children don’t stand out relative to their peers in terms of their age and CTOPP score, and the reason for their higher engagement is currently unknown.

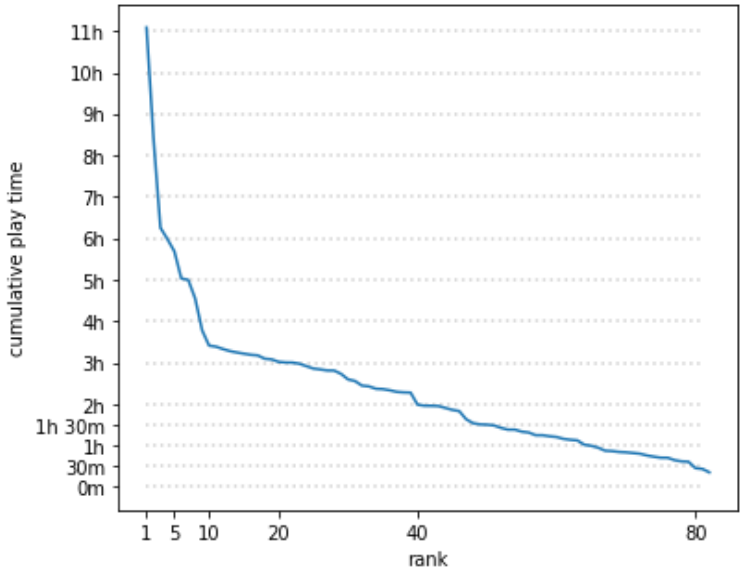


Fig. 5.4. Distribution of play time among children in the home studies

As a brief note, I would like to mention that the presence of literacy coaches, who interacted remotely with children and their families, did increase child engagement. As seen on Fig. 5.5, children in the coach condition consistently played more on average each week, because their play frequency was dropping less steeply. More information about the effect of the coach on play with SpeechBlocks will be available in upcoming publications on this topic.

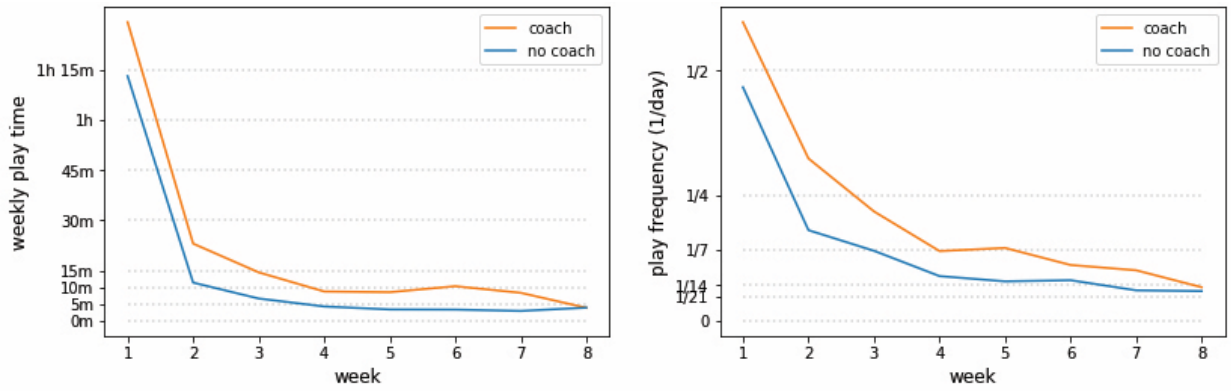


Fig. 5.5. Engagement measures for coach vs. no-coach condition

Why did the frequency of children's play drop quickly? While it is difficult to pinpoint the exact reasons for that, the cause could have been a combination of "high floors" and "low ceilings" (in the words of Papert (1980)) - in other words, high demands on engaging in basic meaningful activities, combined with difficulties of expressing oneself at a sophisticated level. While making nonsense words was a successful "low floor" entry-level activity, it didn't allow for a natural transition to the next level of difficulty: building real words. Although many children in the home studies were able to build real words to some extent, it appears that it was still a difficult and often tedious procedure for them. It is plausible that many of the children found the activity exhausting, and thus didn't want to return to it too often, hence the "high floors". On the other hand, the "ceiling" - the complexity of expression possible in SpeechBlocks - was also quite limited. The limited space on the canvas meant that children couldn't produce anything beyond short sentences, and even those required hard effort, a significant expense of time, and weren't easy to save.

This point of view is indirectly supported by the fact that PictureBlocks (Makini, 2018), a literacy medium which arguably provides a greater expression range, seems to maintain better engagement. In Makini's (2018) study, which had comparable conditions (5-9 y.o. children playing at home), the median cumulative play after two weeks was 203 minutes - about 2.7 times higher than 75 minutes for SpeechBlocks after the same amount of time. A qualification needs to be made that these numbers may not be directly comparable. First, families in PictureBlocks studies were recruited via flyers, MIT study mailing lists and parent groups, and could have been more enthusiastic about play with digital learning technology. Second, it is unknown what fraction of the play time in PictureBlocks was devoted to literacy activities, as opposed to arranging the pictures. Nevertheless, this comparison suggests that a picture-based form of expression could be beneficial for children's engagement with digital literacy media. Additionally, in the second home study, both coaches and several parents suggested increasing expressive capabilities of SpeechBlocks by incorporating pictures.

In the next generation of SpeechBlocks, SpeechBlocks II, efforts were made to both "lower the floors" and "raise the ceilings" by incorporating more sophisticated built-in scaffolding and a scene-making capability to the app. I hypothesized that SpeechBlocks II would be able to better sustain engagement in home conditions, but testing this hypothesis remains a subject for future work.

5.5. Social Play

One of the interesting observations from the classroom study was the children's strong willingness to interact socially around their play with the app. These interactions happened despite the fact that SpeechBlocks had no explicitly designed social features. It is possible that the high level of social interactions is connected to the open-ended nature of play. This aspect aligns the present work with the Peers component of the Resnick's (Resnick, 2017) "Four P's" framework of creative learning. We also saw that hearing utterances produced by their peers' devices raised

children's curiosity about their peers' play. However, this apparently was a less crucial factor than we suspected: in the latter study with SpeechBlocks II, children continued to actively interact despite the introduction of headphones (section 6.1.2). Social interactions that we observed can serve three functions potentially useful for learning: (1) inspiring each other's ideas; (2) maintaining mutual engagement; (3) directly learning from each other. Let us look at which social interactions facilitated each of the three functions.

1. Inspiring Each Other's Ideas:

Children often looked at what their peers were doing while playing with SpeechBlocks. Hearing each other's phones contributed to their curiosity[28], as can be seen in this example phrase: "Can I see? Can I see how you spelled CUPEAR?" Their observations of others likely helped some children gain confidence and ideas for what to do with the app. For instance, one child barely played with her phone during the course of several sessions, but she intently observed others. Eventually, she became an active player herself. Other cases involved borrowing words from their peers. LOV (love) was one such word. When a child spelled it for the first time, a few others heard the sound, and attempted to build it, too, as soon as their turn came.

2. Maintaining Mutual Engagement:

Displaying their work. As described in section 5.3, children were eager to share what they made. But to do so, they needed an audience. Without peers (and also us, the researchers) to fulfill this role, it is likely that they would not be able to enjoy their sense of self-efficacy as much as they did.

Conversation around their works. Participants talked about the words they or others had made in SpeechBlocks. For instance, when one child made DOG, another responded: "I have a dog". It is likely that the interest of others in the words they made stimulated children's engagement.

Shared play. Pairs of children sometimes invented ways to play together. Such joint play was exciting for them, and stimulated mutual engagement. For instance, one child made a nonsense word WCAT and showed it to a peer. The peer by chance spelled VWCAT at the same time. She exclaimed: "You did the same thing as me! You have to do exactly the same thing". She then bent over her friend's screen to help her find letter V in order to match her word. When the letter was found and added, she said: "Another V! Let's add another V!", and each of the children did so on their own phone. They continued to make words together for some time afterwards.

[28] However, note that the later study with SpeechBlocks II suggests that hearing each other's phone is not a necessary condition for flourishing social interactions.

3. Directly Learning from Each Other:

Children asked each other questions such as: “Oh! Do you know where T is?” (where to find it in the letter drawer) or “Can you help me to spell your name on my phone?” In such cases, their peers did indeed provide assistance. Such assistance can aid the helper in solidifying her/his knowledge, while providing scaffolding for the child being helped.

Although I haven’t directly compared SpeechBlocks with more conventional literacy apps, I find it likely that the open-ended, expressive nature of the app contributed to the above-mentioned social interactions. In a conventional setup (e.g. in a game), a player typically has well-defined, individual-oriented goals. Such an individualistic setup likely provides less motivation to pay attention to others’ play. The goals of the game are not self-imposed, and accomplishment of them is typically more mechanical, which likely is less stimulating for the sense of self-efficacy. The reduced sense of self-efficacy, and the fact that children’s accomplishments in conventional apps are not connected to their lives, likely reduces motivation for sharing their work. Furthermore, because the outcomes in conventional games are more predictable, there is less reason to pay attention to what others share. The rigid mechanics of conventional games also preclude children from inventing shared play scenarios.

5.6. Scaffolding

The classroom pilot with SpeechBlocks I was heavily scaffolded. First, the scaffolding was provided via materials: character, action and sentence cards; the cards’ importance has already been discussed. Second, a significant amount of support was provided by adult facilitators. During each session, one of the researchers focused on interacting with the children. During a few initial sessions, children mainly interacted among themselves, but they soon started to reach out to the facilitator more and more often. During the latter sessions, the facilitator communicated with children for almost the entire session, and often had to split attention among several children who tried to talk to him/her at the same time. The types of communication between the researcher and the children were:

- Maintaining children’s engagement, e.g. by being an audience. Researchers acknowledged children’s accomplishments when they demonstrated their creations, e.g. by saying: “Wow, that’s a really long word!”. They also tried to encourage conversation about children’s works. For instance, if a child spelled DOG and said “We have a dog too”, the researcher could ask the child about her dog, and then suggest spelling some words to develop upon this theme.

- Mitigating disruptive behaviors of children. Such behaviors didn’t occur very often, but this function was still important for creating a conducive environment for play.

- Providing technical help. The researchers had to fix technical issues that emerged from time to time, or help children if they were confused with the interface. Initially, they also wrote words down in journals upon children's requests.

- Providing literacy help. The researchers provided some general information about the functioning of print. For instance, one child spelled her name backwards and asked the facilitator: "Why doesn't it sound right?" The facilitator used this opportunity to tell her about print direction. Researchers also assisted children in spelling specific words upon their request.

Let us have a closer look at the latter type of assistance, dealing with spelling specific words. Children's spelling requests arrived at a steady, although not very high, rate. On average, we received about one such request per session, meaning on average, each child asked how to build a word once every four sessions. However, the reader should note that such requests required a high overhead for the child: s/he needed to overcome shyness, wait for the adult's attention, and then go through a long back-and-forth exchange with the adult that was frequently interrupted by other children. It is possible that this overhead was responsible for the moderate rate of spelling requests. Indeed, in SpeechBlocks II study, where built-in scaffolding was readily available, children used it very frequently.

The researchers tried not to simply tell children what to do, but rather to guide their thinking so they could gradually develop their ability to spell on their own. As a result, typical scaffolding exchanges looked like the two examples below:

Example 1.

Child: How do you spell CAT?
Researcher: CAT? It has a *[k]* sound in it. What makes the *[k]* sound?
Child: This *(points)*
Researcher: Mhm, and then *[æ]*, *[æ]*
Child (confidently)**:** C.
Researcher: *[æ].* It's the same sound as in APPLE and...
Child 2: A!
Researcher: You have that letter in your name: A sound...
Child: E!
Researcher: There is an E, but that's not what I want. There is another vowel in your name, *[æ].*
Child 2: A!
Researcher: Yea! Where might the A be?
Child (looking at the keyboard)**:** Uhhhhh... I don't know... *(a few moments later)* Look, I found it!

Example 2.

Child: I need I (letter).
Researcher: You need an I? Well, let's look at the alphabet, see if we can find it up here.
Child 2: Hey, look at that one!

Researcher: You found an I? Well, you can tell [Child 1], you found an I.
Child 2: It's right in the corner...
Researcher: What letter it's next to?
Child 2: H!
Researcher: Perfect! She knows H, because H is in her name.

In these examples, the researchers tried to help children think of other words that contain the same sounds, as well as of the association of sounds with letters. They also tried to encourage mutual assistance between children. Even in the case of short words, such exchanges typically lasted for about a minute. During this period, other children were likely to start interacting with the researcher as well, causing the child who tried to spell the word to have to wait. Furthermore, with the adult's attention being divided, the child had to proactively communicate to her/him about all the difficulties that s/he encountered. This was a tall ask for less sociable children. Below is an example from one of the Mad Libs sessions which illustrates these problems:

Researcher (introducing a Mad Libs card)**:** When [the dog] played with his favorite toy, it made a... What kind of sound?
Child 1: SQUEAK! (excitedly jumps in her seat)
Researcher: Perfect! You can do SQUEAK.
Child 1: I don't know how to spell.
Researcher (trying to encourage invented spelling)**:** However you like!

While the researcher hands out the materials to other children, the child pulls out the keyboard, looks at it with uncertainty, and pulls it back up. Meanwhile, the researcher is having conversations with other children, such as:

Child 2: AWAY!
Researcher: Is that the name of the dog? That's a pretty funny name for a dog!
Child 2: That's because he goes away a lot!
Researcher: Exactly, sometimes dogs do go away a lot.

The first child restlessly moves her hands up and down, exhibiting some frustration.

Researcher: Can you sound out his word? You said SQUEAK? What do you think s-s-SQUEAK starts with? s-s-s-s-s-s-s-s-SQUEAK?
Child: S.
Researcher: Here we go! Can you find S on there? To put on SQUEAK!

The child looks at the phone, but doesn't do anything.

Child 3: I spelled THREETURN!
Researcher: You did? What does it say?
Child 3: THREETURN.

Child 1: I've got an S!
Researcher: You've got an S. Here you go.
Child 4: I can't find it! (regarding his own interaction with the phone)
Researcher (assisting)**:** You can't find it? Oh, here they are! Just slide in here...
Child 3: I did THREETURN!
Researcher: You did? That's a funny word.
Child 2: AWAY!
Researcher: That's a good name for a dog.

This interaction lasts for some time. Meanwhile, child 1 is growing restless. She stands up and sits down a few times, rocks in her chair, plays with a sheet of paper in front of her.

Researcher: Did you do your SQUEAK word?
Child 3: I've got THREETURN!
Researcher: What do you think? SQU-u-u-u-u-EAK. SQU-u-u-u-u-u-u-EAK.
Child 2: I spelled BOWAWAY.
Researcher: That is very funny.
Child 2: When he [the dog] comes, he bows.
Researcher: Exactly.
Child 4: I spelled CHEWCAT.
Researcher: That's so funny! That's his favorite toy? A CHEWCAT? He is chewing something.
Child 3: But I don't see WOOF [the sound he wanted the dog to make]
Researcher: How do you think it sounds? w-w-w-w-WOOF.
Child 3: W.

Child 1, after some hesitation, speaks up again.

Child 1: Um, I don't know what's after S...

From these examples, we can see that (1) scaffolding word building in SpeechBlocks was a non-trivial procedure that would require a qualified adult, and (2) the adult's limited attention could be a significant bottleneck disrupting the flow of children's play. The latter issue manifested itself even when there were just four children per one adult in the session. These observations raise concerns about scalability of such an approach.

5.7. Learning

Direct assessment of learning gains caused by SpeechBlocks I is not possible, as none of the studies involving the app had treatment and control groups. A few indirect bits of evidence point in different directions. The early classroom study shows that children had notably high phonological awareness (PA) gains. However, in the second home study, I didn't find any correlation between the time spent in SpeechBlocks and PA changes, even after trying to account for possible ceiling

effects. It is possible that the type of scaffolding provided by facilitators in the first study was essential for children's learning. Alternatively, it is possible that the available indirect evidence simply doesn't effectively capture the learning dynamics that took place.

Analysis of CTOPP scores from the classroom pilot showed an interesting trend: 14 out of 15 children tested showed an increase in their composite phonological awareness score. This score is age-adjusted, meaning that increase in the score indicates PA growth beyond what normally occurs in this span of time. The difference between the pre- and post- scaled scores was statistically significant. While on the pre-test, four children had very low CTOPP scores, on the post-test, all children scored above 50th percentile for their age. This change occurred despite lack of dedicated activities targeting phonological awareness in the classroom in a relatively short period of time (10 weeks). An optimistic assessment of these results is that some combination of SpeechBlocks and a supportive facilitator had an impact on children's PA learning.

The second home study included pre- and post- CTOPP assessments. Although there was no control group in the sense that all participants played with SpeechBlocks, an estimate of learning effect can be made by regressing the play time with CTOPP gains. This regression didn't show apparent patterns. For instance, Fig. 5.6 (a) shows the change in PA composite score regressed against play time, with no visible slope. Given that some of the children were as old as 8 years, it is possible that there were ceiling effects. In an attempt to account for them, I included interactions with age or with initial PA composite score into the regression. I also attempted to exclude outliers. Still, no meaningful patterns emerged. The only bit of evidence suggesting a possible effect of SpeechBlocks is the positive correlation of the play time with the gain in elision score (Fig. 5.6, b). This correlation is significant ($p = 0.01$), and nearly significant ($p=0.052$) after removing an outlier (the child with 11 hours of play). Without removing the outlier, the estimated elision score gain is between 0.13 and 1.03 points per hour of play. However, this might be a spurious correlation that emerged as a result of multiple comparisons. Overall, there is no strong evidence that playing with SpeechBlocks I at home benefitted children's phonological awareness. One possible reason for this is the absence of scaffolding by an adult (or comparable mode of scaffolding) in the home studies.

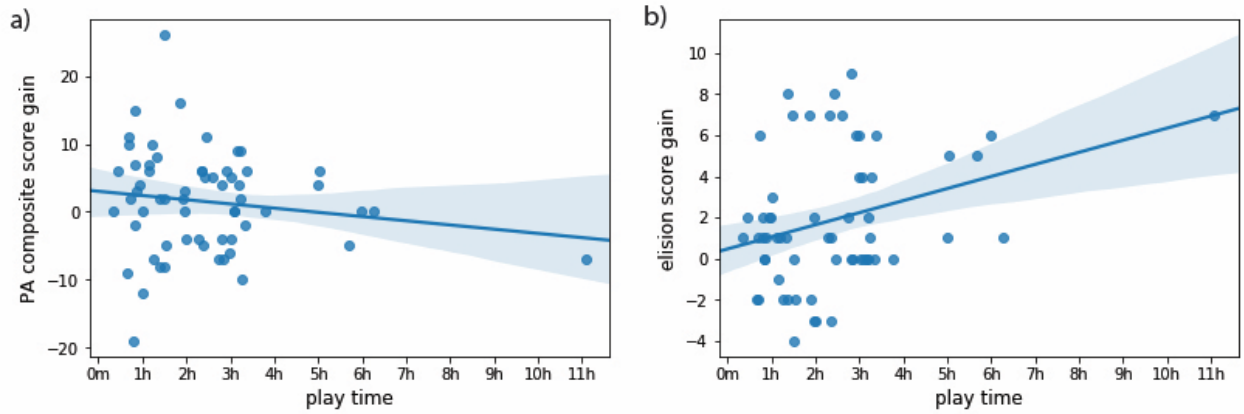


Fig. 5.6. Regressions of phonological awareness gains vs. play time in the 2nd home study

Chapter 6. SpeechBlocks II in Action

This chapter looks at children's experiences with the second SpeechBlocks. Relative to the first version, a number of modifications were made in order to (speaking in terms of Papert (1980)) "lower the floor" (reduce literacy requirements for meaningful play) and "raise the ceiling" (increase expressive capacities of this medium). This chapter focuses on how these modifications affected the play.

With SpeechBlocks II, I saw three primary play types: (1) impulsive exploration, (2) word crafting and (3) imaginative play. Once again, social play was flourishing, but this time there appeared to be more idea sharing and mutual help. Most of the words built in the app were real words built with the help of various scaffolding mechanisms. Numerous nuances about the functioning of these mechanisms that are of interest to a designer were revealed; the details can be found in section 6.5. Various scaffolding systems fulfilled three key roles: (1) responding to specific word requests, (2) facilitating search for new ideas, and (3) acting as a fall-back option in case more sophisticated technology did not work for the child. Because most words were made in the direct guidance mode, the difference between the letter and phoneme blocks turned out to be less important than originally thought. However, the sound creatures (originally designed for phoneme blocks) did turn out to be useful for a sizable fraction of children, facilitating finding sounds on the keyboard. However, for some other children, conventional letters worked better.

6.1. SpeechBlocks II Classroom Study

There was one study conducted with SpeechBlocks II that had three goals. The first was to evaluate various design features of the app, particularly different scaffolding mechanisms. The second was to assess the efficacy of SpeechBlocks for developing phonological awareness. The final goal was to see how children's individual differences affected their experience with the medium.

6.1.1. Study Setup

The study was conducted at five kindergarten classrooms in one public school in the Boston area. Children were between the ages of 4 and 5, with no diagnosed speech or hearing disorders. The school predominantly served low- and middle-SES students of color.

To get a sense of SpeechBlocks II performance, we assigned the classrooms to treatment and control conditions: two classrooms with 24 consented children were in the treatment condition, and three classrooms with 32 consented children were in the control condition. In the treatment condition, we introduced SpeechBlocks (in a way that is described further below), while in the control condition, all classroom activities proceeded as usual. We chose to assign the entire classrooms, rather than individuals, into treatment and control conditions, because of the following

logistical reasons: (1) it made it easier to collect qualitative observations on the maximum amount of children playing SpeechBlocks, and (2) introducing certain activities only for a subset of children within a classroom would likely feel unfair to many children, and thus would disrupt normal classroom operations. The downside of this assignment is that there might be effects on the teacher and classroom that have not been accounted for in the analysis. I considered this problem tolerable, since rigorous assessment of SpeechBlocks efficacy was not the primary goal of the study.

Kindergarten classrooms at the school had a highly structured routine that focused on building early literacy and math skills. A part of this routine was literacy stations—a period of 30 minutes a day during which children rotated between different stations in groups of 4 or 5, performing literacy activities. Two stations were led by teachers and their assistants, while at the other stations children independently performed predefined activities, such as sorting letters. SpeechBlocks was introduced into the classrooms by replacing one of the independent activities at the literacy station. To evaluate how well scaffolding mechanisms could function on their own, and to make the comparison to the control condition more fair, researchers avoided providing any literacy help to the children and tried to let them interact with the apps as independently as possible. Typically, two groups rotated through the SpeechBlocks station every day, making the duration of each session 15 minutes. Each group of children usually interacted with SpeechBlocks twice a week. The main study period lasted for 10 weeks with an additional 3 weeks being granted by the teachers to collect additional qualitative data.

To help the children get familiar with the various design features of SpeechBlocks, each feature was gradually introduced throughout the study period. Most of the sessions started with us demoing a new feature on a tablet and modeling its use. After that, children received their own tablets and started playing.

We ran SpeechBlocks II on tablets, as opposed to previous studies with SpeechBlocks where phones were used. Tablets were chosen to help children see the sound creatures on a larger screen. A key criterion in selecting a particular model of tablets was quality sound, so that children could clearly hear the prompts of the scaffolding systems. Based on this consideration, we decided to use Lenovo Tab 4 tablets. The tablets were equipped with protective cases to guard them from falls. All irrelevant apps were removed from the devices, and parental control software was installed.

We collected log data from SpeechBlocks and qualitative observations of children's behavior. Similar to SpeechBlocks I, the new app was instrumented to record all activity that happened within it. Due to privacy considerations, we were unable to use video recording in the classrooms and instead relied on two observers per session to take notes. Six Northeastern University Speech-Language-Pathology (SLP) graduate students were recruited to be the observers as part of their professional training. We decided to keep the structure of the notes open-ended to account for unanticipated interesting events. However, keeping in mind that the observers'

attention would be limited, we assigned each observer a target child, for whom observation would be prioritized. We also trained the observers to prioritize observations pertaining to:

- Verbalization of intended or completed content (e.g. children saying words that they would like to make or have already made), or literacy concepts (e.g. children pronouncing letter sounds);

- Literacy-related help requested or received by a child (e.g. spelling help from adult);

- Social interactions of the target child;

- Confusion about literacy concepts or usability;

- Children's affect (happy, bored, frustrated) with evidence (laughs, smile, yawns, random swipes).

We selected these focal aspects for observations so that observational and log data would compliment each other, allowing us to recreate the most complete picture possible of what happened in the classroom.

In addition to these data sources, we conducted two assessments before and after the study. One measured children's phonological awareness using the corresponding subset of CTOPP-2. The other quickly estimated children's executive function (EF) using the "hearts and flowers" test (Wright & Diamond, 2014), which was administered on a tablet. We conducted the executive function assessment because we hypothesized that this variable would be relevant to children's engagement in a sophisticated, child-driven, and mentally-active technology such as SpeechBlocks.. The tests were administered by the same six SLP students, while I was assisting them.

A note needs to be made on what I refer to as "CTOPP scores" below. The PA component in CTOPP for 4-6 year-olds consists of three tasks (elision, blending and sound matching). The aggregate score for the component is computed using age-adjusted scores. However, in some cases, I would like to look at the "raw" level of PA development, irrespective of age. To do that, I simply summed the three scores and referred to it as "raw CTOPP score". An alternative to this is to use CTOPP age equivalents of children as a measure of their PA skill.

There was a methodological flaw to the administration of the tests that was noticed too late. Administration of tests stretched over several days. To minimize disruption to the flow of classes, we pulled children sequentially from one classroom before moving to the next. Therefore, children from the treatment and control conditions were tested on different days. The school had constraints on available space, so we had to administer tests in one of its halls and the corridors. Although these environments were relatively quiet, there was always some level of noise and movement, which could have affected the results of different children. These background

distractions varied on different days. In addition, the dynamics within the classrooms changed day to day: e.g. children might have been more excited as weekends approached. Thus, there is a possibility that the effects of the testing day and effects of the conditions were conflated, similar to the possible varied effects of the different classrooms. A more appropriate procedure would be to test the children in random order. The resolution of this issue remains the subject of further, more rigorous studies.

6.1.2. Alterations to the Environment and Related Learnings

Based on our observations, we made a few changes to the environment throughout the course of the study, with the intent to create conditions more conducive to learning. We introduced headphones and changed the duration of the sessions in one of the classrooms from 10 to 15 minutes. There were also a few modifications to the app design, which are detailed in Chapter 3. The observations that motivated us to implement these changes, as well as the results of the changes, are among the learnings of the study.

Originally, we introduced tablets without headphones, believing that the headphones would isolate children and severely limit social interactions between them. Furthermore, from the experience of the first SpeechBlocks study, we judged that the ability to hear peer's devices would pique children's curiosity about their peer's activities. However, we found that this setup made it very difficult for children to work with scaffolding. Because the children were simultaneously using the scaffolding system, the overlapping chatter of the tablets made it challenging for them to determine which prompts were being directed at them. As a result, children exhibited a lot of chaotic, distracted behaviors. Introduction of the headphones led children to become visibly more focused. Moreover, contrary to our expectations, headphones did not disrupt social interactions in a noticeable way. The proliferation of social interactions around the app is described in section 6.4.

The two treatment classrooms originally had different rotation schedules. In one treatment classroom, the teacher divided the children into four groups of five children, while in the other, the teacher chose to have five groups of four children. The smaller groups in the second classroom rotated quicker, so initially the sessions with them lasted only 10 minutes. We noticed that the children in the second classroom were much less focused, and suspected that faster rotations contributed to this. Indeed, a part of the 10 minutes was consumed by demos, so by the time the children settled into play, it was often already time for them to move onto the next station. Their verbal expressions communicated their frustration. At one point, the transition between sessions caught a child in the middle of building the word TANK, and he lamented, "Awww, but I want to finish it!" The same thing happened a few weeks later, and he exclaimed: "Oh, come on!" Fortunately, the teacher kindly agreed to adjust the rotation schedule, and the duration of the SpeechBlocks sessions were extended to 15 minutes. Immediately, children became visibly quieter and more focused. The child who had expressed his frustration in the examples above now expressed his satisfaction at being able to complete his plans: "Yes! I finished all the words I wanted!"

These observations suggest the importance of giving children sufficient time to play. Anecdotally, we saw that children were able to engage with SpeechBlocks for prolonged periods of time. One day the teacher asked us to combine the two sessions that were planned for that day into one 30-minute session, for scheduling reasons. Three out of the four children present continued to enthusiastically play throughout the 30 minutes, and one of them even expressed a desire to continue, and to bring the tablet home to play there.

6.2. Play Types

The play with SpeechBlocks was not uniform. Variations existed in what children did with the app and how they used various features. Children also varied in the emotional tone of their play, in the type and amount of assistance they needed from adults, in what they found fun, and in their apparent motivations. I distinguished three broad types of play:

- **Word crafting**—focused on building words for the sake of doing that;

- **Imaginative play**—focused on creating scenes and stories;

- **Impulsive exploration**—characterized by engaging in seemingly chaotic, short-term-reward driven actions.

There are several reasons why fewer (and somewhat different) play types were observed in SpeechBlocks II compared to SpeechBlocks I. First, the interface of SpeechBlocks II is not optimized for remixing, which was a distinct form of play in SpeechBlocks I. Second, one of the SpeechBlocks I play types, *Communicative Play*, had been exhibited only in home conditions. It is likely that either children's skill level, environment, or both, were not conducive to this type of play in the present study. There is a rough parallel between Narration in SpeechBlocks I and Imaginative Play in SpeechBlocks II, although the latter is primarily imagery-focused. The other two play types are roughly the same.

Each play type consistently attracted a particular group of children. These children shared certain traits, making each type associated with a profile of a player who preferred it. However, the association of kids with play types was not clear-cut. Many children mixed several types of their play, in various proportions. There was also a child who exhibited an idiosyncratic form of play: he focused on tapping on the sound creatures and observing their actions. Because this was a singular case, I did not include it in this analysis.

The following sections will focus on each of the play types.

6.2.1. Word Crafting

Word crafting is a type of play characterized by intrinsic interest in building words but devoid of attempts to use these words in any way except collecting them. For instance, children who focused on this activity did not attempt to build sentences, stories or scenes with words and related sprites. Some of them collected the words they made by arranging them (and corresponding sprites) on the canvas pages. Two interesting categories of words were associated with word crafting: various names (children's own, their peers, relatives, friends, etc.) and complicated, unusual words. Despite word crafting being a limited activity, it was quite engaging for some children, who gravitated towards it as their primary activity. They typically scored high on executive function, but did not otherwise exhibit obvious commonalities. In this section, we will first look at word crafting activities common to everyone, and then look at the children who "specialized" in word crafting.

One universally appealing word-crafting activity was making names. Upon discovering their own names in the system for the first time, most children were surprised and tremendously excited. These emotions were manifested in their exclamations: "What? That was my name!"; "Did it just say my name?"; "Zack! Is that my name? That's my name!" When they completed their names for the first time, they were particularly proud to share it with the researchers and their peers. Many children expressed a desire to make their name several times: "I want to do my name again!"; "I'm going to make Alex again!" When names of their peers became available in the app, children were eager to point it out to their friends, saying things like, "Listen, Joe! It said your name!" while simultaneously turning the tablet towards their peers and tapping on it to trigger the sound. After spelling their own names a sufficient number of times, children switched to the names of their friends. A few children also attempted to spell their family names, names of their relatives (e.g. an uncle), and the names of the researchers. For that, they had to use the open-ended mode.

In SpeechBlocks II, names didn't have any associated pictures, since we were wary of potential privacy issues associated with collecting children's photographs. Although this did not diminish children's interest in making names, they were expecting to see pictures, and asked questions like: "Why no picture for Abigail?" If portraits had been included, name spelling could have become associated with imaginative play rather than word crafting. The possibility of including both themselves and their friends into their scenes could have changed imaginative play in interesting ways and potentially boosted it. In addition, pictures would have made it easier for children to find names in the word bank, so their inclusion could have reinforced name spelling as well.

In SpeechBlocks I, children tended to keep words they made in the word drawer; in SpeechBlocks II, children used canvases for this purpose. In many cases, they simply stored the words they made in a pile, without any apparent system or order (Fig. 6.1). Sometimes they created collections of items arranged by similar themes (Fig. 6.2). One child explained this behavior as: “I’m making a fruit collection.” A relatively common pattern was also labeling: dragging both the word and the related image onto the canvas and arranging them next to each other (Fig. 6.3).

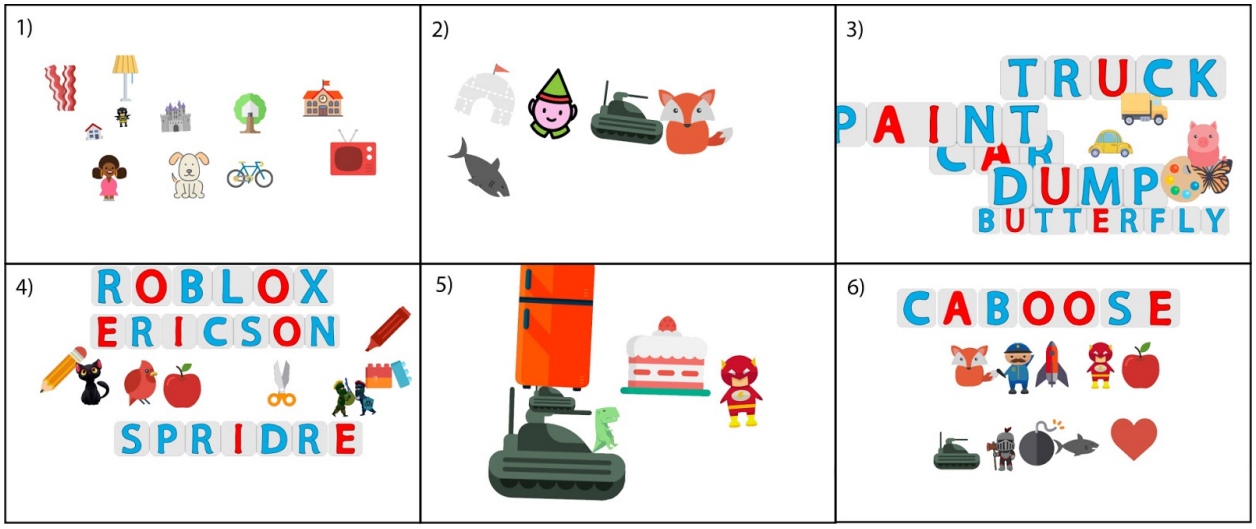


Fig. 6.1. Using the Canvas as a Storage Space

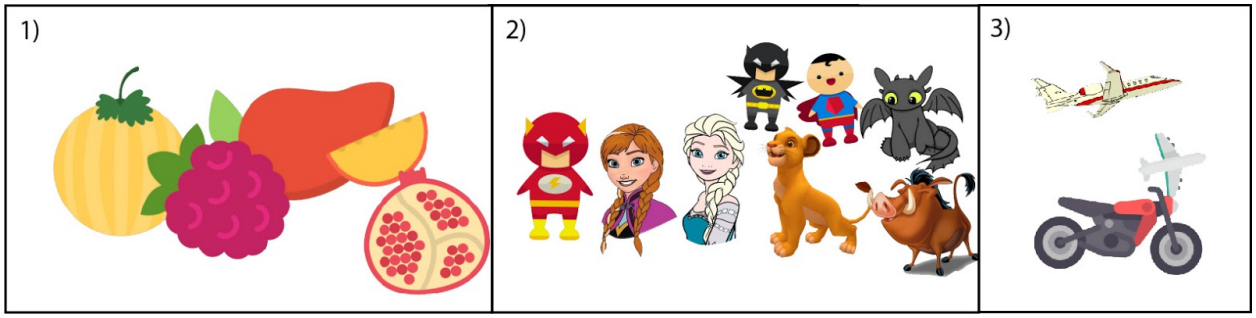


Fig. 6.2. Thematically Arranged Collections of Sprites

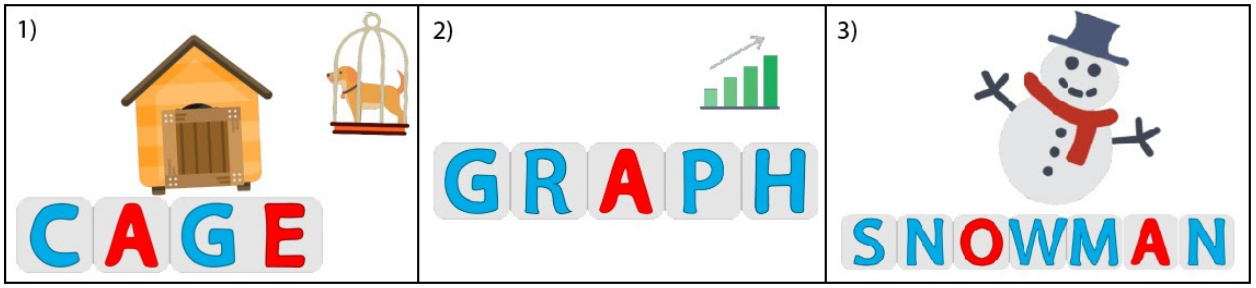


Fig. 6.3. Labeling

For several children, word crafting was their predominant mode of play. They mostly scored high on executive function pre-test (Fig. 6.4, c) but differed in their literacy skills. Three children in particular—Zack, Ericson and Ulisses—observationally stood out as quite "literacy-savvy" and keen to demonstrate their knowledge to adults. For instance, while building his name, Ericson told a researcher: "After that goes O, because my O actually makes *[a]* sound!"; and Zack introduced himself in this manner: "Zack. Starts with Z." Ulisses was able to read polysyllabic words fluently and had a very high CTOPP score for his age. Two other children—Mary and Mack—did not have higher than average literacy skills. While Zack, Ericson and Ulisses simply seemed interested in the process of making words, Mack and Mary might have been word crafters because their skill levels prevented them from engaging in more sophisticated play.

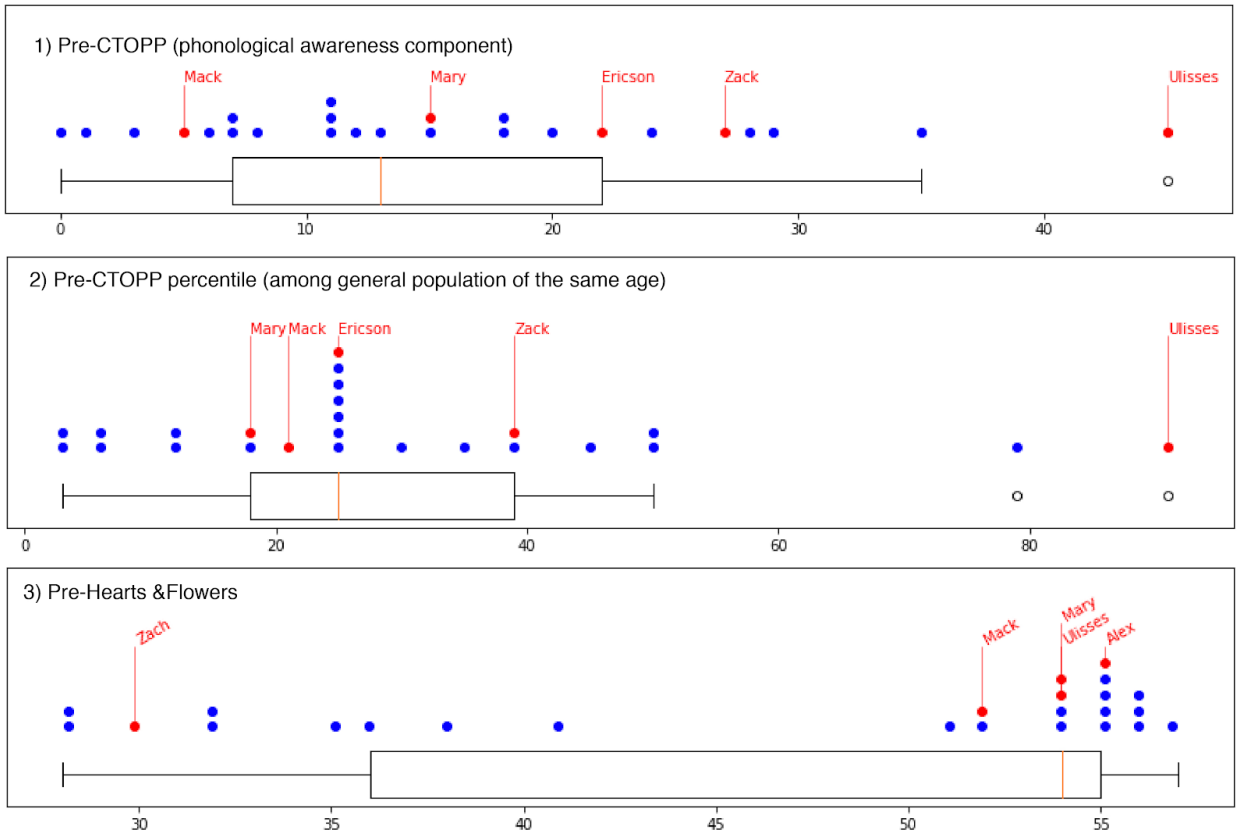


Fig. 6.4. Word Crafters Relative to Their Peers on CTOPP and EF pre-scores

Looking at one of Zack's play sessions, we see how word crafters engaged in building words without any apparent system. At first, Zack looked through the Word Bank until he found DINOSAUR. He built it while saying: "Dino, I need one more", and built the second one. He then continued to browse through the word bank until he stumbled upon BATMAN. He exclaimed: "Batman! My favorite! *[b]-[b]-[b]-[b]*...", as he started to assemble the word. He completed BATMAN, then built FLASH. He announced: "Now I'm moving Mr. Flash", and played with the FLASH sprite for a little while. He then proceeded: "Now I'm doing SIMBA." Upon adding SIMBA to the album, Zack exclaimed: "Yes! I finished all the words I wanted!"

Despite creating seemingly random words, word crafters were quite engaged in their play. Zack, for instance, was quite frustrated whenever he had to end his play prematurely because his session ended. In the beginning that happened quite often because his group used to have 10-minute sessions. Erickson and Ulisses skillfully used "high-tech" interfaces (which generally were somewhat challenging to use) to make words they were interested in. For instance, there was a day where they were engrossed in exploring the classroom using text recognition, talking with each other about the words they picked before building them.

Word crafters occasionally explored unusual words. For instance, Zack requested GOD, JESUS and FORTNIGHT from the speech recognition. Ulisses and another child built ROBLOX, the name of an online game platform, and were somewhat disappointed that no picture appeared. It also appears at times they were simply attracted to the complexity of some words. A friend of Ulisses once proudly showed him that she spelled TRANSPORTATION (a word that she picked up from a classroom shelf via text recognition). Ulisses was originally disinterested: "Transportation! I wouldn't spell that!" But soon after, he requested TRANSPORTATION using speech recognition, built it, and said "Yay! I built TRANSPORTATION."

While the three boys engaged in word crafting because of their interest in building words, Mack and Mary might have done that because of their limited capacity to engage in more sophisticated forms of play. All the same, these two children were focused enough to avoid chaotic exploration. Mary was not very strong at using scaffolding. Her average time to fill a single slot was more than one minute, and she struggled to pick the correct blocks. It might be that she had difficulties building words "by ear", so instead she focused on recreating words using her visual memory. For instance, she attempted to make her name in free mode 10 times over the course of the study, and eventually learned how to do it. Mack, on the other hand, might have had some difficulties arranging sprites on the canvas — possibly because his fine motor skills had not sufficiently developed yet. On one occasion, he asked his peer: "Can you help me? How did you do it?", referring to a scene she made. She tried to guide him, but eventually said: "Just do what you want. Just don't make it messy."

6.2.2. Imaginative Play

Imaginative play is the most sophisticated among the observed play types. It goes beyond merely making words and uses the resulting sprites to tell a (simple) story. This was done in two ways: either by composing a static picture out of sprites, or by enacting a story via moving sprites akin to physical toys. These two ways were often combined. Within-app play was also complemented by verbal narration. A diverse range of themes were explored by the players, such as fantasy, city, jungle, home life, and family. Many of the kids' creations were quite complex, involving as much as ten to fifteen sprites. Making such works required deliberate, focused effort. Children's intentionality and rich imagination were revealed by the comments that they made in the process of construction. Planned efforts coexisted with serendipity and externally inspired ideas.

The sources of such ideas were both technology and other children. Social interactions between players enriched and invigorated imaginative play. Players actively shared their creations with each other, borrowed ideas from peers, and helped each other with both software and literacy issues. Although imaginative players used scaffolding heavily, this form of play was still demanding, requiring construction of large numbers of words without veering off-track. As a result, children who gravitated towards it also displayed high executive function and at least moderate CTOPP scores. By the end of the study, they exhibited a high amount of autonomy.

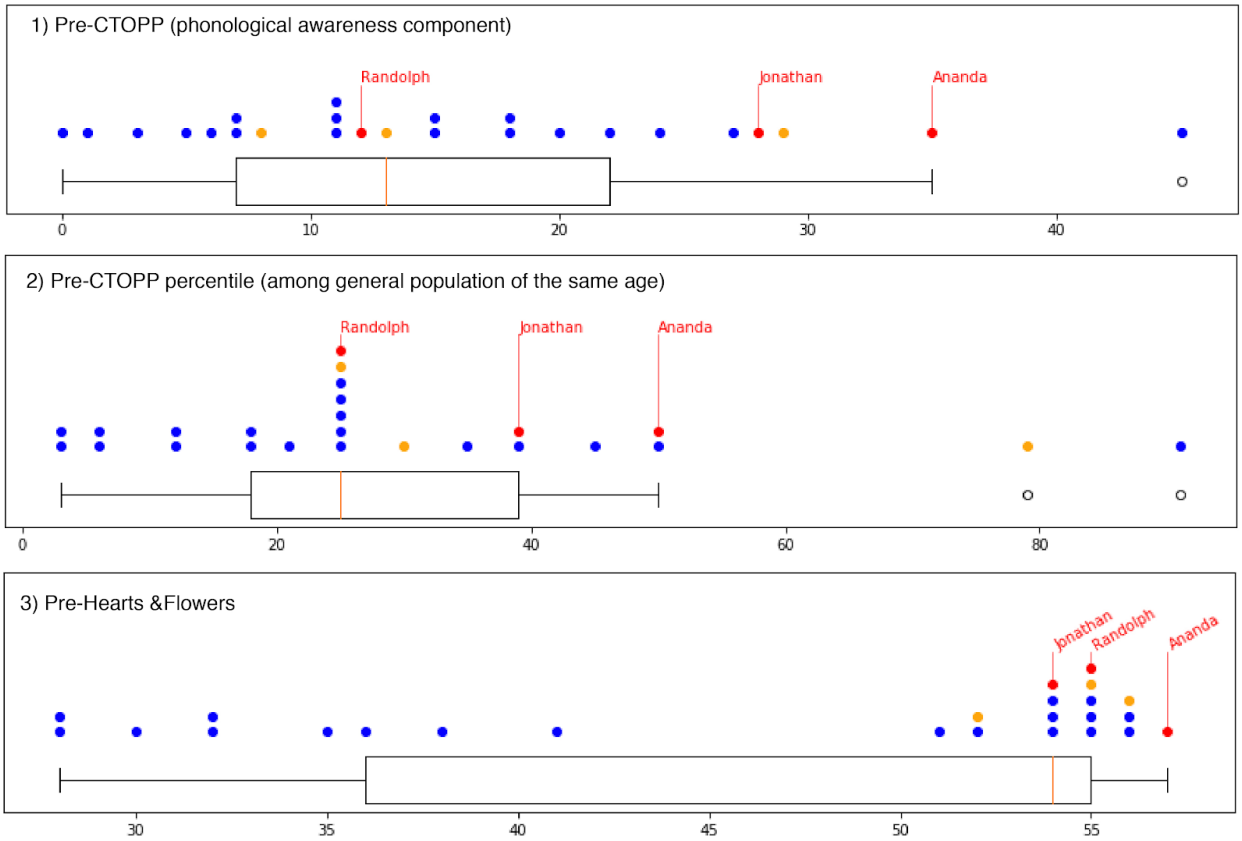


Fig. 6.5. Imaginative Players Relative to Their Peers on CTOPP and EF pre-scores. In red are the three children selected as examples. In orange are three more children whom I considered particularly active imaginative players, based on both the log data and observation records.

I made an attempt to estimate how widespread imaginative play was and to select avid imaginative players. Identifying imaginative play was at times challenging: in some cases, it was difficult to determine whether sprites were arranged in a certain way to form a scene or just by chance. I used my judgement to find some scenes that appeared clearly intentional to me (such as the scenes on Fig. 6.9 and 6.10). I then inferred additional cases of imaginative play using evidence of children's intent appearing in the records of their conversations and behavior. I found that almost every child in the study exhibited some imaginative play behaviors on at least one occasion. However, there were only six children (about a quarter of the total number) who clearly engaged in imaginative play more than three times throughout the study. Only four children (a sixth of total

number) exhibited imaginative play consistently, during most of the sessions. Furthermore, they all belonged to one group and were seated at the same table. This arrangement was the result of the teacher's perception of them as "advanced" children and her intent to put them together so that she could do some special activities with them (such as invented spelling). I selected three children from that group, denoted by fictional names Ananda, Jonathan, and Randolph, as the primary sources of examples. Personality-wise, Ananda appeared open and sociable, but disciplined and relatively quiet, while Jonathan and Randolph stood out as rowdier, but still capable of focused work.

Fig. 6.5 shows positions of the six avid imaginative players relative to their peers on CTOPP and EF pre-scores. All six children scored high on the EF assessment and none had particularly low CTOPP scores. These values are likely not coincidental. To build scenes, children needed to coordinate long, pre-planned sequences of actions and suppress all impulses that could disrupt the process. The less literacy knowledge they had, the longer and more strenuous the process became, and the more strain was placed on their executive functioning. Therefore, low CTOPP and EF scores likely presented a barrier of entry to the imaginative players "club." Still, it seems that demands of imaginative play on phonological awareness are moderate: some avid imaginative players had CTOPP scores well below the population median for their age.

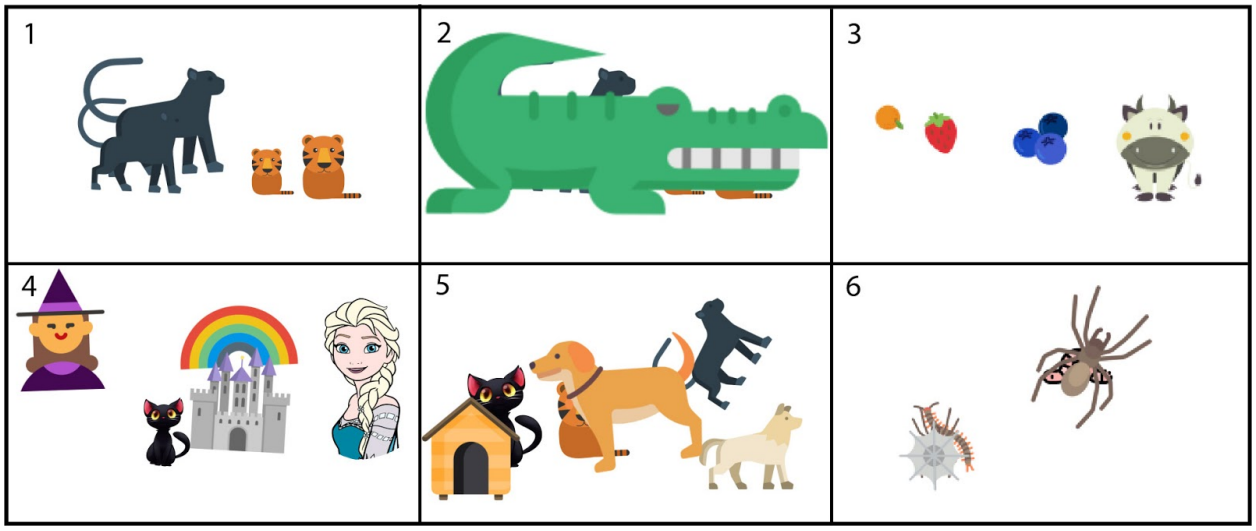


Fig. 6.6. Examples of Enactment. (1) Ananda's jungle scene before and (2) after addition of the crocodile, (3) Feeding BLUEBERRY to a COW, (4) "Witch, I need you to come", (5) a DOG and a PANTHER fighting, (6) Critters. Scenes 3 and 4 slightly edited to improve readability.

Imaginative play is manifested in two ways: (1) building static images and (2) enacting the scene using sprites akin to physical props for dramatic play. Figure 6.6 shows several examples of the second form of play. On Figure 6.6, (1), one can see several animals built by Ananda. She put a crocodile over them and started to rock it back and forth while saying "Chomp! Chomp! Chomp! Chomp!" (Fig. 6.6, (2)) to enact it devouring the other animals. Figure 6.6, (3) reflects a rare but curious example of a child expecting the app to go along with the enaction. The child "fed" the

BLUEBERRY sprite to the COW and was somewhat disappointed the BLUEBERRY didn't disappear. Enactment was often accompanied by commentary and dialogue. Figure 6.6, (4) and (5) reflect such cases. On Figure 6.6, (4), the child role-plays Elsa: "Witch, I need you to come!" Witch "responds": "I need the cat today." Figure 6.6, (5) corresponds to a long and rich in commentary enactive play, depicting a fight between a DOG and a PANTHER. After building the PANTHER, the child said: "Roar! He is going to eat my buddy", told the researcher: "Look! The dog and the panther are fighting!" and started to move both sprites around the screen, imitating a brawl. The researcher asked: "Why are they fighting?" The child responded: "Because the dog was watching the panther, and then they started to fight. And then the cat came out. Look! Cat is biting her [panther's] tail!" After imitating the sounds of battle for a while, he role-played the dog: "Panther, I'm going to kill you!"; then the cat: "Oh dog! I'm going to help you!" Sometimes the enactment was conducted through physical movements of the entire tablet or body movements. For instance, after constructing the arrangement shown on Figure 6.6, (6), Jonathan menacingly "walked" the tablet by rocking it from side to side, to show the crawling of dangerous critters. The enactive behaviour matches similar observations by Makini (2018). The boundary between making static images and enactment is blurry. Children were seen arranging a few sprites in a static composition, then enacting some action (with the enactment typically being directed at their peers as an audience), then continuing with the composition.

The theme of children treating sprites more as physical toys on a rug than images on a sheet of paper manifests itself in other ways as well. Although SpeechBlocks allowed players to easily start a new page for each new composition, many children created new compositions in the middle of old ones—as seen in Fig. 6.7, where a jungle scene is started amid a composition of vehicles. At times sprites from the old composition were repurposed to serve a different role in the new one. Similarly, when children play with toys, they exclude from their attention the toys that they are not currently interacting with, even if they are lying in front of them. Another toys-related behaviour observed was children moving all the sprites from one page to another and rearranging the entire composition on a new page, even though such reorganization served no visible purpose. On the other hand, children created no imitations of physical text forms, such as comic strips, newspapers, cards and letters (even though the possibility of doing that was incorporated into the design of the medium).

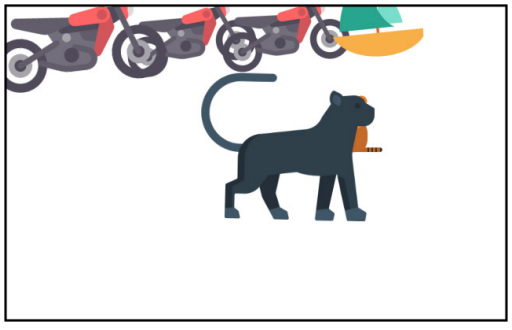


Fig. 6.7. Starting a New Scene Amid an Old One

Children's in-app expression was often complimented by verbal narration or commentary directed at themselves, their peers, and researchers. This commentary often revealed the complexity of children's imaginative intent, highlighting characters, their roles and their actions, as well as the logic of the scenes. Occasionally, a hidden structure was revealed behind what otherwise looked like a collection of random sprites. For example, while building the lynx scene shown on Fig. 6.8, (a), the child said: "This is a father, a mother and a baby." While building the royal family scene on Fig. 6.8, (b), the child put the BABY sprite on top of the BED and said: "The baby goes to sleep. Goo goo." During construction of the Elsa and Witch scene on Fig. 6.8, (d), the child said: "There is a friendly witch living in the castle." On Ananda's scene in Fig. 6.8, (c), one can see superheroes guarding a castle, while hidden behind the castle are figures of Anna and Elsa. Ananda said to researchers: "they each got their own room!" While building the scene in Fig. 6.8, (e), she commented about the big and small ninjas: "They are father and son. They are practicing." She reflected a little bit on the weaponry each of the ninjas should receive: "He's got a sword, and he is going to have a shield." Before adding the third character (a sprite corresponding to the word *prisoner*), she said: "Now I'm going to make a villain to fight them!"

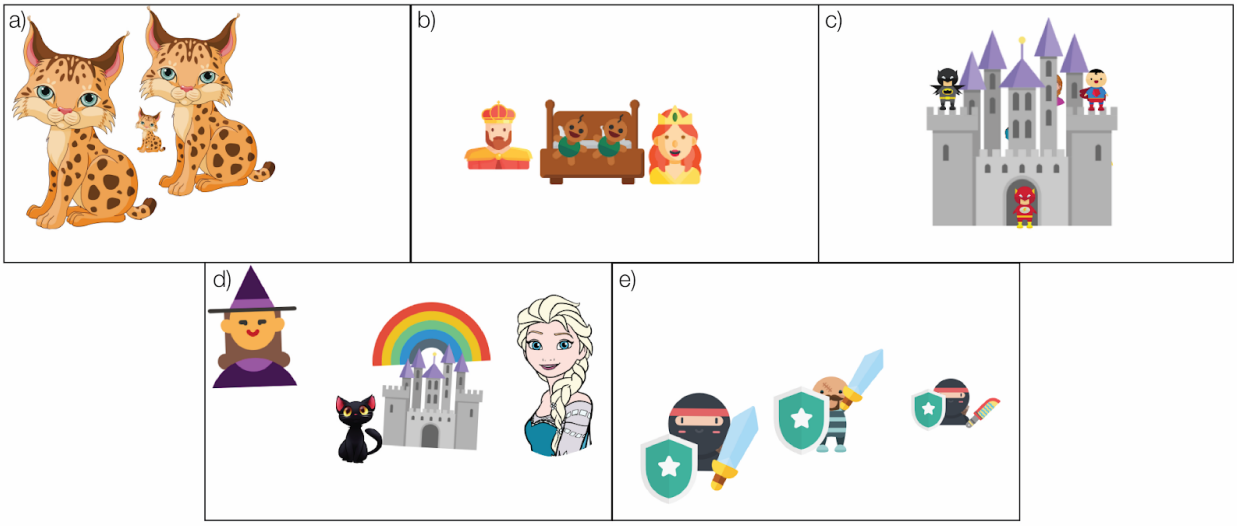


Fig. 6.8. Some scenes accompanied by verbal commentary

Children's works were quite diverse in their themes. Fig. 6.9 shows all the scenes created by Ananda, the most prolific of all the scene-makers. In her works alone, one can see themes related to fantasy, science fiction, wildlife, domestic life, and outdoors. Each of these themes has been explored by multiple children. Fig. 6.10. shows varying approaches to the same theme, a theme of family, by different players. One can see ordinary families, a royal family, and an animal family. Along with diversity, one can see some recurring elements, such as the presence of small babies.

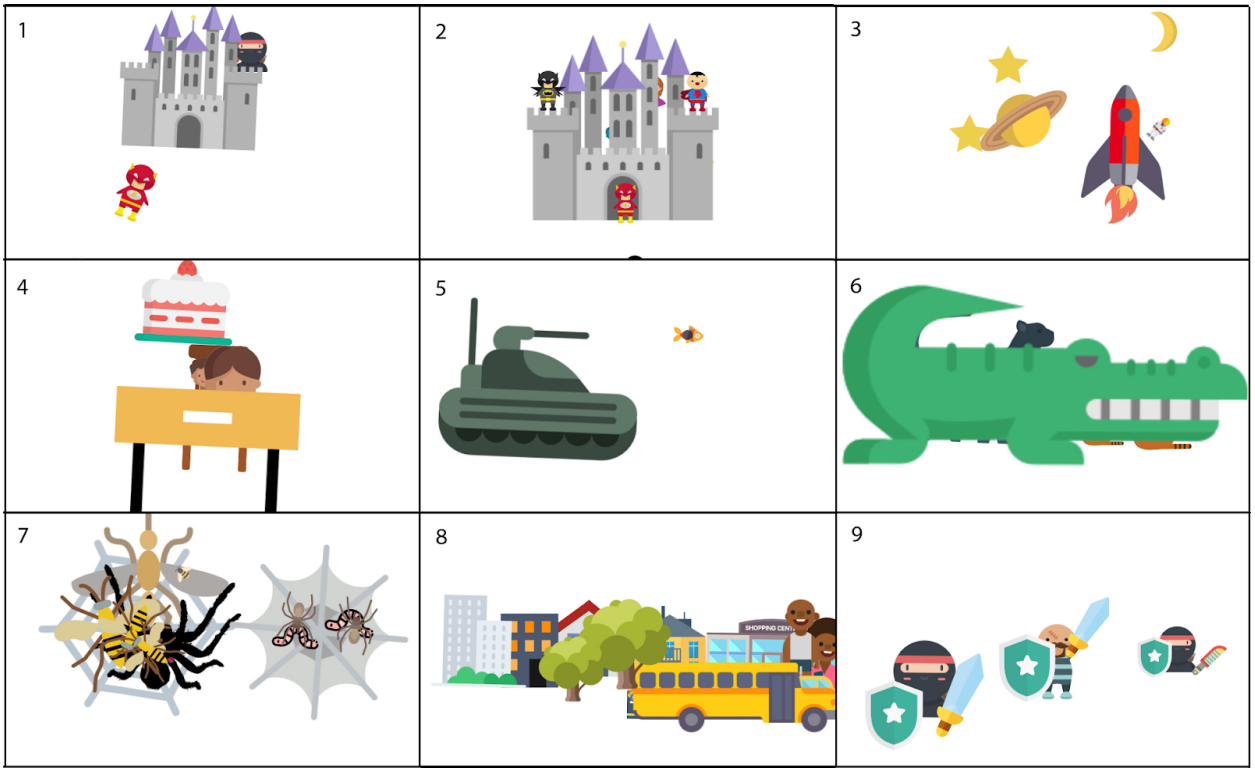


Fig. 6.9. Scenes created by Ananda. (1) a ninja living in a castle (2) royal family (hidden behind the castle) living in a castle and guarded by superheroes (3) a space scene (4) a boy and a girl having a cake on a table (5) a little fish being blown up (6) a crocodile devouring other animals (hidden in crocodile's belly) (7) insects caught in a spiderweb (8) a town (9) ninjas fighting a villain

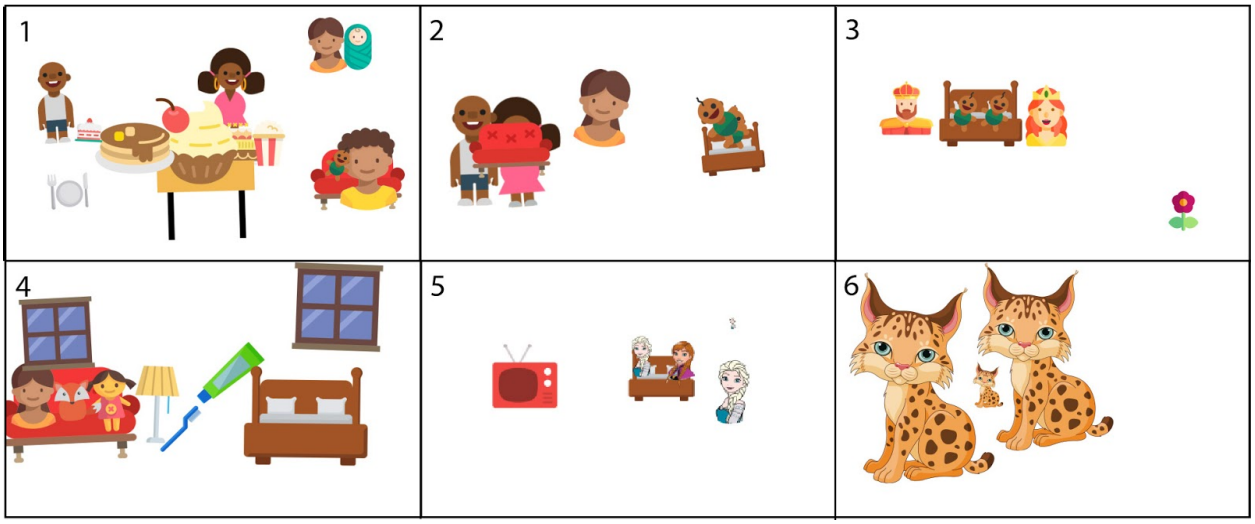


Fig. 6.10. Family scenes by different children. (1) a family feast (2) a family at home (3) a royal family (4) a girl, a fox and a doll on a couch, getting ready to brush their teeth (5) Anna and Elsa (cartoon characters) watching TV (6) lynx family: a father, a mother and a baby lynx (according to the creator of the scene)

Many of the scenes shown on Fig. 6.9 and Fig. 6.10 are rich in detail, sometimes having 10 to 15 individual sprites. Constructing such a scene required commitment, given that even good spellers were usually able to produce only about 1 to 2 words per minute. Sometimes children continued playing with elements of the same page for several sessions. On the other hand, meaningful expression was possible even with very few sprites. A selection of such simple scenes is shown on Fig. 6.11. Note how children used sprites to compliment other sprites — for instance, to add an exhaust plume to a rocket, or to give Elsa (whose sprite only showed head and shoulders) a body. Such simple arrangements may be an entry point into expressive play.

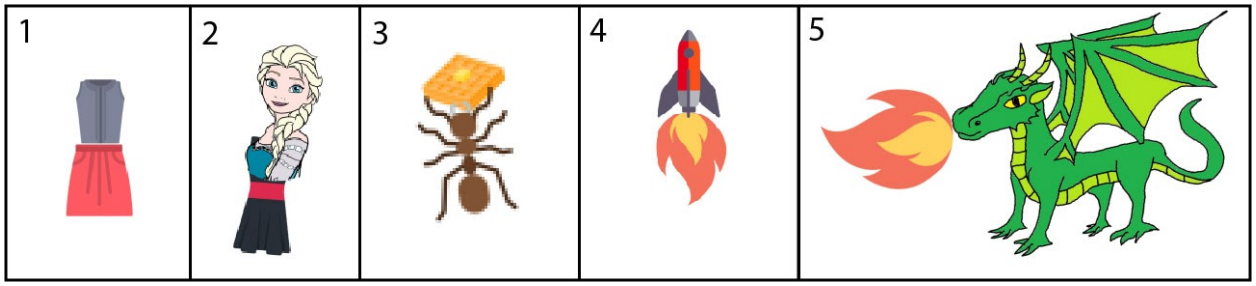


Fig. 6.11. Simple aprite arrangements. (1) a suit assembled of BLOUSE and SKIRT, (2) ELSA wearing DRESS, (3) an ANT carrying a WAFFLE, (4) a ROCKET spewing FIRE, (5) a DRAGON breathing FIRE

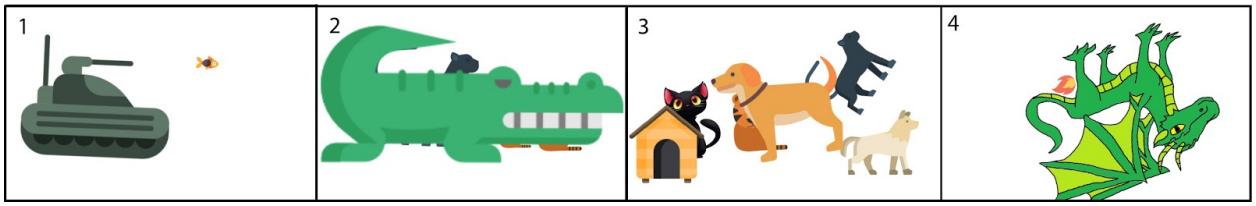


Fig. 6.12. Scenes with elements of grotesque: (1) a tank "blowing up" a fish, (2) a crocodile devouring jungle animals, (3) a brawl between a panther and a dog, (4) "dragons fart"

A sizable portion of the imaginative play was somewhat violent or grotesque. We have already seen two examples of such play, with a crocodile devouring jungle animals (Fig. 6.12, (2)) and the brawl between a panther and a dog (Fig. 6.12, (3)). Another example is Ananda saying that the fish is being "blown up" by the tank while constructing the scene on Fig. 6.12, (1). In other cases, children aimed to convey grossness, such as the portrayal of bodily functions. After making the dragon shown on Fig. 6.12, (4), the child said loudly in a deep, gravely voice: "I'm a destroy dragon! Big, big, big! Giant! Giant-er! I'm humongous ugliest beast." He made a flame in order to make the dragon breathe fire, but then flipped the dragon upside down, so that the fire was now located under the dragon's tail, and said: "Dragons fart." This type of play seems to be related to an interesting phenomenon described by Sutton-Smith in his studies of children's storytelling. He noted a strong trend of "phantasmagoria" — grotesque, violent, gory, gross, obscene, and absurd themes that challenge adults' stereotypes of what a story told by a child ought to look like (Sutton-Smith, 2009). There are multiple points of view on why it is prominent, and even whether this phenomenon is specific to childhood (Bickford, 2017). For instance, Mizuki Ito suggests that

these themes manifest children's reaction to adults' attempts at "sanitizing" and structuring their play (Bickford, 2017). In the case of SpeechBlocks, another possible reason for "phantasmagoria" might simply have been the children's interest in impressive, dramatic, and out-of-the-ordinary things. There also seemed to be a social element in such play: these themes were likely to evoke a response from peers, which motivated children to explore them. Different educators and parents may have varying opinions on whether "phantasmagoria" in SpeechBlocks play requires any special handling, or whether it should be discouraged. In any case, it is valuable to know that such play is likely to happen.

Imaginative play was accompanied by an abundance of social interactions: sharing the scenes and talking about them, borrowing ideas from peers, and helping peers with both word building and technology. These interactions are described in section 6.4.

Flourishing imaginative play involved creation of many sprites. Consequently, it stimulated word building. Imaginative players tended to spend a lot of time in scaffolded mode (Fig. 6.13), but because of their relatively high skills, they tended to spend little time per slot (Fig. 6.14). As a result, they tended to construct a lot of words during each session (Fig. 6.15). This high amount of word building could be beneficial for their learning (e.g. phonological awareness learning).

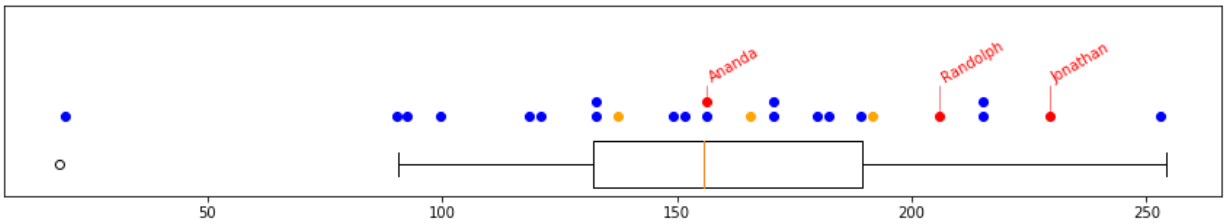


Fig. 6.13. Imaginative players relative to peers on time in scaffolded mode (seconds per session)

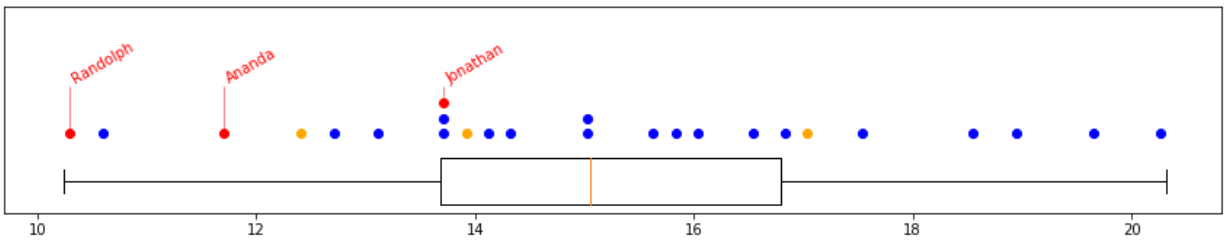


Fig. 6.14. Imaginative players relative to peers on time spent per scaffolded block, in seconds

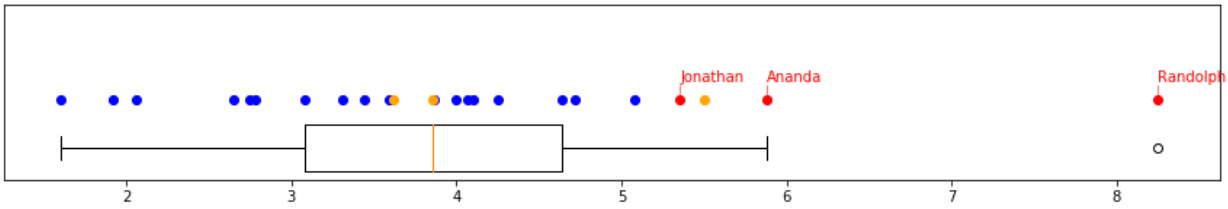


Fig. 6.15. Imaginative players relative to peers on number of words built per session

6.2.3. Impulsive Exploration

This style of interaction is characterized by a lack of systematicity and focus on short-term rewards — emotional, social, and cognitive — that can be reaped through the interaction with the system. Long-term plans may be expressed by the players, but are almost never followed through. This form of play is often dynamic and passionate, but chaotic. Players frequently experience difficulties with building words, but compensate for it by coming up with other ways to have fun. This often results in unexpected and unintended use of the technology's features. The primary drive of impulsive exploration seems to be enjoying the agency and the entertainment value provided by the digital technology at the level that the child's current skill allows. Some of the avid impulsive explorers were extremely eager to play with SpeechBlocks and impatient to start autonomous play. Despite having a chaotic style of play overall, they exhibited periods of focused, goal-oriented play — particularly when being assisted by scaffolding. Their play seemed to become increasingly skilled and focused over time. Others were frustrated by demands that the system placed upon them. Limitations of the technology strongly affected children who gravitated towards impulsive exploration. It was typical for impulsive explorers to have either low pre-test executive function scores or low CTOPP scores, or both. Their interaction style likely emerges as a result of their limited ability to interact with SpeechBlocks in a more structured way.

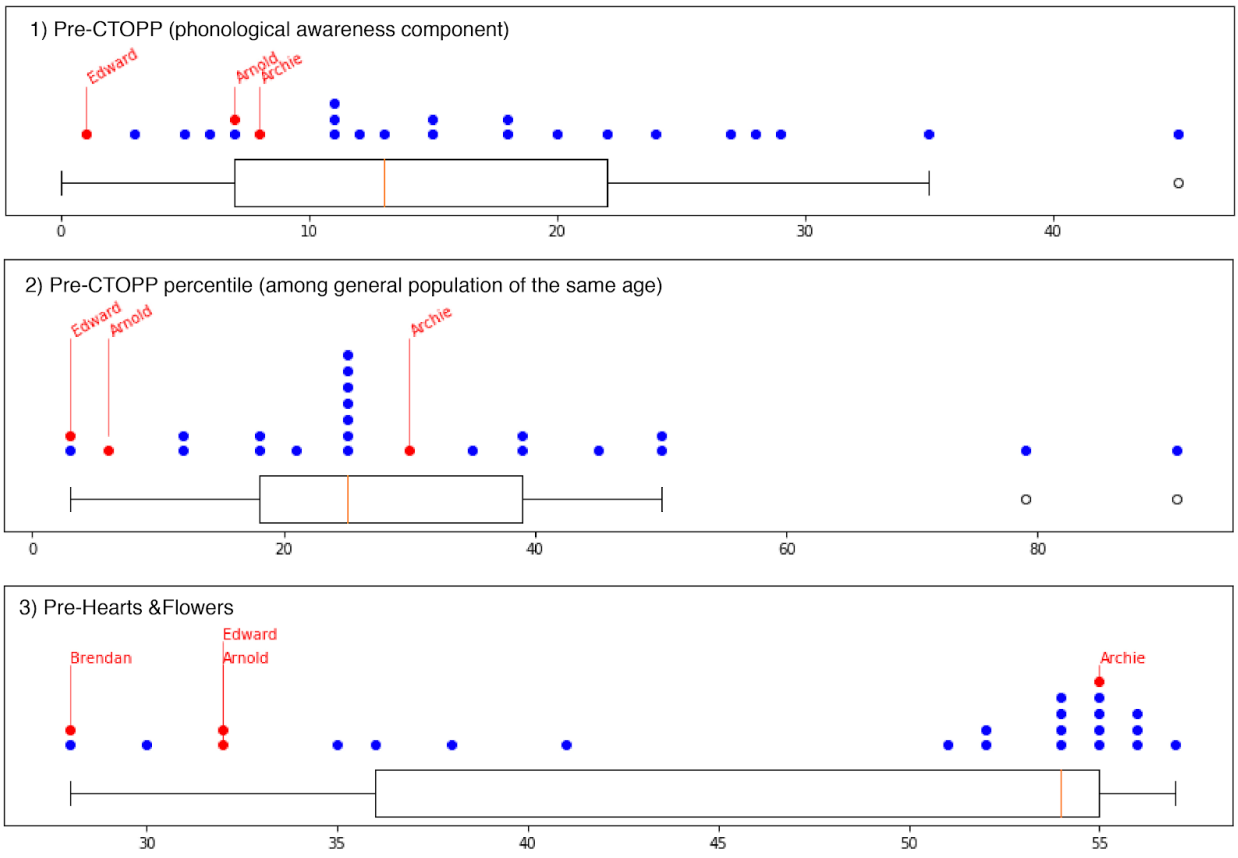


Fig. 6.16. Impulsive explorers relative to their peers on CTOPP and EF pre-scores.

To paint a picture of this interaction style, we will primarily use examples from four children, whom we will denote as Arnold, Edward, Brendan, and Archie. All four are boys — in general, boys exhibit much more impulsive exploration than girls. Fig. 6.16 depicts their relative standing on phonological awareness (PA) and executive function (EF) pre-scores. Brendan is excluded from the CTOPP diagrams, since he refused to collaborate with the researchers on the CTOPP pre-test. However, that in itself highlights the impatient behavior that was common of impulsive explorers. One can see that Edward, Archie, and likely Brendan, scored low on both measures. Archie was an exception, having a high EF score and a moderate PA score (particularly with respect to his age). His play differed qualitatively, combining impulsive exploration with elements of imaginative play. He is an example of how one child can engage in several types of play, and may have been in transition towards the more sophisticated type.

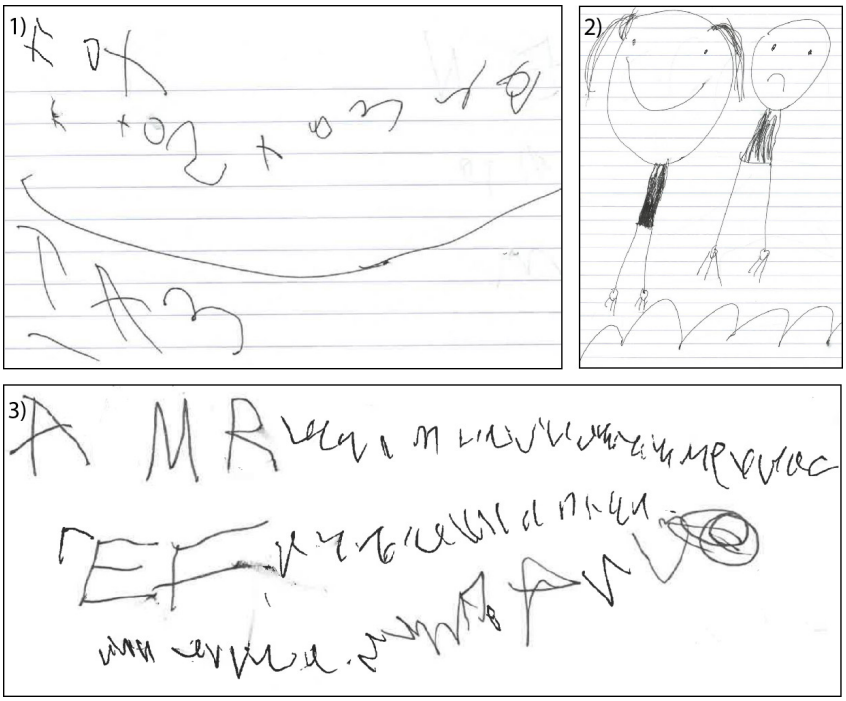


Fig. 6.17. Scribbles of Brendan (1, 2) and two other impulsive explorers (3).

Observationally, the four children stood out as restless and vocal. They exhibited a lot of self-stimulating behaviours during their play, such as making jerky motions, waving their hands, rocking in their chairs, murmuring to themselves what appeared to be fragments of songs or random phrases (e.g. Arnold mumbling “What is hit? What hit is this?” while building something in SpeechBlocks). In one particularly interesting case, Archie had an exchange with his peer via what seemed to be invented words, while building something unrelated in SpeechBlocks. Archie was mumbling “fashi fashi butchi bushi”, to which his peer said “bamutacola!”; then they both laughed. Despite the low CTOPP scores, Arnold, Edward, and Brendan exhibited some interest in literacy-related items. Arnold and Edward looked very curious in the books that were laid out on a table in preparation for a session involving text recognition. “How many books do you have?”,

asked Arnold, and Edward noted: “Books are for reading.” They were delighted to learn that they would be able to use the books. Brendan and two other impulsive explorers showed interest in scribbling and sketching using traditional media — their scribbles are shown in Fig. 6.17.

Many impulsive explorers seemed to appreciate considerably the agency of their individual interactions with the devices. One way this manifested itself was in their desire to get a hold of the devices as quickly as possible. In the beginning of most sessions, we held a small demonstration of new features and modeled how to interact with SpeechBlocks. We tried to make these sessions as interactive as possible, taking input from children on what to make and asking them what phonemes they heard in the words we were building. Arnold, Edward, and Archie tended to be impatient during that period. Archie sometimes kept standing during the demo, and Edward and Arnold leaned on the table towards the demo tablet. They reached towards the demo device at random moments and tried to tap and swipe on it unsystematically, often disturbing the demo. Sometimes they grabbed the device still in the researcher’s hands and tried to orient the screen towards themselves. On one occasion, Archie exclaimed: “Gimme! I want!” When engaging in the collective deliberation on what to build, they were often frustrated when their ideas were not followed through. But even if their idea was chosen, they didn’t always pay attention to the laborious process of building the word. Typically, they were quick to lose attention of whatever was being demoed and instead would start to look around, most often in the direction of the pile of devices that were being prepared at a different table. They also entertained themselves in other ways, such as playing with observation clipboards and materials that were placed on the table, as well as with researchers’ name tags. Arnold was once reproached for not paying enough attention, but appeared to be offended and ceased to watch the demo completely, as if in defiance of the structure imposed by the researcher. Noticing their urge to interact with the devices, we tried to recruit their help in presenting the demos. However, this only partially worked, since the children seemed to be interested in carrying out their own ideas of what to do with the devices, rather than following our instructions.

A counterpart to the desire of getting the devices as soon as possible was the reluctance of many impulsive explorers to put them down when their session ended; this behaviour was particularly noticeable with Arnold and Edward. At the end of one session, Edward was observed crouching on his chair, holding the tablet close to the floor and under the table, as if he wished to be overlooked by the teacher and therefore continue to play. At the end of another session, he said about the tablet: “I want to keep this.” Most notably, that was said after an unusually long, 30-minutes play session, during which he was engaged the entire time. Similarly, Arnold once said during his play: “I want to do this at home.” During sessions, Edward and Archie were occasionally noticed holding devices in a private or protective fashion: instead of keeping them on the table, they held them close to their body, under the table or (in Edward’s case) on top of their shoes. In some cases, impulsive explorers, who were deeply engaged in their private play, treated attempts of other children to communicate with them as an unwelcome distraction. For example, on one occasion Edward’s friend started to tell him excitedly about the scene he made: “Look! Look, Edward! They are jumping in the....” Edward responded impatiently: “I don’t want to!”, and did not move his gaze away from the screen. On another occasion, he turned away from his group to

minimize distractions to his play. Impulsive explorers also did not appreciate unsolicited interventions in their play, even with the intent to help. For instance, a researcher wanted to show Arnold how to use the album. Arnold exclaimed: “I can do it!” and pushed the researcher’s hand away. Combined, all of these observations create an image of impulsive explorers valuing their personal agency during their play.

Getting SpeechBlocks in their hands, impulsive explorers often expressed ambitious plans of what they wanted to do with it. However, the impatient nature of their interactions prevented most of these plans from coming to fruition. For example, Brendan started one of the sessions saying: “I’m going to have king and queen and Elsa and Anna...” A researcher noted: “You first have to spell them! What do you want to make first?” Brendan exclaimed: “All of them!” while tapping buttons on the screen all at once. The researcher remarked: “You have to do things slowly, or it [SpeechBlocks] gets confused”, and a peer said to Brendan, too: “You need to build them first!” After some deliberation, Brendan and the researcher decided on a word to build. However, instead of trying to build the word, Brendan passed the headphones to the researcher, asking that the word be made for him. The researcher rejected his request and said: “Listen! You have to be patient! Which one makes that sound?” (referring to the initial sound of the word that had just been played by the scaffolding system). Brendan picked a random letter and exclaimed: “I got it! I got it! R!” He was also restless in his chair, and the researcher said: “You need not to move so fast, and to listen.” Working together, they managed to build one word by the end of the session.

Another variation of this pattern can be seen in a play sequence by Edward. It started when a peer sitting next to him announced: “I’m making ten BATMANs!” Edward was puzzled: “Why are you making ten BATMANs?”, but the peer provided no rationale. After pondering on the ten-BATMANs idea for a moment, Edward apparently liked it too, and exclaimed: “I wanna make ten BATMANs!” He went to the Cartoon Characters section in the word bank, but something else attracted his attention, so he didn’t choose BATMAN at first. However, he eventually returned to that section and picked that word. The system went to the scaffolding mode and started telling Edward which sounds the word consisted of. He tried to drag a few blocks into the slots, but apparently picked them randomly, so the system rejected his choices. Edward gave up, went to the open-ended keyboard and started building a random nonsense word. Meanwhile, his friend completed one of his BATMANs and showed Edward what he made. Edward responded to the friend: “I wanna make BATMAN!” The friend showed him how to select BATMAN from the Cartoon Characters section. Edward objected: “BATMAN is not working!”, referring to his previous unsuccessful experience building it. His friend said: “You just gotta spell it! Can you hear it?” He pointed to the blocks on the scaffolded keyboard and started helping Edward make the word. Edward himself was, however, looking away and not listening to what his friend was saying. The friend then asked Edward if he heard what sound would come next. Edward pointed to one of the blocks without much confidence. The friend confirmed: “Yeah! That guy!” Together they finished BATMAN, and the corresponding sprite appeared. The friend said: “Yeah, now you can put it anywhere you want.”

These two examples share interesting behaviour patterns that seem to be indicative of thinking and underlying motivation of impulsive explorers. First of all, in both of the examples, there is a passionately expressed desire to get a certain outcome — a certain set of sprites. However, the commitment to this outcome may not be very strong. In the second example, we see that the child almost immediately became distracted by something else. After returning to his initial plan, he proceeded for a little while, but then abandoned it. However, after his friend provided help, he returned to that plan and was able to fulfill part of it. Such switching between ideas was common. For instance, once Archie was trying to use text recognition on some words written on the board, and said: "I want to do MONDAY!" However, he held the camera unsteadily, and MONDAY was not picked up by the system. Archie immediately abandoned his plan: "I don't want to do MONDAY!" On another occasion, Edward reinterpreted the sprites that he had already made. He built three SUPERMANs in a row and arranged them on the page. A researcher asked him: "Who is Superman playing with today?", and he corrected her: "Batman! No Supermans." It is plausible that such lack of consistency was impulsive explorers' way to compensate for frequent inability to accomplish their plans.

The second commonality between the two examples is not paying attention to directions. We saw not only that impulsive explorers often ignored demos, but also that they would ignore directions which were intended to help them in achieving their own goals. For some children, this lack of attention led to deep misconceptions about how the word building process worked — at least during the initial part of the study. For instance, in the above example, Edward attributed the inability to complete BATMAN to technical issues with the system ("BATMAN is not working!"). On another occasion, while building DINOSAUR, Edward said: "I'm going to put TH here!" and persistently tried to drag the TH block into the final slot. He did it four times, despite the system rejecting it each time. We have observed similar behaviors both with other children during Study IV as well as during play-testing at a children's museum. Apparently, these children assumed that they simply needed to fill the word box with blocks, and didn't pay any attention to acoustic feedback or features of the blocks.

The third commonality was the apparent lack of interest among both children in the word building process — they were only interested in the outcome. In fact, both children had to be reminded that they needed to build the word first and afterwards, they both attempted to outsource the spelling part to the researcher and their friend, correspondingly. Another child in the study (not one of the four) did, in fact, succeed with outsourcing attempts: on several occasions he managed to get a peer to build words for him. Such behaviour stands in stark contrast to the behaviour of the word crafters.

The reader might wonder whether this avoidance of spelling was caused by the rigid nature of the scaffolding mechanisms, which placed constraints on the agency-loving impulsive explorers. I think, however, that there is evidence to the contrary. First, scaffolding-free open-ended mode was available to impulsive explorers at all times, but they only used it to build random sequences of letters and sounds. Second, in both of our examples (as well as many other observed cases) children approached building words through picking random blocks and not paying attention to the

acoustic feedback of the system. This shows that without scaffolding, they would likely do no better, becoming lost in assembling chaotic combinations of blocks.

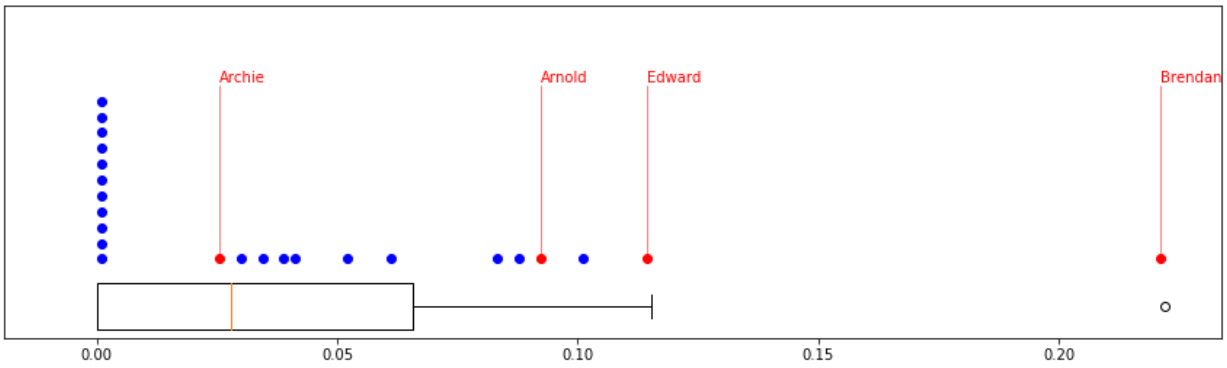


Fig. 6.18. Impulsive explorers rate of repeated incorrect attempts (per slot), during the first 3 weeks of the study

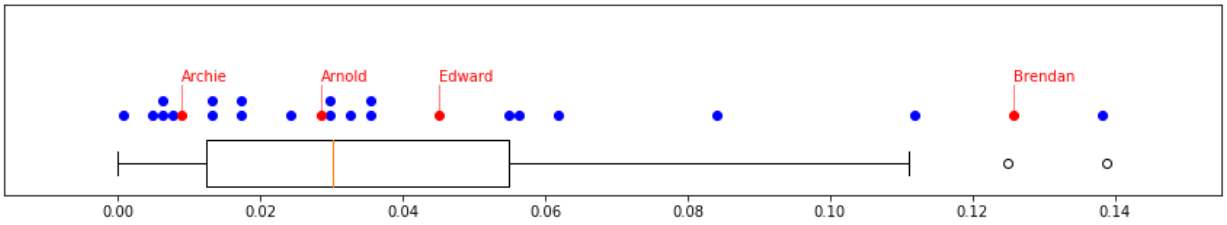


Fig. 6.19. Impulsive explorers rate of repeated incorrect attempts (per slot), overall

To estimate the frequency of such trial-and-error word building, I measured the number of slots (in the scaffolded WordBox) for which the child needed more than two attempts to put the correct block in place. One can see that in the early phase of the study, our four impulsive explorers were relatively high on this measure (Fig. 6.18). However, towards the end of the study, many impulsive explorers started to exhibit a good amount of purposefulness in building words. In line with that, the four children did not stand out anymore on the frequency of such repeated mistakes (Fig. 6.19). Their increased purposefulness in word construction was manifested by two strategies. One was tapping on different blocks on the keyboard in order to hear their sounds before choosing one to drag into the word box (as opposed to dragging blocks at random and seeing whether the system rejected them or not). The other was tapping on the "help" button to receive an extra bit of guidance (to hear the target phoneme and the corresponding onomatopoeic mnemonic again). One can see that impulsive explorers stood out among their peers on usage of both strategies (Fig. 6.20 and 6.21). Fig. 6.22 shows how word building strategy of the four boys evolved over the course of the study. For Arnold, Edward, and Archie, we see the fraction of blocks found via taps picked up during initial weeks and then balanced back and forth with "instant hits" (immediately putting the correct block in the slot), while the frequency of mistakes, particularly repeated ones, and unfinished slots, steadily decreased. Less can be said about the dynamics of the question button usage, which appeared to be quite erratic (Fig. 6.23). Although impulsive explorers remained low relative to their peers on the number of instant hits throughout the study, the reduced

randomness in their interactions with SpeechBlocks suggested ongoing learning (phonological and/or related to the app functioning). This gradual transition towards more focused play may have been related to the phenomenon of normalization described by Montessori (P. P. Lillard, 1972).

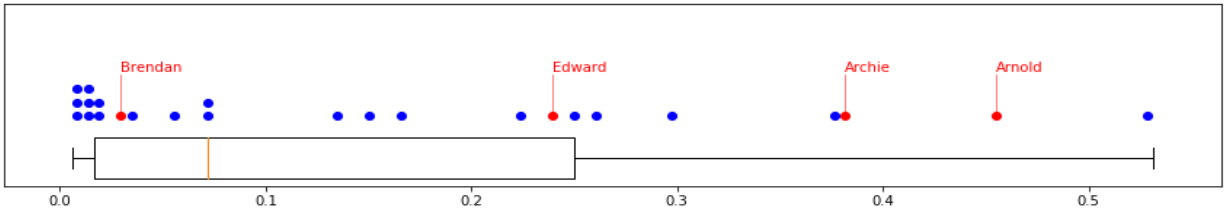


Fig. 6.20. Impulsive explorers relative to their peers on number of blocks found via taps (per slot)

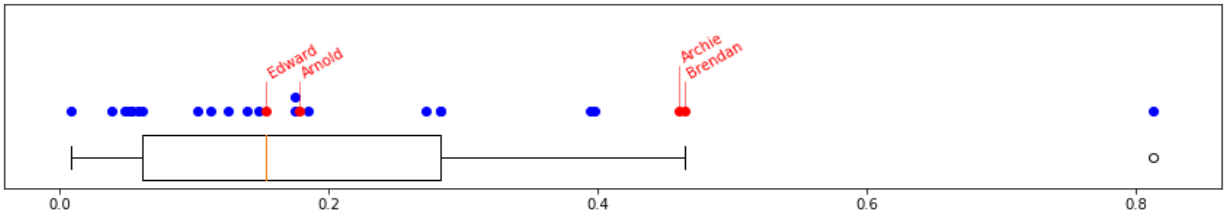


Fig. 6.21. Impulsive explorers relative to their peers on question button taps (per Slot)

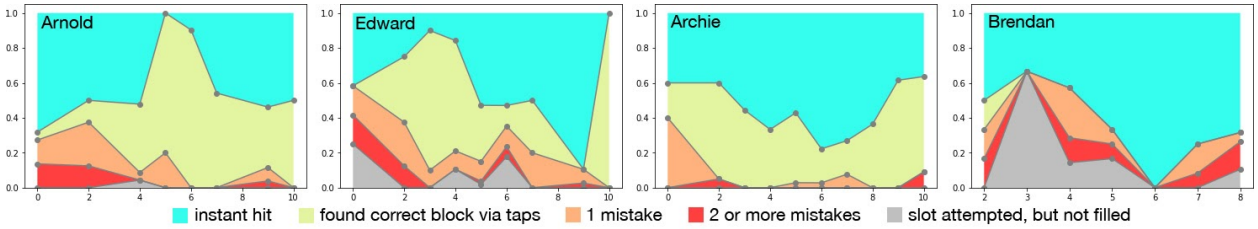


Fig. 6.22. Impulsive explorers statistics of slot filling (by week)

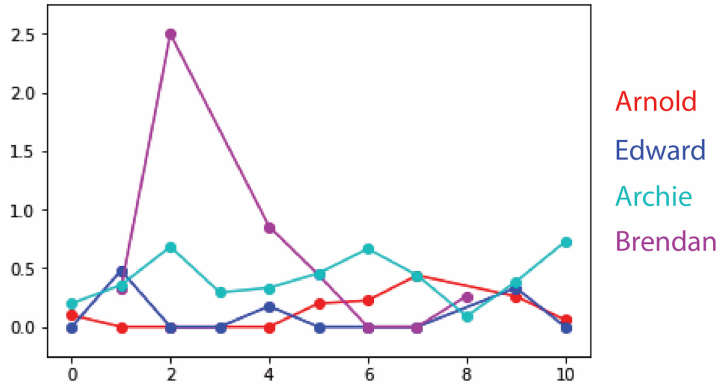


Fig. 6.23. Impulsive explorers statistics of question button usage (per slot, by week)

Unfortunately, the learning trends were less consistent for Brendan. He was never an active user of the probing strategy, but did actively use the help button for a period of time (Fig. 6.23). Then his usage of it reduced, but it seemed that it was being replaced with instant hits, the rate of which

reached 100% on week 6 (Fig. 6.22). This is remarkable, because observations collected early in the study suggested that he had difficulties matching sounds to blocks even when he recognized sounds in the word correctly. His sound-to-grapheme matching skill seemed to improve a lot. However, such deliberate efforts slowed down his word building significantly: on week 6, he was only able to make two words. Possibly because of that, on subsequent days, he started to use random guessing again, and his automatic correction rate started to go back up (Fig. 6.22).

Observations of Brendan showed that he was frequently frustrated — not only with building words, but with operating the sophisticated medium in general. Early in the study, he became lost navigating different screens and spent the entire session impatiently trying to explain what he wanted with phrases like: “I want to do sounds” and “No! The blue thing!” (referring to the turquoise background of the scaffolding screen). During another session, after seeing the demo involving a boy and a girl standing under a rainbow, the boy exclaimed: “I want to make people big!” The facilitator asked him: “What are you trying to spell?”, but Brendan continued to exclaim with growing frustration: “But I want to make people big! I want to make people...” Eventually he said: “I want to make rainbow boy.” The facilitator and Brendan started to make RAINBOW together, but after a few steps, the boy said that he wants to build ELSA (a character from the animated movie Frozen) instead. After experiencing a few issues with building the character (one of which was unfortunately caused by a bug), the boy passed the tablet to the researcher and said: “Make it!” However, by that time, the end of the session came. Brendan pulled out the headphones from the audio jack and threw the tablet on the table in frustration; then he angrily pulled on the headphones’ cord, as if wanting to break it. In one case, struggling with technology may have pushed Brendan to use the more docile paper medium instead. He put SpeechBlocks aside and started to copy words from the book *Fox the Tiger* (Fig. 6.17, (1)).

Another frustrating encounter with imperfections of the technology occurred when Brendan tried to use an early version of the speech recognition interface. In that version, the player needed to hold the recording button throughout the duration of speech. Brendan made several attempts at using speech recognition, but did not manage to align his voice and holding of the button, so the system kept returning random results. Meanwhile, one of his friends showed him a heart she made with the help of the system. Brendan gave speech recognition another try: “Heart! Please give me a heart!” When a set of random results appeared again, he exclaimed: “Noooooo! Nooooo, please!” and switched to random activities with the app.

These incidents with Brendan shows how important it is for a learning technology to have the interface that minimizes the need for focused effort. While children with higher levels of executive function (such as Ananda and Jacob) didn’t have much trouble with the early speech recognition interface, it presented another barrier to meaningful interaction for impulsive explorers. Following these observations, the speech recognition interface was adjusted to automatically detect the intervals when the child was speaking. After that, Brendan was able to use it successfully.

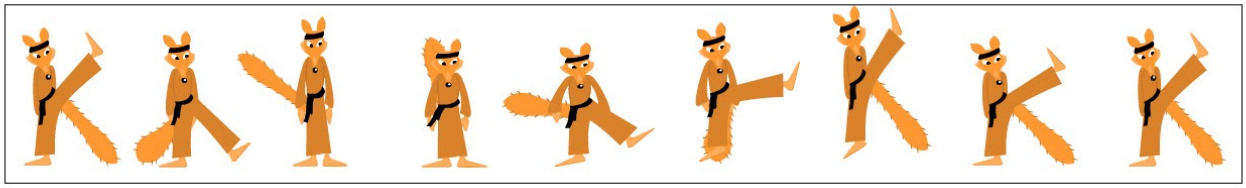


Fig. 6.24. "I'm kicking my tail": animation for phoneme [k]: Kathy the karate kicker (paired with grapheme K in this example) that evoked a curious reaction from a child

While struggling to enjoy their interactions with SpeechBlocks in the same way as imaginative-players or word collectors did, impulsive explorers came up with alternative ways to have fun. Edward and Archie were observed playing while holding the devices upside down, and Archie was even able to successfully spell words this way. They responded emotionally and at times viscerally to the animations of sound creatures. For instance, Edward jumped in his seat when he saw the snake animation for *[s]* sound for the first time, feigning fear of the snake. Arnold wiggled in his chair while watching the animation for *[w]* sound ("Wally the boomerang spins") and afterwards lifted his tablet in the air and wiggled it in a similar manner. Archie heard the prompt for *[eɪ]* sound: "A waves to a friend" — and waved as well. Edward was very amused to see the karate animation for *[k]* sound, and declared: "He is kicking his own tummy!" He kept returning to that creature on many days and narrated what he saw from the first person point of view: either "I'm kicking my tummy!" or "I'm kicking my tail!" (Indeed, the animation was construed in such a way that it could appear that the creature was kicking its own tail, Fig. 6.24.) On one occasion, Edward even enacted this by pretending to hit himself in the cheek with his fist.

Impulsive explorers occasionally engaged in simple, but passionate, enactive play with however limited sets of props they were able to make. For instance, after picking TIGER to spell, Archie loudly exclaimed: "Tiger, RAWR!" Edward made FLASH and started moving it around the screen rapidly, saying "He is so fast!" On another occasion, he gave a post-hoc interpretation to a set of sprites he made (two BATMANS): "They are sister Batmans. Actually." Archie scanned ANT from the box with the *Ants in Pants* game on a shelf, made it, and put it in his album. He then said to his peers: "Guys, look at this! Look at the bug! I'm hiding from you, bug!" He enlarged the ant, and said from the ant's perspective: "I'm going to bite; I'm humongous!"

On several occasions, children were observed "gamifying" their play with SpeechBlocks by inventing rules and rewards. Archie and Edward were seen arranging unrelated sprites on a page strictly on top of one another, as if playing a game. After completing the word ANT, Archie grabbed the *Ants in Pants* box and said: "Guys, this is how you earn the ant." The idea of "earning" sprites was expressed by another child as well. After completing a word that she picked up through text recognition, she excitedly exclaimed: "I earned it! I earned it!" This idea might have been borrowed from video games.

Another way for impulsive explorers to have fun was to involve other people — peers and researchers — into their play. Archie, for instance, at times bent over and randomly tapped on his peers' screens; a researcher had to intervene to prevent disruption of the peers' play. While all

children enjoyed showing their creations to their peers, impulsive explorers also rushed to show other things that attracted their attention within the app, such as interesting sound creatures or classmates' names. In their desire to share something, they sometimes removed their headphones and offered them to other children. Two children were observed swapping their headphones frequently. In doing so, they were sometimes confused as to why they could not hear sounds from their own tablet anymore, and after each exchange, their headphone cords had to be disentangled in order for them to continue individual play. Archie and Edward repeatedly offered headphones to the adults to include them in their play. They also tried to pique the adults' interest by saying, "You'll never guess what I'm making!" In one case, Archie declared that he wanted to spell something as a surprise for the researcher. He protectively held the tablet close to his body and said several times: "You can't look!" When the researcher said: "I don't know what you are spelling", Archie responded: "Justice! League of justice!" He, however, was actually in the process of spelling ELEPHANT for a wildlife scene.

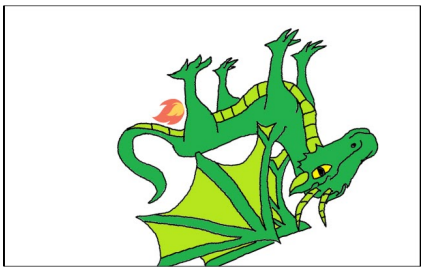


Fig. 6.25. Archie's dragon scene

It appears that on one occasion, Archie attempted to involve the researcher in the play by focusing on "teasing" her. The "teasing" commenced through exploring the theme of bodily functions that he thought the researcher might find inappropriate. We already took a brief look at this example while discussing the theme of children's phantasmagoria; now we shall have a closer look. Archie started by building DRAGON and "walking" it around the screen while saying loudly in a low, gravely voice: "I'm a destroy dragon! Big, big, big! Giant! Giant-er! I'm humongous ugliest beast." He then noted: "Dragons fart." He proceeded by building FIRE and saying "Fire in my chicken nugget!" ("Chicken nugget" was a catchphrase used by Archie on many different occasions and could have been an euphemism.) He placed FIRE next to the dragon's mouth and announced: "Fire in my mouth!" However, he soon flipped the dragon, moved the fire under its tail and reiterated his earlier statement: "Dragons fart!" (Fig. 6.25) He raised the degree of drama by exclaiming: "Fart in my face!" — while moving his tablet around in the air. A peer responded: "Ew! Don't say that!" — which seemed to delight Archie. He continued to act silly by selecting the *[r]* sound and saying "R starts with chicken nugget." At the end of the session, he showed his scene to the researcher and once again pointed out: "Look! Dragons fart."

Yet another unintended way to have fun with SpeechBlocks practiced by impulsive explorers was "probing" both the software and the hardware via unusual interactions and seeing what would come out of it. For instance, one child was curious to see what would happen if he long-pressed two sound creatures at once. Their sound pages opened simultaneously, and the content of the

pages jumbled up on the screen. The child was delighted by the glitch he caused. Other forms of probing interactions were “tickling” the screen with five fingers at once, tapping on multiple buttons in rapid succession, etc. Children also probed the hardware by pressing various buttons on the tablets and the headphones, unplugging headphones from the jack and plugging them back, and adjusting the headphones’ size. Pressing the power button caused the tablet to go into sleep mode, and a researcher’s help was necessary to return to SpeechBlocks. The call button on the headphones caused a minor distraction, since some children switched to a pretend play of making phone calls to each other. Some impulsive explorers did eventually manage to purposefully control the hardware to some extent, although not very confidently. For instance, Edward used the volume buttons to make the sound louder, but then complained that it was too loud.

Probing behavior poses interesting challenges to a developer. It can lead to bugs that adult testers might find hard to catch, because the latter do not usually think to use the system in this way. It also asks for an interface design that minimizes possibilities of unintended use. Educators who deploy expressive media in the classroom should look for minimalistic hardware that would not distract children with a selection of buttons to press.

Some of the alternative ways of using SpeechBlocks that children employed to entertain themselves were both pervasive and somewhat distracting for them. The first such activity was alternating between making one sprite extremely big and extremely small for a prolonged period of time. For instance, a child spelled EGG and made the sprite extremely large. She said: “I made it big and gigantic!” and showed it to a peer: “Look, I made a gigantic egg!” She then decided “Now I’m going to make it teeny-tiny”, and did so. She continued: “Now it’s medium big”, “Now it’s huge-huge-huge-big!”, “Now it’s small” — until her session ended. A variation of this activity is “hiding” things by making them so big that they appear as a monotonous field of color on the screen. For instance, Arnold spelled MOUSE, put it on the canvas and enlarged it so that the whole canvas turned into a uniform brown field. He then said to a researcher in mock confusion: “Where did the mouse go?” The researcher shrugged: “I don’t know...” Then Arnold rapidly shrunk the mouse to normal size, exclaimed: “Look! Here it is!” and smiled at his trick. Another child tricked a researcher by similarly covering the screen with a sprite of a black cat; he then turned to the researcher and said: “Look! My iPad broke!” Seeing the researcher rushing to get a spare tablet, he laughed and reduced the sprite to normal size. This size changing activity was exhibited by almost every child, aside from the most dedicated imaginative-players and word collectors, such as Ananda and Jacob.

The second distracting activity was using the camera preview of the text recognition interface to “take pictures” of other children. As described in section 3.2.4, the interface had a “freeze” button that made the image still. Its unintended side effect was the capacity to “take a portrait.” Children said to their peers words like: “Smile! I’m taking your picture!” The peers often paused their activities, smiled, and waved. The “photographer” then pressed the “freeze” button and showed the “portrait” to the peer. That often caused laughter and the desire to either continue posing, or to reciprocate by taking a picture of the “photographer.” Occasionally children also took “self-portraits” and showed them to peers. In our example group of four, Edward and Brendan

engaged in this type of play. To mitigate these effects, we started to display the "freeze" button only when there were words within the camera's field of view. That improved the situation, but often the system still picked up words on the classroom walls and allowed children to "freeze." Even if they were unable to "take pictures", they simply found it entertaining to look at their peers and the classroom through the camera's display. During the first few days children generally did not mind being "photographed" in this way, but later on some of them started to express discontent with others "taking pictures" of them. They told their peers to stop — a request that the peers often disregarded. Thus, the camera feature is, unfortunately, not only a potential distractor, but also a potential source of tension between children.

There were other secondary uses of the interface, less disruptive, but still not involving building words. Several children, such as Brendan, enjoyed walking around and *detecting* words. They pointed the camera at some text and waited until the words turned green. As soon as that happened, they moved on without trying to figure out which words had been highlighted. This interaction was purely a play with technology and did not involve a literacy component.

There are some quantitative traces of the behaviors described above in the log data. To estimate "big-smalling" behavior, I have looked at the distributions (one for each child) of scales that various sprites on the canvas have assumed throughout their lifetime. Intuitively, the bigger is the entropy, the higher is the spread of scales, and the more abrupt are changes in the scales. This corresponds to the behavior of interest. I used the *m*-sample-spacing entropy estimator (Vasicek, 1976) to assess entropy without the need to assume an analytical expression for the distribution. In this approach, the entropy of a continuous random variable *e* can be estimated from N sorted samples using the formula:

$$H(e) = \frac{1}{N}\sum_{i=1}^{N-m} log(\frac{N}{m}(e_{i+m} - e_i)),$$

where m is the step between samples. The estimate converges to true entropy when $N, m \to \infty$, $\frac{m}{N} \to 0$. I used $m = 20$, since in my computational experiments I saw that entropy estimates stabilize after this number. To ensure that the results were not an artifact of the estimation approach, I also assessed the spread of scales in two different ways (through the entropy and the standard deviation of distribution of their logarithms), with the same results. There was a statistically significant correlation between the target measure and executive function, and an indication of the same trend (but not significant) for CTOPP (Fig. 6.27), even though the four exemplary impulsive explorers didn't end up being prominently positioned on this scale (Fig. 6.26).

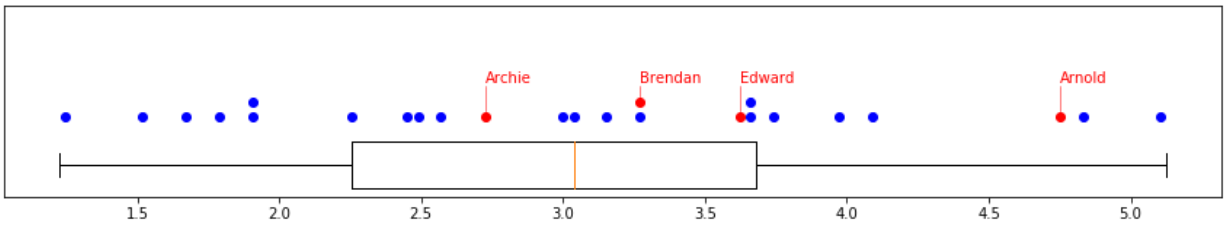


Fig. 6.26. Positions of impulsive explorers on the entropy measure of zooming behavior

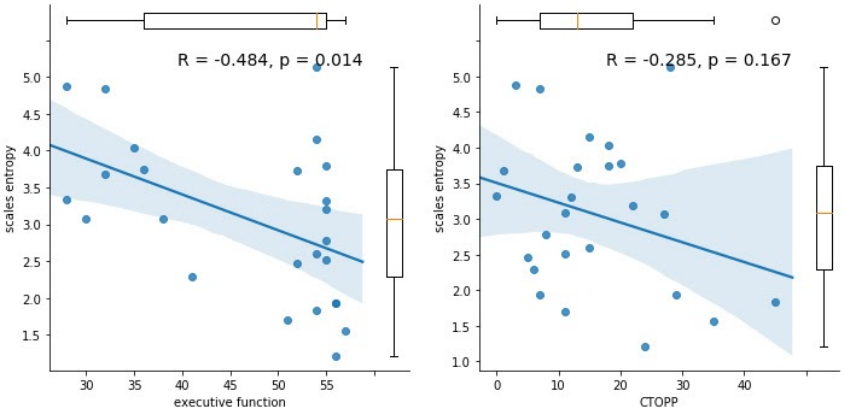


Fig. 6.27. The entropy measure of zooming behavior in relation to executive function and CTOPP

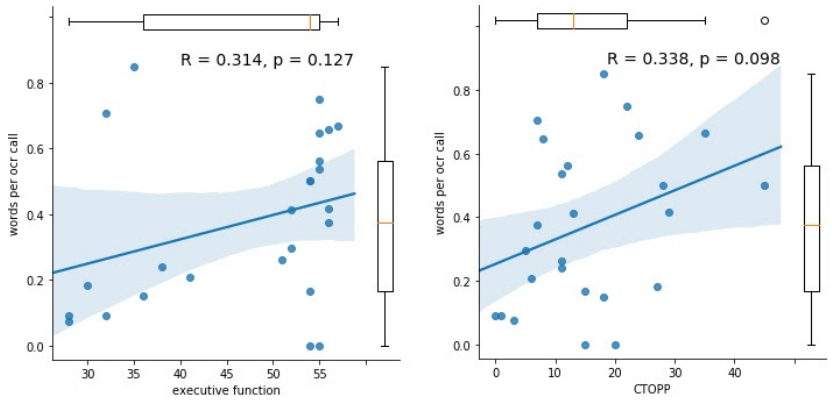


Fig. 6.28. Correlations of the fraction of text recognition calls that led to building a word with executive function and CTOPP.

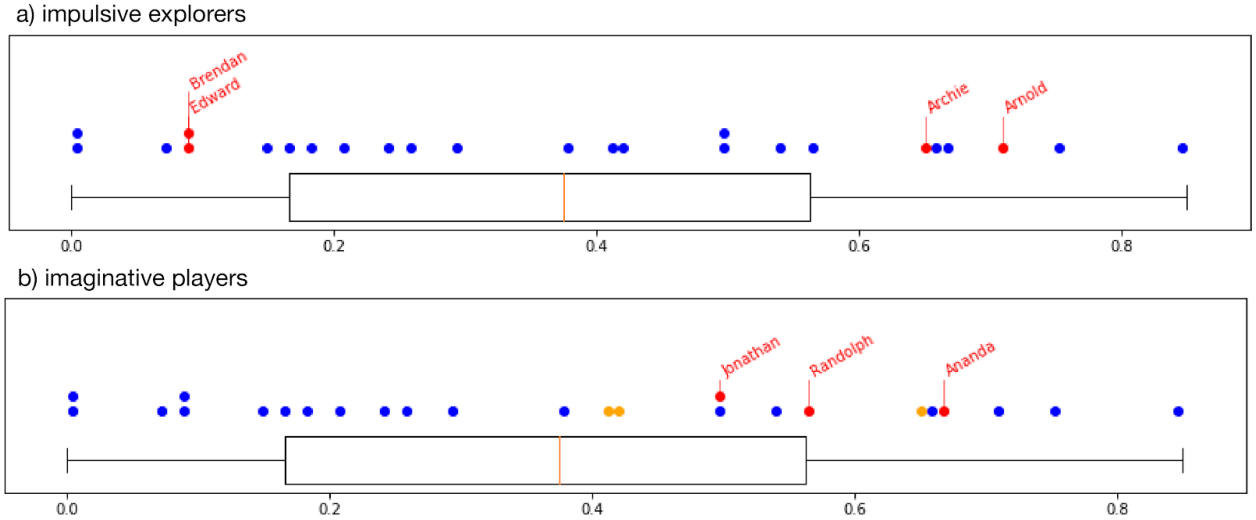


Fig. 6.29. Position of various children on the fraction of text recognition calls that led to building a word

To estimate the rate of non-purposeful use of text recognition, I looked at what fraction of text recognition calls resulted in word construction for different children. Indications of correlation of this measure can be seen with executive function and CTOPP (Fig. 6.28), although the correlation is not significant. The four sample impulsive explorers are not positioned in a straightforward way on this value's range. It is curious to note, however, that imaginative players are all above the median (Fig. 6.29). This fact fits the overall image of imaginative players as less chaotic and more goal-oriented in their usage of SpeechBlocks.

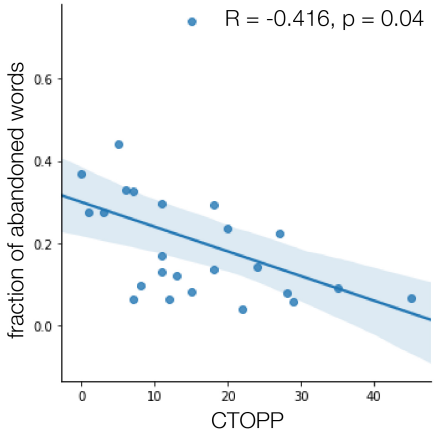


Fig. 6.30. Fraction of abandoned words in relation to CTOPP

Matching the above-mentioned tendency of impulsive explorers to frequently give up or switch plans, there was a correlation between the fraction of abandoned words (words which were started, but not completed, in scaffolded mode) and CTOPP, which can be observed on Fig. 6.30.

Interestingly, prominent patterns corresponding to low CTOPP and executive function scores also manifested in children's touch interaction with the devices. I measured the mean numbers of touches per session, the mean durations of touches, and the mean finger speeds and accelerations. To account for both speeding up, slowing down, and turning, I computed accelerations as magnitudes of vector derivative of speeds. When computed in this way, higher accelerations correspond to "jerkier" finger motions. We see that lower CTOPP and EF scores are associated with more numerous touches (Fig. 6.31, EF—approaching significance; CTOPP—significant) which are shorter (Fig. 6.33, significant for both). Corresponding finger motions are faster (Fig. 6.35, significant for EF) and "jerkier" (Fig. 6.37, significant for both EF and CTOPP). Our four sample impulsive explorers tended to have a lot of touches per session (Fig. 6.32) and are particularly prominently positioned on the scales of finger speeds and accelerations (Fig. 6.36 and 6.38), being at the top of both scales. One way to interpret these results is to presume that these measures capture various qualitative behaviors described in this section — "tickling" the screen, randomly tapping on the keyboard, probing different buttons, and quickly zapping sprites around. The general picture which emerges in this case is that lower executive function and lower CTOPP tend to be associated more with such interactions. I need to qualify, however, that I haven't observed similar results in the second home study with SpeechBlocks I, where CTOPP data was also available.

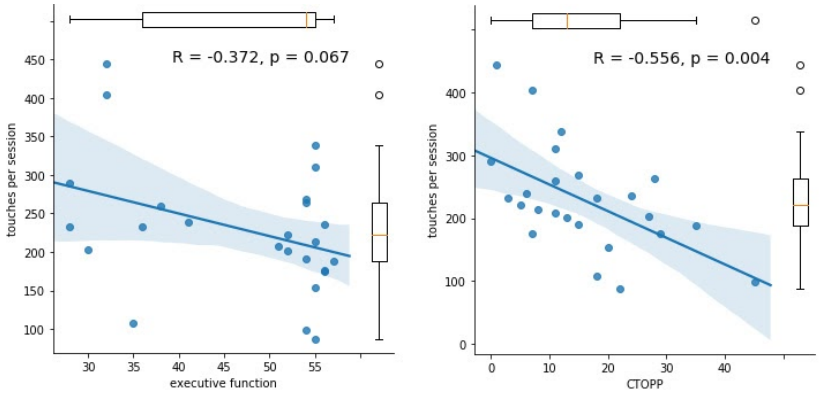


Fig. 6.31. Number of touches per session in relation to executive function and CTOPP

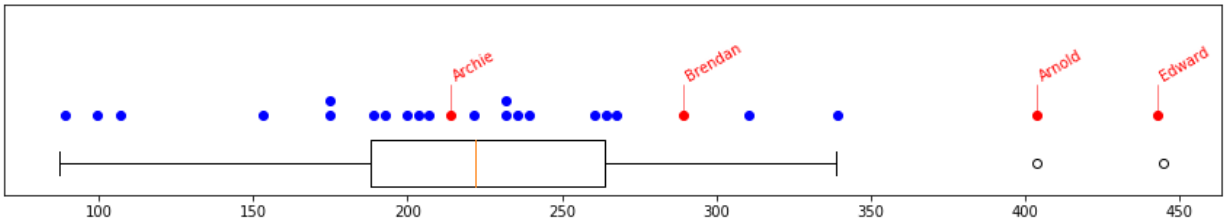


Fig. 6.32. Position of impulsive explorers on touches per session

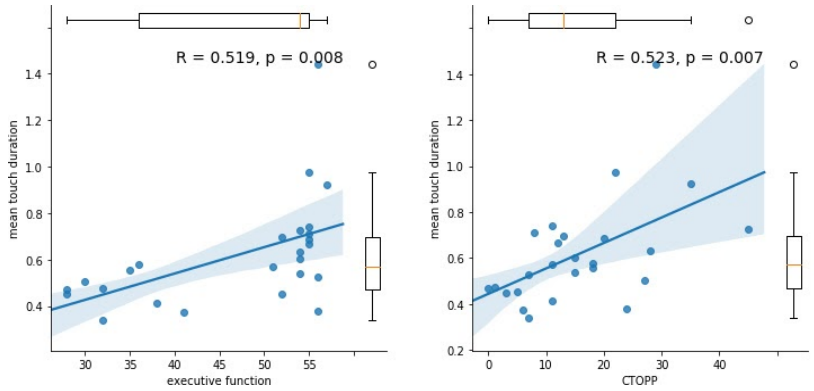


Fig. 6.33. Mean touch duration (in sec) in relation to executive function and CTOPP

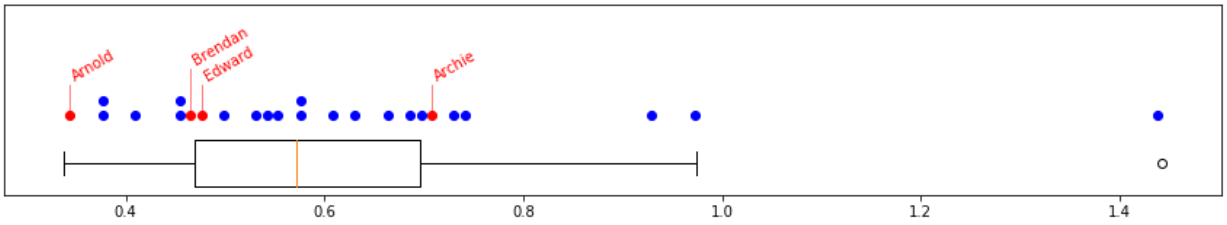


Fig. 6.34. Position of impulsive explorers on mean touch duration (in Sec)

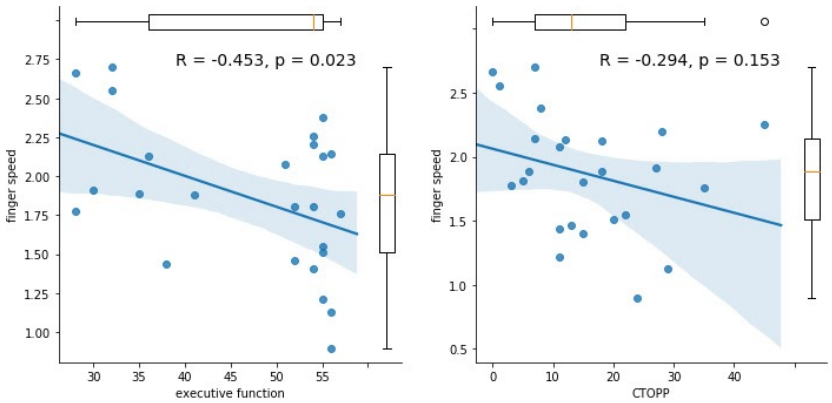


Fig. 6.35. Mean finger speeds (in inch/sec) in relation to executive function and CTOPP

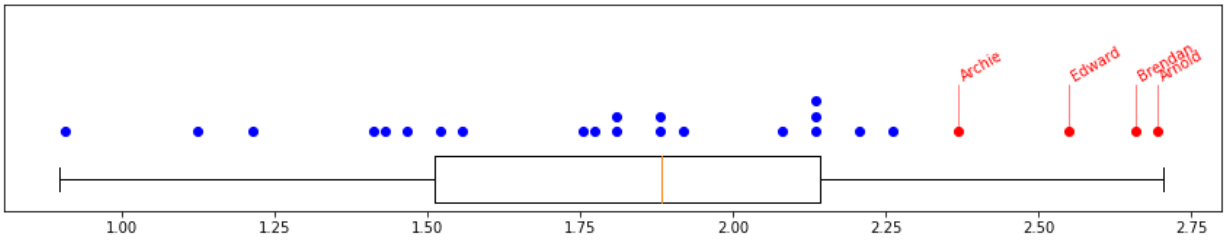


Fig. 6.36. Position of impulsive explorers on mean finger speeds (in inch/sec)

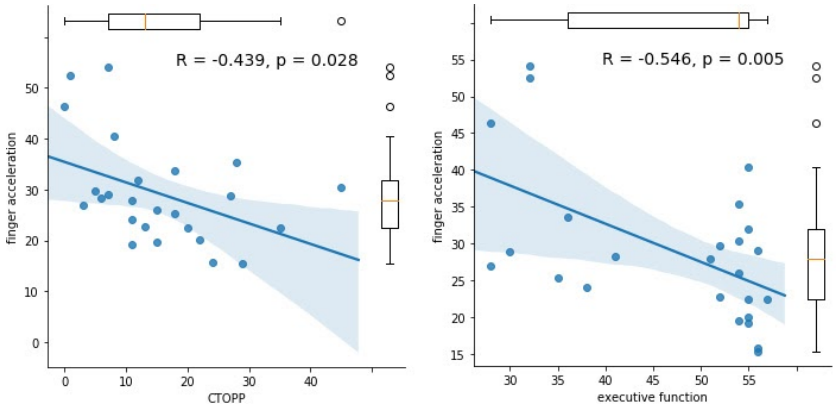


Fig. 6.37. Mean finger accelerations (in inch/sec²) in relation to executive function and CTOPP

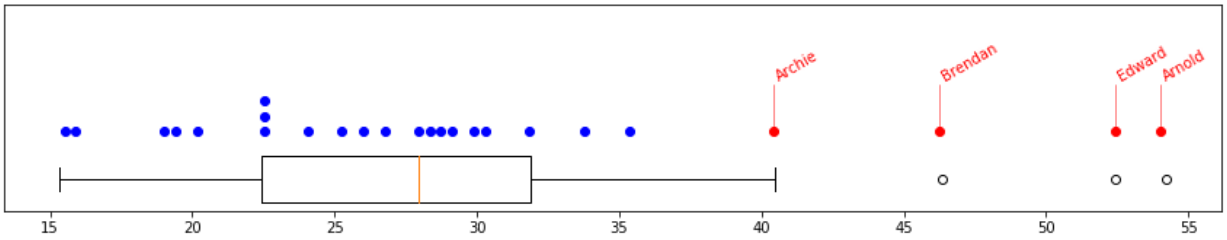


Fig. 6.38. Position of impulsive explorers on mean finger accelerations (in inch/sec²)

6.3. Agency, Self-Efficacy and Ownership of Work

Similarly to SpeechBlocks I, children frequently displayed behaviors related to their agency, self-efficacy, and ownership of their work. With the design of SpeechBlocks II offering new capabilities, these behaviors manifested themselves in some new ways.

While designing SpeechBlocks II, I was concerned that the sense of self-efficacy would be inhibited by scaffolding mechanisms, because children would not perceive words as their own work anymore. Qualitatively, this did not appear to be the case. Children still frequently produced expressions related to self-efficacy and referred to words as their own creations, e.g. “Look! I spelled SHARK!” Impulsive explorers often had exaggerated displays of self-efficacy: they kicked, stomped, jumped, and exclaimed “Yay!” after finishing words they intended to build. This may be a reflection of both their less quiet nature and of the greater effort that they had to put into constructing words.

In a limited number of cases, children apparently intentionally challenged themselves to master a difficult task by deliberately avoiding scaffolding. For instance, children sometimes removed the headphones to make sure that they received no sound prompts from the scaffolding system. When one child finished building a word in this manner, he said: “I did it all by myself!” Yet another child refused to use speech recognition to make a word, and asked how to make it “with words” (he actually meant “with letters”, in free mode). Children experimented with making familiar words (typically their names) in free mode, sometimes going through a tinkering process before reaching the final result. Mary (a word crafter) tried to spell her name in free mode about 10 times over the course of different days, and eventually learned to do it. Another word that she repeatedly tried to spell might have been the name of her friend or relative. She also repeated some other letter combinations over and over again, suggesting an intent to make other real words. In several cases, after constructing a word in the scaffolded mode, children repeated the same word in free mode. Possibly, they were trying to solidify their knowledge, or simply found scaffolding excessive now that they knew how to spell the word. These cases hint the capacity of scaffolding to naturally go away once the child acquires sufficient skill to make words on their own.

Just like in SpeechBlocks I, children tended to keep the words (and associated sprites) they made, using the canvas instead of the word drawer as a means of storage. In the case of several impulsive explorers, we observed their apparent pride in amassing large amounts of sprites on one page. This was another self-efficacious behavior, since the players apparently perceived the mass of sprites as a testament to their hard work. For example, several times Edward proudly told a researcher that there is no more space left on his page. When Archie similarly started to run out of space on his page, a researcher tried to show him how to start a new one, but Archie immediately went back. After a few minutes he proudly showed his crowded page to the researcher and said: “A lot of stuff!”

Children's ownership of their work manifested in their delight at seeing the scenes they made. For example, one child, noticing for the first time that SpeechBlocks saved the scenes he did during the previous sessions, exclaimed in amazement: "Wait a minute! I did this!" He scrolled through more of his scenes and shouted: "I did this too!" Similar behaviour was observed when we printed out the scenes children made and brought them to the classroom. Children immediately started to recognize the scenes they made: "This one is mine! I made this." They were also curious about who made the other scenes. Perhaps the most interesting form of this behavior was exhibited by a child who said to his peers: "Look! This is all my stuff! This is my page! This is my book!" It appears that he conceptualized himself as an author.

Children's sense of ownership over their works was also indirectly illuminated by their reactions to several incidents. First, when Randolph's and Jonathan's tablets accidentally got switched, Randolph asked his peer: "Are you using my tablet? On my account?" On a different day, a software glitch caused all saved scenes to be lost. A child noticed the loss and asked: "Where did my little scene go?" On a third occasion, the same child said to a researcher: "Next time, give me the right iPad with my pictures in there."

6.4. Social Play

Similarly to what we observed with SpeechBlocks I, children playing with SpeechBlocks II in groups actively engaged in social play. This play can support children's play by serving three functions for learning: (1) inspiring each other's ideas; (2) maintaining mutual engagement; (3) directly learning from each other. However, the different capabilities of the app — in particular, the addition of scaffolding and imagery — introduced new forms of social play. Let us have a look at how this play occurred using SpeechBlocks II.

Inspiring Peers' Ideas

Sophisticated forms of idea exchange accompanied imaginative play. The group of four children who were the most prolific imaginative players exchanged ideas frequently. Throughout fifteen days in which they were building scenes, their play was only disjointed for a single day. Below is a few examples of idea exchange between Ananda, Jonathan, and Randolph.

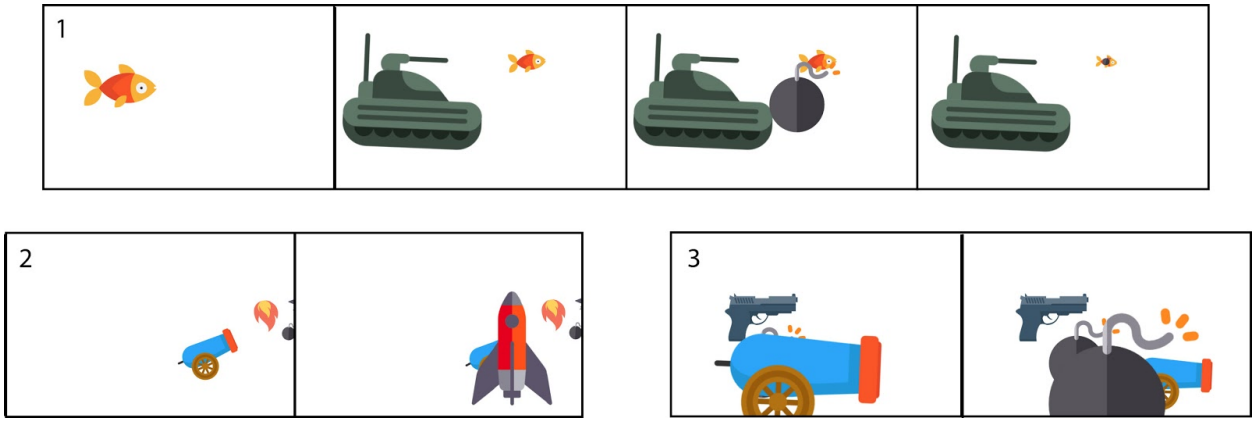


Fig. 6.39. Interaction between players - example 1. (1) Ananda's play sequence. (2) Frames from Jonathan's and (3) Randolph's play sequences on the same day.

On the day Ananda built her fish and tank scene (Fig. 6.39, (1)), the demonstration conducted by the researchers involved building a marine scenery. Following that theme, Ananda built FISH and attempted to build OCEAN using invented spelling. However, two of her friends instead focused on the theme of "explosions" and "blasting", as they referred to their work (Fig. 6.39, (2) and (3)). They excitedly chatted about what they made. Jonathan showed his scene to Ananda and according to him, it was a rocket ship blasting aliens. Ananda responded with a similar idea: "I'm going to blast the fish!" She made a tank, and said: "Tank! I made this big tank! I'm going to blast this fish!" Then she repeated it to Jonathan: "Look [at] it! I'm going to blast the fish with the tank! I'm going to blast the little ball!" Afterwards, she asked her friend how he got to the word BOMB within the app. However, Jonathan was engrossed in his play and did not respond. Instead, Ananda created the word via invented spelling (with some help of a research assistant), added the BOMB to the scene, and placed it over the FISH. She showed the composition to Jonathan and said: "Look [at] it! He [the fish] is scared."

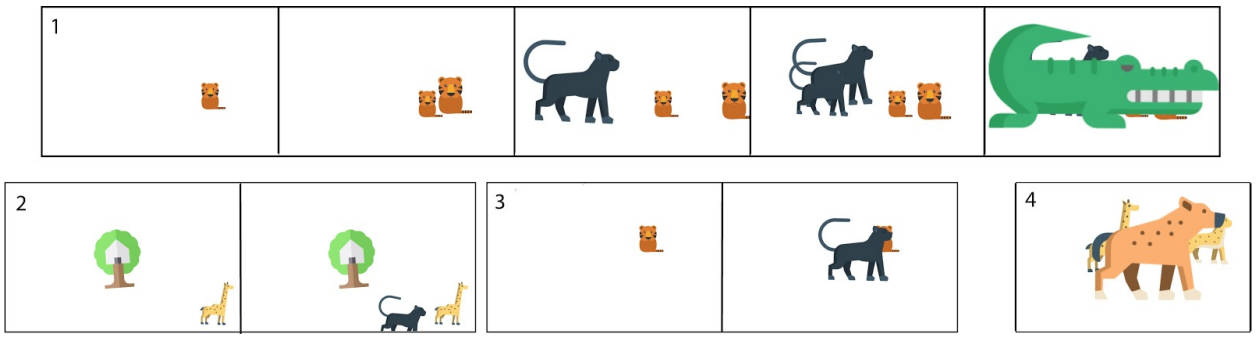


Fig. 6.40. Interaction between players - example 2. (1) Ananda's play sequence. (2) and (3) Fragments of Jonathan's parallel play sequence. (3) A shot from Randolph's play sequences on the same day. Jonathan's frames in (3) are slightly edited by removing irrelevant sprites.

During construction of Ananda's animals scene (Fig. 6.40, (1)), she first built two tigers (using speech recognition) and then added a group of panthers sneaking up on them via associations.

She showed the composition to Jonathan and enacted the scene by saying: “Oh! Run away!” to her tigers. Her friend, who by that time had built a giraffe, immediately embarked on building his own panther (via speech recognition). Once it was finished, he showed his scene to Ananda and said: “Look! He [the panther] is eating a giraffe. His [the giraffe] whole body is eaten” (Fig. 6.40, (2)). He then added his own tiger (also through speech recognition) and said: “This dude [panther] is going to eat this guy [the tiger]!” — and, assuming the role of the panther, said to his tiger: “I will eat you alive!” (Fig. 6.40, (3)) Ananda looked at his scene and pondered: “What can I do? Alligator? Crocodile? To eat them up.” She decided in favor of the crocodile, used speech recognition to obtain the desired word, and then enacted the crocodile eating the other animals (Fig. 6.40, (1), the last frame). She returned to that scene and continued enacting the same story more than a month later. Notably, Randolph was also building a wild animals scene on that day (Fig. 6.40, (4)).

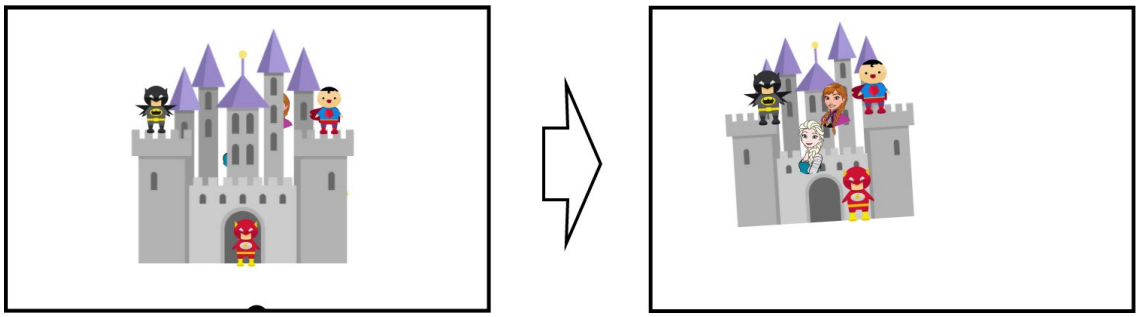


Fig. 6.41. A scene built by Ananda and its imitation by another child

Transfer of ideas occurred not only between children sitting at the same table, but also, although to a much smaller extent, via printouts of scenes that we made. For the printouts, we selected scenes that we considered most sophisticated, well-structured, and beautiful. These well-executed scenes attracted a good amount of children’s attention, and in rare cases led to imitation. One such imitation is shown on Fig. 6.41. Children also occasionally suggested ideas to each other. For instance, when Ananda made a dragon, Jonathan suggested that she should make the dragon eat someone. However, we have not seen any occasion where a child actually followed a suggestion given to them by a peer.

Maintaining Mutual Engagement

A new social game children played using SpeechBlocks II was to jointly decide on complimentary words to build. For instance, one child proposed to another: “I’m going to spell your name, and you should spell my name.” Her friend agreed. After they both finished with their respective words, the friend proposed: “Now let’s both spell Abigail’s name.” Both children were excited at these ideas, and there is little doubt that it contributed to their engagement.

Another component of maintaining engagement was being an audience for each other. Understandably, it is much more motivating to build something when you can share your creation and as expected, children actively shared the words they made. Imaginative players could demonstrate not just standalone words, but entire scenes, which their peers sometimes watched

with interest. For instance, when Ananda said that she was going to put a NINJA in her CASTLE, Jonathan responded with amazement: “Ninja in the castle!” Ananda repeated: “Ninja in the castle!” and laughed. A few minutes later, she said: “I’m going to put... Look [at] it, look [at] it, look [at] it!” — requesting Jonathan’s attention. Jonathan did look, and soon announced his own development: “I put a rocket ship in the castle.” Ananda responded with something indistinct, but excited.

Children also talked about the words and the scenes they made, both with peers and researchers. For instance, Zack told us, after making MOUSE: “My mom is scared of mice, but I’m not. Even my grandfather and grandmother [are scared].”

Learning from Each Other

Peers often assisted each other in building words or built them jointly. It is likely that this process was catalyzed by the presence of the direct guidance system, which reduced the complexity of building words and thus also made it easier for children to help one another. Sometimes we saw that children not only told their peers what to do, but took steps to make sure that their peers would be able to do it on their own in the future. For instance, a peer assisted Edward in building BATMAN when Edward complained that “BATMAN is not working.” He highlighted to him the key aspects of the building process, such as the need to place correct blocks into the slots and to pay attention to the sounds: “You just gotta spell it! Can you hear it?” He started to help Edward by pointing to the blocks he needed, and when Edward got distracted, asked him to find the next block himself. When Edward hesitantly pointed to the correct block, his peer confirmed: “Yeah! That guy.”

We saw several naturally formed pairs of children in which one child learned from another by imitation. “The follower” repeatedly chose to build the same words as “the leader” in the pair, and if s/he encountered any difficulties, s/he asked “the leader” (who by that time had completed creation of the same word) what to do. Such pairs were engaged in active conversation about their process, what they made, and what they should make.

Some children acted as “literacy experts.” One such expert was Ananda. For instance, on one occasion Ananda and Randolph each started building Jonathan’s name (as a result of a joint deliberation on what to make); with her more refined literacy skill, Ananda progressed through the word faster. At one moment, Randolph bent over to see which blocks she put together, mumbling “Let’s see... JONA...” Ananda responded: “Let me show you. Put TH there. The T and the H. No... Yes. Good, there.” Starting construction of the next word, Randolph notified Ananda right away: “I’m on the same one as you”, and she helped him again: “Put the S there. No, I mean N. [Now] put the D there.” Next, Jonathan asked Ananda: “What is the last letter of your name?” — and she helped him, too. Then, Randolph showed Ananda his scene and explained how he built it: “First I did FLASH, then BATMAN, then SUPERMAN, then I did that DRAGON, then…. Do you know how I spelled FLASH?” Ananda paused to think for a little, then replied sound-by-sound: “[f]... [l]... [æ]... [ʃ]... Yes, I do.” It appears that in this exchange, Ananda enjoyed being a literacy expert, while

Jonathan and Randolph acknowledged her competence and were glad to receive support from her. In turn, Ananda inquired of her friends about certain aspects of using the software — for instance, how to invoke the scaffolding for BOMB, as shown in one of the examples above.

6.5. Building Words With and Without Automatic Scaffolding

The key innovation in SpeechBlocks II was the introduction of built-in scaffolding. Scaffolding appeared in two forms: as special blocks (phoneme blocks, onomatopoeic mnemonics) and as automatic scaffolding routines. This section examines the effects of the latter on children's play, while questions pertaining to the former are discussed in the next section. Here, we will look both at how scaffolding routines were used and how their introduction affected open-ended word building. This analysis is performed through the lens of so-called word sources. Word source denotes a way through which construction of a word originated, e.g. via open-ended mode, word bank, speech recognition, etc. Each word source was associated with different play patterns. We will discuss both general learnings regarding each word source as well as ones pertaining to their specific design and implementation.

Scaffolded word construction dominated children's play, with almost 79% of words being made in direct guidance mode and an additional 6% of words being made via invented spelling interpretation. Open-ended mode was primarily used for nonsense words. Scaffolded modes were likely popular because they supported children's desire to make specific words. We noted how in earlier versions of SpeechBlocks, adult's attention was a bottleneck for letting this desire come to fruition. By eliminating this bottleneck, scaffolding routines played an instrumental role in supporting imaginative play and word crafting, which both rely on making specific words of a child's choice. They also facilitated smooth flow of ideas between children and certain modes of shared play.

Looking closer at scaffolded word construction, we find that different word sources under this umbrella played different roles. Some were used with a pre-defined plan, while others were used in the process of tinkering. There were at least three important roles fulfilled by different word sources: (1) directly responding to specific children's ideas, (2) helping to generate new ideas, and (3) serving as a fallback mode when a child encountered issues with more sophisticated systems. Because of these complementary functions, different word sources were sometimes used synergistically. This synergy helped the creation of sophisticated scenes.

Various word sources performed differently in practice. Speech recognition turned out to be an ideal word source for children who had specific words in mind, despite its less than perfect accuracy. Associations network fulfilled the complimentary task of being the main word source for tinkering. Word bank played an important role by being an easy to use and reliable word source that children could fall back on if they had issues with more complicated mechanisms. On the other hand, usefulness of some word sources turned out to be questionable in light of issues that showed up during the study. While text recognition did prompt children to explore the texts

surrounding them, it was also challenging to use and generated a large amount of distractions due to non-purposeful uses. Some observations of play suggest that object recognition might be a more natural word source than text recognition for children of the target age. Invented spelling interpretation also ended up being challenging to use, primarily because of children's limited phonological awareness. This situation might be different for slightly older children.

6.5.1. Popularity of Different Word Sources

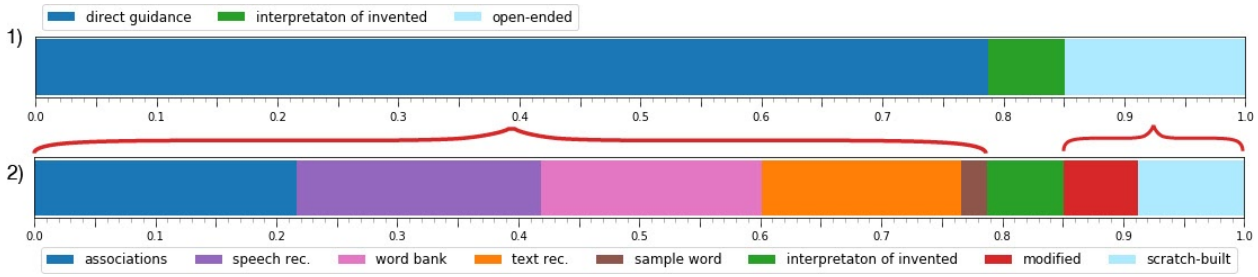


Fig. 6.42. Fraction of constructed words by source, averaged among children

Popularity of each word source is a proxy to how much benefit children perceived in using it. Fig. 6.42 shows what fraction of words originated from each word source during the last two weeks of the study. The top half (Fig. 6.42, (1)) shows the breakdown into the three principal categories: open-ended words and two principal scaffolding types, invented spelling interpretation and direct guidance. The bottom half (Fig. 6.42, (2)) breaks down open-ended words into scratch-built ones and modifications of scaffolded words. It also breaks down the direct guidance words by what the guidance routine was called. One can see that an overwhelming majority of words were produced in direct guidance mode. Compared to the share of direct guidance, open-ended mode was not very popular. Particularly unpopular was the invented spelling interpretation mechanism. Further breakdown shows that associations, speech recognition, text recognition, and word bank enjoyed roughly comparable popularity. And although text recognition was sufficiently popular, there were some problematic effects seen during further quantitative analysis.

A note needs to be made on technical details of how this popularity was computed. Readers not interested in these details can skip this paragraph. First, we needed to take into account that word sources were introduced gradually over the course of the study. So, to compare them more fairly, we needed to look at the statistics from a period when children were sufficiently familiar with the usage of each word source. Assuming that a week (two sessions) was needed for a child to gain such familiarity, that gave us the last two weeks of the study as the target period. Second, we needed to define what should count as one word built in open-ended mode, because there was no natural start and end of the process of word construction in this case. For the purpose of this analysis, I decided to count construction of a single word as an interval from the moment an empty word box received an initial block to the moment the word box cleared again, or the session ended. That method of counting could, however, underestimate the number of open-ended words in the cases when a word was constructed and then modified into another word. To avoid such a

possibility, all traces of open-ended word construction have been manually examined, and no instances of such behavior were found. “Words” consisting of a single block were also excluded from the open-ended word count. In addition, there were cases when children used open-ended mode to modify a word built with scaffolding. Such cases were counted towards the number of open-ended words if (a) during modification, at least one block was added, and (b) the result of modification was not equal to the original word. These two rules were introduced to avoid counting (a) cases when the child simply manually cleared the word box, and (b) cases when the child accidentally removed a block from the box, but then restored the word to the original condition. Fig. 6.42 was obtained under these assumptions.

As mentioned earlier, guided mode dominated word building. Given that children were free to choose what mode of spelling to use, the popularity of this mode suggests its significant value to the players. This was the case despite the rigid nature of the guidance system, which might appear as less playful than open-ended tinkering. This popularity can be linked to children’s desire to build specific words of their choosing. Manifestations of this desire were already documented in the analysis of SpeechBlocks I play. In SpeechBlocks II, it was further fueled by imaginative play. However, as we saw in the SpeechBlocks I analysis, neither adult’s guidance nor word cards were sufficient in fulfilling this desire. Cards were limiting children to a small vocabulary, while adult’s limited attention turned out to be a bottleneck that caused children to compete for this limited resource and interfere with each other’s word building. Such interference occurred even when each child was building just a few words during a session. Automatic scaffolding allowed for much higher rates of word construction and eliminated frustration from constant mutual interruptions. For imaginative players, it opened a possibility to create sophisticated scenes consisting of 10 or more objects. For word crafters, it created a possibility to explore a multitude of interesting and unusual words. Not having to wait for an adult also allowed children engaged in both types of play to fluidly respond to emerging ideas. In this way, automatic scaffolding was instrumental in supporting these two types of play.

Automatic scaffolding also facilitated social play. It allowed for fluent borrowing of ideas from peers: children could easily copy a word they were interested in, or to alter their peer’s idea with their own word. Examples of such process are given in the section on imaginative play. Specific types of joint play also were supported by the scaffolding. For instance, children engaged in the following game: “I’m going to spell your name, and you spell mine. Then, let’s spell X’s name together.” This form of play is based on simultaneous spelling, which would have been impossible if children had to compete for adult’s assistance. Furthermore, the guidance system made it much easier for children to help each other, because it greatly reduced demands on the helper to know the correct spelling of the word s/he assisted with. In the end, the automatic guidance mechanisms increased children’s agency in the areas of play that seemed to matter most to them: creating scenes, stories and collections of words, and engaging in social interaction with other children via SpeechBlocks.

Different children gravitated towards different word sources, with no compelling overall patterns emerging. Fig. 6.43 illustrates this variation by showing the breakdown for various children

introduced in section 6.2. I compared usage of word sources by various groups of children: with high and low CTOPP, high and low EF, boys and girls, imaginative players, impulsive explorers, and word crafters. However, the variation between children was so high that it is hard to talk about general regularities.

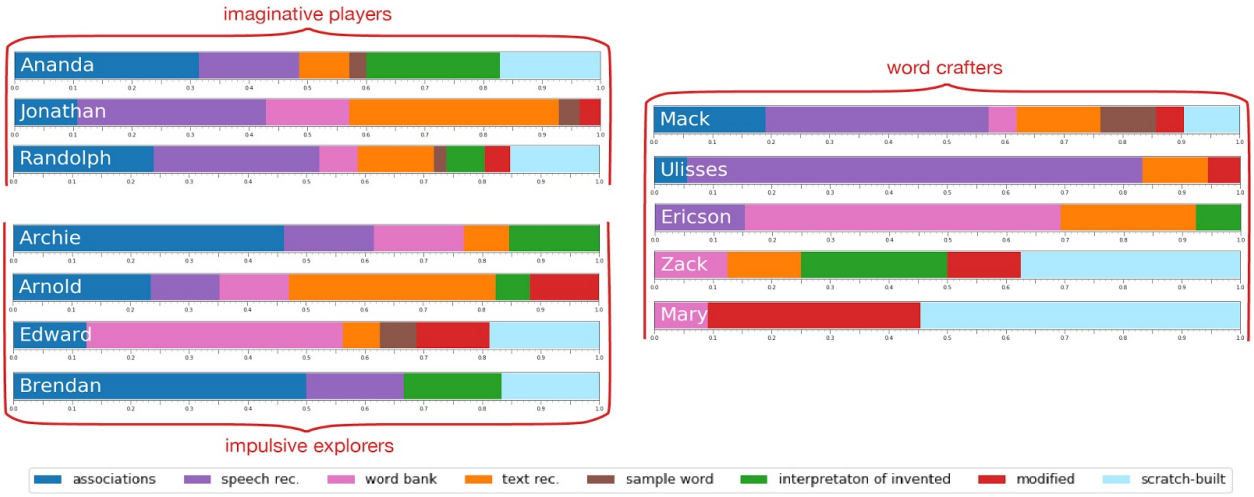


Fig. 6.43. Usage of word sources for different children

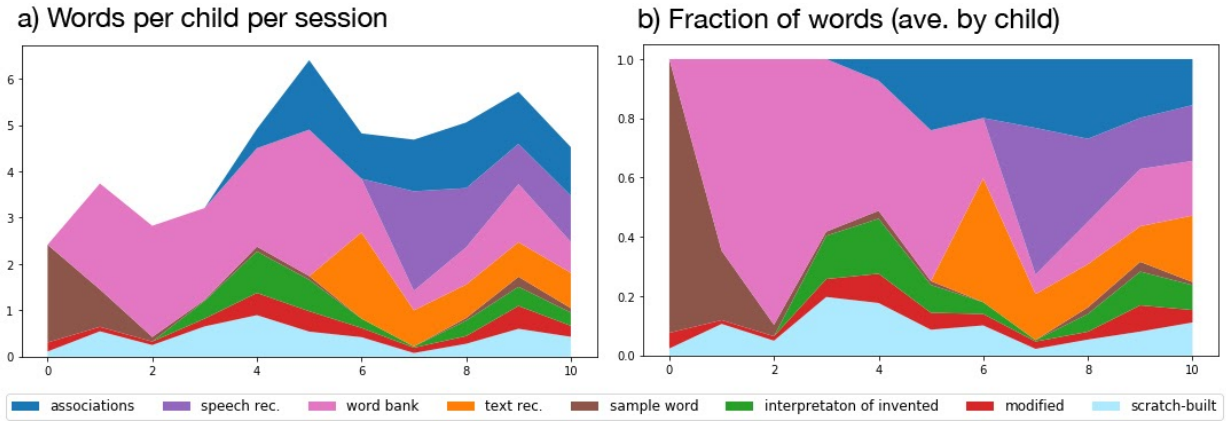


Fig. 6.44. Dynamics of word source usage by week

Fig. 6.44 shows how usage of different word sources developed over time. Several aspects are of interest. The first is the pronounced notch in word building starting at week 6, when text recognition was introduced. This is a consequence of various non-purposeful uses of text recognition, which created significant distraction. Some examples of distracting uses were mentioned in section 6.2.3, and will be examined further in section 6.5.7. Second, the use of sample words (appearing on the phoneme and letter pages) as a word source dries out almost as soon as alternatives to them are introduced. This is understandable: sample words were intended to be used as a means for children to get to know letters and sounds, and not intended as a word source. Third, there is a gradual displacement of word bank as more “high-tech” scaffolding inputs are introduced. This is likely the consequence of word bank’s limited vocabulary. Yet word bank is

still actively used even at the end of the study, likely because of its relatively convenient and quick-to-use interface. Fourth, during weeks 3 and 4, one can see an uptake in word sources other than direct guidance. This is the consequence of the researchers actively encouraging children to try invented spelling. Once the encouragement stopped, the fraction of these word sources dropped. Fifth, usage of each word source peaks shortly after its introduction, reflecting the novelty effect, and stabilizes afterwards.

The following will consider the specifics of usage of each word source.

6.5.2. Sample Words

Sample words on letter and phoneme pages were intended to help children get familiar with sound creatures. They were not intended to be used as a practical word source, and indeed they weren't. Children only used them serendipitously, when they accidentally stumbled upon a word they liked.

6.5.3. Open-Ended Mode

About 1 in 6.5 words constructed by children either was entirely built in open-ended mode, or was an open-ended modification of a scaffolded word. A vast majority of these words were random or patterned combinations of letters. Unfortunately, limitations of the word box interface precluded children from making nonsense words by remixing real words — an activity that was popular and much enjoyed with SpeechBlocks I. Most nonsense words in SpeechBlocks II were apparently created for their own sake, as a simple, impulsive-exploration-type fun. However, a few nonsense words were created in the process of pretend play in order to imitate making real words. Among actual real words constructed in the open-ended mode, a majority were names. There was also a small number of words resulting from self-challenge, and a small number of rare personally meaningful words that were unknown to the system (e.g. ROBLOX). More details on these categories of real words are provided in section 6.2.1. Finally, we saw a few examples of out-of-vocabulary invented spelling.

Most nonsense words constructed in SpeechBlocks II were random or patterned letter combinations. The reader might remember an additional type of nonsense words enjoyed by children in SpeechBlocks I: remixes of real words, such as CUPEAR and ZOOBALLBALL. These words played a prominent role during children's initial experience with SpeechBlocks I. Unfortunately, this type of playful experience was lost in SpeechBlocks II, as a tradeoff of switching to the Word Box word building mechanics. Consequently, although children were still delighted in making nonsense words, this process didn't cause nearly as much laughter, excitement, or social interaction as with SpeechBlocks I. Future designs may need to consider how the advantages of both approaches can be combined.

Most nonsense letter sequences were apparently made as an end in itself, which is evidenced by the behavior of the players. For instance, when Arnold (an impulsive explorer) showed a researcher one of the nonsense sequences, the researcher asked: "What are you trying to spell?" — and he responded "I don't know." Despite their simplicity, these sequences were still able to evoke the senses of agency and self-efficacy. Children showed them to researchers, saying "Look what I made!" Edward (an impulsive explorer) once said: "Look what I made: a mess!" — and appeared proud of that.

Some nonsense words were constructed in the process of pretend play, imitating building real words. One child put together AB and said: "These letters are in my name." Her name indeed started with A, but didn't contain B. She continued by putting together ABCDEF, and then said: "Now I need to clear my name" — indicating again that she pretended that the string was her name. Another child started ORANGE with scaffolding, but exited the scaffolded mode and started to assemble the string LDWOXYI. When asked what she was making, she responded: "I'm making orange!" Such pretend play resembles children's scribbles: children similarly pretend that scribbles hold a certain meaning (Strickland & Morrow, 1989).

6.5.4. Invented Spelling Interpretation

Invented spelling interpretation enjoyed a very limited success. It accounted for only 6% of the words built. Moreover, most of these words were either created by accident, or by imitation of demo examples, or with a significant amount of adult's help. Only one child, Ananda, consistently and independently used the invented spelling interpreter. Limited usage of this scaffolding mode is caused by children in our sample having a challenge identifying sounds in words and locating corresponding blocks on the full keyboard. These difficulties are exacerbated by children's tendencies not to undo their construction steps and to confuse the order of letters in the word, both of which mislead the interpreter. These problems might be less pronounced for older children, for whom invented spelling interpretation might still be of value. Furthermore, observations suggest that the interpreter can be improved by taking into account contextual information.

For most children, their experience with the interpreter was limited to serendipitous finds resulting from spelling nonsense words. This is how AU became ASPARAGUS, AVM became AQUARIUM, KMD become CAMBODIA[29], BP became BAGPIPES, BMD became BADMINTON, and BDU became BODYGUARD. Given the unusualness of these words, their complexity, the fact that they were not associated with anything on the canvases, and that children have expressed no verbal intent to make them, it is very unlikely that children were purposefully trying to build these words. This is further illustrated by the case of a child who built MOZAMBIQUE but exclaimed: "I've got it! It's a Rozenback!" Although children often expressed delight at the results and sometimes incorporated them in their word collections or scenes, the value of such use of the interpreter appears to be minimal, since it is akin to random word suggestions.

[29] The list of countries was included in SpeechBlocks vocabulary because the author thought that children from migrant families might attempt to spell their home countries. In the actual study, that never happened.

Purposeful usages of invented spelling interpretation often were limited to copying the demo examples or involved heavy adult guidance. This seems to be to a large extent caused by limited phonological awareness of most participants. E.g. when we asked one child what sound the word BATMAN starts with, he responded "It starts with Batman!" Several weeks later, he correctly identified the first phoneme in BOAT, but in response to what *other* sounds can he hear, he kept saying "*[b]! [b]!*" Many other children responded to such questions with a mixture of correct phonemes and random guesses. As a result, as the invented spelling procedure involves identifying at least two sounds in the word, most children weren't able to use it without an adults' help. Such help consisted of (a) sounding out the word slowly and clearly, so that each individual sound can be heard, (b) rejecting wrong guesses and confirming the good ones, to prevent children from going astray, (c) helping children to find the desired blocks on the keyboard, and (d) making sure that they arranged blocks in the correct order. In a way, adults acted as a more sophisticated and intelligent version of the direct guidance system coupled with the open-ended keyboard. Even though the above-mentioned practice required one-on-one adult supervision, there still likely is value in it, associated with practicing invented spelling. Nevertheless, the original goal of the scaffolding system — supporting autonomous word creation — hasn't been met in this case.

The above-mentioned difficulties were exacerbated by a tendency of children not to undo the steps of their construction process. They were never observed removing a block that they were not sure of from the box. As a result, misleading blocks accumulated in the word box and led the interpreter away from the desired word. For instance, Ananda was building BOMB for one of her scenes, and was not sure about the medial vowel. Instead of skipping it, she put together BA. When she didn't see the desired result, she tried another representation of the same vowel, without removing A, and got BAO. Still not seeing BOMB in the list of suggested words, Ananda continued looking for the right vowel. In order to get the system back on track, a researcher's intervention (suggesting to remove A) was necessary.

Other misleading cases for the interpreter occured when children put in correct blocks, but in a wrong order. The confusion about the order of blocks was consistent with the phenomenon of spelling words backwards that appeared in SpeechBlocks I play for children of this age.

Although the main issue with the invented spelling system seemed to stem from children's limited literacy knowledge, there were some issues related to operating the interface as well. They stemmed from difficulty to present multiple guesses of the interpreter in the limited screen space. I tried several ways to address this challenge, but have not arrived at a satisfactory solution.

The only child who did eventually derive a significant value from the invented spelling interpreter was Ananda. She learned to operate this system autonomously (or nearly autonomously) and used it to spell a variety of words for her scenes, displayed on Fig. 6.44. She used it to build STAR, MOON, FIRE, and SATURN for scene (1), CHAIR, TABLE, CAKE, and GIRL for scene (2), BOMB for scene (3), BUS, ROAD, and TREES for scene (4), and SWORD for scene (5). As it was mentioned earlier, Ananda combined strong literacy skills, enjoyment in exercising them, a good

executive function, and the need to build words for the sake of imaginative play (the type of play that among all children she was the most active at). She was an unusual child in our sample, but it is possible that as children grow and their literacy skills develop, more of them will eventually reach the same stage as Ananda. In other words, it is possible that our sample was simply too young for the invented spelling interpreter to function properly. Evaluating it with older children remains a subject for future studies.

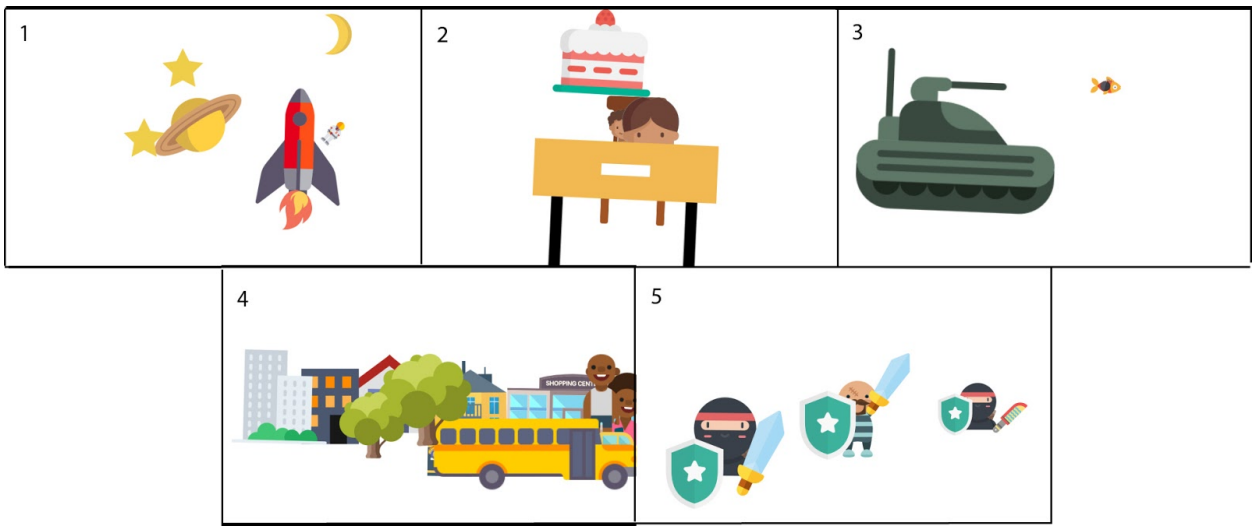


Fig. 6.44. Scenes built by Ananda with the help of invented spelling interpreter

The performance of the system can also be improved by taking into account contextual information. For instance, during the construction of a space scene (Fig. 6.44, (1)) Ananda wanted to place SATURN among the stars. She said “I’m going to make Saturn”, and was able to independently identify the initial sound *[s]*. However, she was not sure about other phonemes and their locations on the keyboard. She turned to a researcher (the author) for help. I assisted by carefully sounding the word out. She first was able to make SR. She spent some time scrolling the list of guesses in search for the target word, but was not able to locate it, and had to resort to my help again. Only after she made SRN was she able to find SATURN. However, by the time she started building the word, she already had STAR, ROCKET, and ASTRONAUT on her scene. Moreover, she said the word out loud. If the interpreter used the content of the scene and the background conversation (passed through speech recognition) as contextual cues, it would be able to figure out the child’s intent from shorter inputs, reducing demands on the child's literacy knowledge.

As an interesting side note, some invented spelling phenomena described in the literature showed up in the study. Children tended to skip medial vowels to some extent. Fascinating examples of using letter name for letter sound were witnessed, such as ALLE for ALLIE (the name of an observer; here E was used to denote *[i]*), HH for CHURCH, and HR for CHAIR (here H - *[eɪt͡ʃ]* - was used to denote *[t͡ʃ]*).

6.5.5. Word Bank

Word Bank enjoyed great popularity among children at the time of its introduction. Despite subsequent appearance of more sophisticated word sources, it remained relevant through the end of the study. Its value arises from the combination of two properties: (a) providing children with sufficient choice of words to roughly match their interests and play intentions, and (b) being easy to use and free of technical issues. The latter property allowed children to fall back on it when more sophisticated word sources didn't work for them. Children also used Word Bank to browse for ideas.

Contrary to some original concerns, the two-level organization of the word bank (categories - words) was well understood by children. At the same time, this structure allowed us to "pack" a sufficient number of diverse words into the bank to cater to children's interests. Two aspects of Word Bank content selection were also helpful in that regard. First, such categories as "Names", "Cartoon Characters", "Food", "Animals" and "Family" turned out to be broadly appealing to children. Second, content that compliments each other (e.g. items in the Food category going along with furniture and dishes in the Home category, going along with people in the Family category) allowed children to build scenes entirely or primarily from the Word Bank (Fig. 6.45). Ease of navigation arising from the two-level organization led to another observed behavior: children browsed the Word Bank in search of ideas. They clicked through categories until they stumbled upon a word that interested them and built it.

Fig. 6.45. Scenes built entirely (1,2) or primarily (3) with word bank

6.5.6. Associations

The network of semantic associations was the most popular word source (closely followed by speech recognition). Similarly to the Word Bank, it was reliable and easy to use. Because of its connectedness to the content of the canvas, it was particularly useful for imaginative play. It served not only as a convenient word source, but also as a source of ideas. Some children browsed the association network for prolonged periods of time looking for suitable objects to develop their scenes. Such interactions can be thought of as a simple form of human-computer co-creativity, and suggest the potential of incorporating co-creative aspects into expressive media.

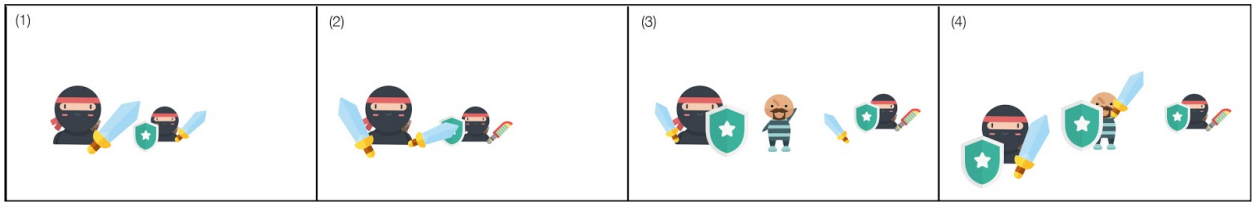


Fig. 6.46. Ninjas Scene Before and After Addition of DAGGER and PRISONER

A good example of using the association network as a source of ideas can be found in the construction of a sophisticated scene by Ananda (Fig. E1). At first, she created two ninjas, a big one and a small one, and said: “They are father and son. They are practicing.” She equipped them with swords and shields. Seeking to further elaborate the scene, she then tapped one of the swords to get the associated words and saw DAGGER. That gave her an idea to arm the small ninja with a dagger; to do so, she transferred his sword to the father. After completing this modification, she tapped on the sword again and began a long journey through the semantic network in search of something interesting to add: SWORD → WARRIOR → HERO → BATMAN → DRAGON → UNICORN → CENTAUR → GOBLIN → WARRIOR → HERO → SOLDIER → PRISONER. Seeing PRISONER, she got an idea of what could enrich her scene, and exclaimed: “I’m going to make a villain to fight them!” Thus, a lucky find in the association network made her develop her story into a new direction, switching from the original scene (a father and son practicing) to a more dramatic setup (the ninjas fighting a villain).

In some cases, children’s imaginative play unfolded in this fashion: a seed sprite was brought to the canvas, and then sprouted into a scene via associations. The initial, seed sprite could well be random. For instance, Archie (an imaginative player and an impulsive explorer) serendipitously created FLOUNDER by tinkering with invented spelling. He then used associations to create OCTOPUS, TURTLE, and SHARK and expanded his creation into an underwater scene (Fig. 6.47).

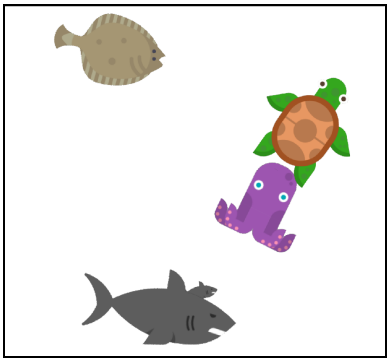


Fig. 6.47. Archie’s Flounder Scene

Such interactions make the present work relate to studies of human-computer co-creativity (e.g. Davis et. al., 2014; Davis, 2017; Ali, 2019). Similarly to the observations above, co-creative systems described in literature had value for human users in facilitating divergent thinking and getting out of creative blocks. This similarity raises a question whether the approaches of that field

can be applied to expressive literacy media. I should note that the idea suggestion offered by the association network was very gentle, giving the child a lot of freedom in interpreting the ideas and incorporating them into his/her work. It might be desirable to preserve these qualities in future co-creative systems for literacy learning.

6.5.7. Text Recognition

Text recognition attracted a good amount of children's interest and prompted them to explore texts around the classroom and in books. However, several factors also made it a significant source of frustration, confusion, and distraction. First, limitations of available technology made it difficult to use. Second, children eagerly spelled random and incorrectly recognized words, creating the potential for confusion about the meaning of the texts around them. Third, it invited several unintended uses, such as "taking pictures" of other children and looking around through the screen, which were distracting from the main activity. Fourth, it might not be well aligned with the main expressive activity, making words. Taking both its advantages and disadvantages into account, it is debatable whether text recognition is a worthy addition to an expressive learning medium. A more natural word source for children of that age might be object recognition.

The tablet's ability to read text impressed many children. "That was amazing!" — said Archie (an impulsive explorer) several times during the text recognition demo. "That was so awesome and cool!" — said another child. But unfortunately, that technological impressiveness didn't translate very well into practicality. Despite the significant effort invested in making text recognition a useful word source, its usage was still plagued by a multitude of technical issues. These issues were amplified by children's insufficient motor skills, limited understanding of how both technology and texts function, and impatience. A likely incomplete collection of problems is below:

- Children held the camera sideways relative to the text. Real-time text recognition libraries were not capable of processing text appearing at a large angle. To partially mitigate this issue, thin horizontal lines were added to the text recognition screen, and children were asked to align the text with these lines;

- Children held the camera too close to the text, making the device unable to focus. In extreme cases, children put the camera right on top of the page they were trying to scan, believing that it would be able to see better;

- Children sometimes were not able to align the camera with the words that they were interested in;

- Children shook the tablet because of their insufficiently firm grip on it. As a result, the camera was not able to focus and pick up words;

- The system had degraded performance on decorative letters and unusual fonts. Unfortunately, materials designed for children often contain exactly these fonts, because their designers try to make them look informal, warm, and friendly[30], sometimes at the expense of legibility (Fig. 6.48, a). In some cases, such fonts even end up being applied to learning materials (Fig. 6.49, b)[31]. Children's books present a different challenge: illuminated capital letters (Fig. 6.50, c). Such capitals were typically recognized separately from the word following them, and as a result the word got truncated;

Fig. 6.48. Examples of unusual and decorative fonts used in children's materials: (a) samples from designers' blogs, (b) posters similar to ones found in the study classrooms, and (c) illuminated capital letters in children books

- Children were interested in messages written by teachers on the board, but unfortunately these messages often had smudged, underlined, highlighted, and crossed out handwritten letters. Each of these types caused issues for the recognition engine;

- Using the word picking interface required following a sequence of three-four steps. The child needed to point the camera at the word, freeze the image with the "freeze" button, tap the desired word to hear it, and tap the "write" button to build it. As it was noted in section 6.2.3, long coordinated sequences of actions are likely to cause children with low EF to abandon them mid-way.

A set of children, including Ananda, Jonathan, Randolph, Ulisses, Erickson, Arnold, and Edward, were able to overcome these technical obstacles. As the reader might remember, the first five children generally gravitated towards self-structured play, while Arnold and Edward were impulsive explorers who nevertheless started to exhibit a lot of focused behaviors towards the end of the

[30] Examples of some designers' reasoning about child fonts, as well as font samples, are drawn from the following blogs: https://creativemarket.com/blog/fonts-childrens-books, https://bashooka.com/inspiration/45-finest-decorative-fonts/, https://medialoot.com/blog/fun-and-playful-kids-fonts/
Other designers do treat font legibility and ease of imitation by children as paramount priorities, e.g: http://marie-story.com/the-best-fonts-for-childrens-books/.
[31] Example posters are from companies *Creative Teaching* and *Business Basics*.

study. These children held the camera steadily and upright, at the right distance from the words, selected the words purposefully, etc. For other children, however, each of the technical issues contributed to the probability of them giving up attempts to use text recognition for its intended purpose and instead coming up with alternative ways to have fun with it.

When children did use text recognition to spell words, they were at times surprisingly indiscriminate in their choice of which words to spell. They picked seemingly arbitrary words, such as AN, THE, HE, UP, AND, IS, DO, out of context. Each time it happened, they were pointing the camera at a page full of more informative words, which they didn't explore. They also eagerly spelled products of text recognition errors, such as CALE (for CUTE), DRAGO (for DRAGON), ANNAS and EARTHS (for ANNA and EARTH), RZONT, FADER, LOORM, HEELS (for WHEELS), SOO, ODA. In one case, the child (Ericson) even knew that the system didn't pick up the word correctly, but proceeded to spell it regardless: "I'm going to do this, except it doesn't get it right." The tendency to spell seemingly random words is puzzling. My interpretation is that children wanted to use the feature, but struggled to make it work in a meaningful way, and instead chose to use it at least in *some* way. If this is correct, the usefulness of text recognition in such cases is questionable. A factor that might have contributed to this behavior was that children typically didn't know what the text they were scanning was supposed to say, so they likely just relied on the reading of the system. Such reliance might have created confusion regarding the meanings of the texts around them and how texts work (especially if the system's reading changed time after time).

Although text recognition is child-driven in a sense that it gives the child initiative on what texts to explore, it is not entirely child-driven in a sense that its results depend on the source as much as on the child. This becomes particularly true in light of how difficult and cumbersome it was for children to purposefully search for words. This difficulty was exacerbated by the fact that most of the children couldn't read. Therefore, they had only a rough idea, or no idea at all, regarding the content of the texts that surrounded them. How limiting it was depended on the type of play that children engaged in. For word crafters, who seemed to be interested in gathering exciting words without much of a system, this aspect of text recognition wasn't an obstacle. They explored the classroom in search of words (environmental texts were available to them can be found below) and eagerly spelled the words they discovered. In the case of imaginative play, it is important for the player to make words connected to the content of the scene s/he is building. Therefore, text recognition was less conducive for this purpose, although still usable on some occasions. First, in a small number of cases, words were serendipitously discovered via the feature and then used in imaginative play. Fig. 6.49, (a) shows one of the scenes created in this manner. The child was walking around the classroom, exploring what was written on different labels and boxes, when he came across a game named "Ants in Pants." Delighted about the name, he built both words and put them on the canvas. Next, he came across a construction material labeled as "Waffle Builder." He built WAFFLE and added it as a trophy for the ant to carry.

The second way text recognition could be utilized for imaginative play was via texts that provided children with good cues on what they were about. One type of such text were the educational books for beginner readers — they typically contained large illustrations corresponding to the

subject of each page, surrounded by a small amount of text on the subject. Fig. 6.49, (b) shows a scene sourced from National Geographic's children book *Planets*. Another type of books that were very convenient in that regard were Richard Scarry's books. They consisted of scenes arranged out of a great variety of interrelated small objects, with a label written right next to each object. The images helped children to navigate towards the words they wanted. Scene on Fig. 6.49, (c) was sourced from such books. Finally, "words on the week" written on the classroom board ended up being another rich word source. Children were aware of what the theme of the week was, and could therefore source words in a purposeful way. Fig. A2, (d) shows the insect scene built by Ananda, in which the words SPIDER, BEE, and BUG originated from the "words on the week."

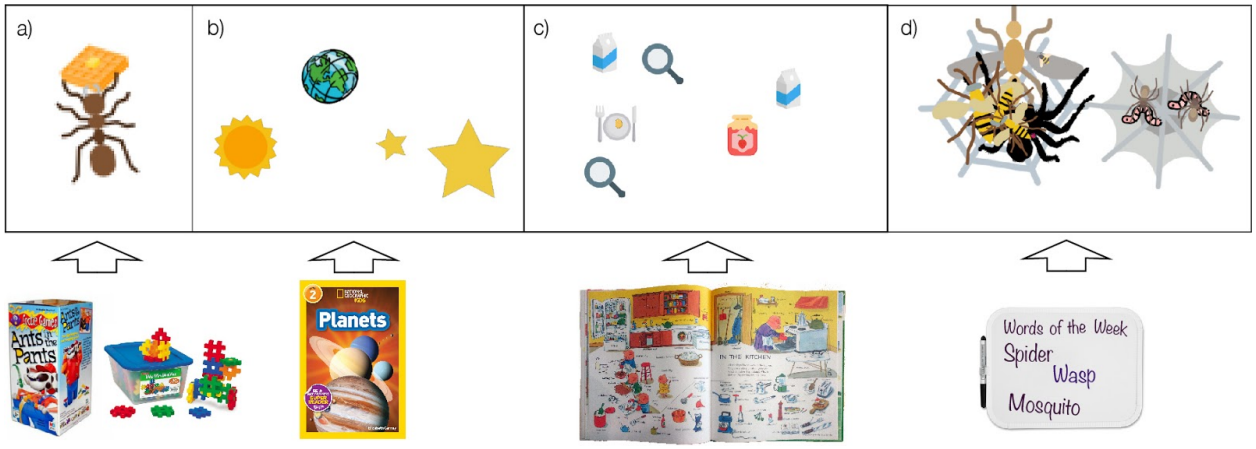


Fig. 6.49. Usage of text recognition for imaginative play

With purposeful use of text recognition being difficult, a variety of unintended usages emerged. Section 6.2.3 describes a few of them: using the "freeze" button in order to "take pictures" of other children and themselves, looking around through the screen (as if through camera obscura), and running around the classroom detecting bits of text, without any attempt to read or build them.

A more literacy-oriented use of text recognition was exploring the texts scattered around the classroom by tapping on them on the text recognition screen (still without attempting to spell them). Children could spend entire sessions exploring the classroom in this manner. Several locations attracted interest from such technology-assisted readers. They explored labels on the cubicles where various classroom materials were stored. Of great interest to them was the board where the greeting message, the "message of the day", and the "words of the week" were written. Children already had some familiarity with them through their classroom routine, so finding these words evoked delight of their confirmed expectations. The board also housed various posters that displayed, among other things, the names of the days of the week and months, weather types, animals, and plants. One child, Edward, used the text recognition to carefully "read" an entire sentence from such a poster, "Fresh air flowers and plants", by tapping on words in order. Unfortunately, unbeknownst to him, the text recognition didn't get the words "air" and "plants" correctly, so he arrived at a wrong understanding of the sentence. More posters were scattered throughout the classrooms. For instance, in Edward's classroom, there was a behavior chart to

which clothespins with children's names were clipped to mark how they were behaving today. Edward scanned the top entry on the chart and heard “Super Student.” He then kneeled to see the bottom one and predicted: “This one is going to say ‘Bad Job’.” He listened to the text recognition and repeated what the text actually said: “Stop! Think!” The behavior chart was not the only location displaying children’s names. For instance, there was a chart called “class jobs” that indicated responsibilities of different children, such as watering plants. Next to the entrance, names adorned the cubicles where children stored their belongings. As usual, names were of utmost interest to children. When Edward found his own, he ran to a teacher, offered his headphones to her and exclaimed: “Mrs. K., hear what it says!” Finally, children tried to scan writings on each other’s T-shirts and researchers’ badges.

Along with environmental text, books were also explored via text recognition. With some help from the researchers, children “read” their titles by tapping on words in order. Children didn’t attempt to read the body of a book in this manner, but they did point the camera at the pages in order to see which pictures would show up. One child combined this activity with pretend reading[32] — making up a story about a tiger to go with the book’s illustrations. When a researcher suggested that she tapped on the words that showed up, she was surprised and delighted that the system said “tiger.”

Such exploration of texts might have certain value on its own. Children were engaged in it, it was accompanied by a positive emotional background, and a sense of self-efficacy. These factors may help in the formation of a positive attitude towards texts, as well as a better understanding of their role and functions. However, such usage of text recognition was disconnected from the core function of SpeechBlocks, that of expressive medium, and was actually distracting children from it. Therefore, a designer who is interested in these capabilities might consider whether they are better suited for a standalone, reading-oriented app.

On multiple occasions, children tried to scan images in books instead of words. For instance, one girl was trying to pick RUBBLE, a dog character from the animated series *Paw Patrol*, from a book page showing his image, but no related text. Such observations raise a question whether object recognition might be a more natural word source for children to use than text recognition. Since the language of images is much more familiar to children of that age than the language of written word, object recognition could help to mitigate the difficulty of purposeful word selection that was observed in connection to text recognition.

6.5.8. Speech Recognition

Despite initial concerns about poor speech recognition accuracy on children's voices, this word source was well-liked by children and successfully used by them both to carry out their creative ideas and to imitate works of their peers. In practice, speech recognition had by far less technical

[32] The fascinating phenomenon of pretend reading is widely documented in the research literature - e.g. (Bissex, 1980; Rhyner, 2009).

issues than text recognition. Children themselves greatly contributed to its success by being remarkably patient and eager to repeatedly try over and over again when technology didn't work as intended. An interesting detail was that their interaction with the system was conversational: they spoke in sentences and addressed the avatar of the system, a big-eared fennec fox, as a person. Results of speech recognition usage are very encouraging, since they open a door for truly open-ended and child-driven scaffolded play. Parasocial interaction with the system is interesting in its own right, since it points to a potential of using virtual agents in expressive learning media.

Children used speech recognition for several purposes. Their earliest interactions with the technology were often simply probing its capabilities. For instance, they requested some words and waited for them to show up, but instead of spelling them, made new requests. As they became more familiar with speech recognition, they started to request words that were of personal interest to them and words that they needed to develop their scenes. Another very important use of speech recognition was quickly borrowing ideas from peers. Many cases of idea borrowing described earlier were facilitated by this technology.

Children exhibited remarkable patience when speech recognition was unable to pick up words they said. If it happened, they repeated their requests again and again. I observed up to six such repeated attempts in a row. Only in rare cases did children express frustration: e.g. "I said nothing of what is here!" Such patience and persistence hints at the value that children derived from the system.

In most cases, using the speech recognition interface wasn't limited to simply pressing the button and requesting the word of interest, as most adults would do. Instead, children conversed with the system: "I want DINOSAUR, I want all the DINOSAURs!", or "Please give me BATMAN and SPIDERMAN and..." Sometimes they deliberated what they would like to have for prolonged periods of time, while holding the recognition button: "I want uh... I want ummmmmmm..."; "I want CANDY! I mean, I need a POLICEMAN!" Occasionally they requested objects of a particular type or even entire scenes: "PURPLE RUG", "RED OCTOPUS", "Can you make a ROOM with TOP?", "I want a HOUSE with TOYS." Handling requests involving adjectives was not supported, but is technically feasible, and might enrich the play.

It is also interesting that children related to the avatar of the speech-recognition system, a big-eared fennec fox, as a person. Although parasocial interaction wasn't originally intended (the avatar was introduced for different reasons, which are described in section 3.3), it appears to be quite beneficial for children's engagement and mitigation of their potential frustration with the system. A name for the character, Mr. Fox, quickly emerged independently in both groups. Children often requested words from Mr. Fox in a very polite manner: "Mr. Fox, can you give me SPIDERMAN?", "Mr. Fox, can I have STRAWBERRY?" In one case of speech recognition failing to deliver the correct result, the child (Archie), instead of getting frustrated, tried to encourage the character: "Oh no! Mr. Fox, we need you! Mr. Fox, spell FIRE!" When he did finally get the result he wanted, he exclaimed happily: "Mr. Fox, you busted it out!" (apparently in the sense "produced"). He then said: "Fox, I love you." Children issued other requests to Mr. Fox, presuming that the

character could control the tablet: “Mr. Fox, what time it is?”, “Mr. Fox, finish off (turn off) the tablet!”, “Mr. Fox, erase everything! Erase-erase-erase!” They talked to the fox about his state: “Are you awake? Go to sleep!”, “Wake up!”, “Mr. Fox, confuse! I said confuse!” (referring to his resting state and confused state, Fig. 6.50). One child also asked about Mr. Fox’s abilities: “Can he fly?” — likely by analogy with the cartoon character Dumbo. The fact that children related so much to such a simple virtual agent as Mr. Fox suggests a great potential for use of relational AI (J. Kory-Westlund, 2019) in applications such as SpeechBlocks. Such AI can be a face of the scaffolding system, a helpful and responsive guide.

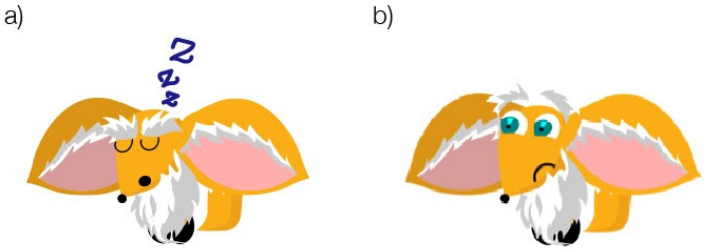


Fig. 6.50. Mr. Fox’s sleeping and “confused” (recognition failure) states

6.5.9. Synergistic Usage of Multiple Word Sources

Multiple word sources introduce complexity into the user interface of the medium. Is it desirable to have several of them, or is it preferable to choose the one that works best? An argument in favor of multiple word sources is that they fulfill different roles (response to specific ideas, idea generation, fallback mode). Furthermore, because of their complementary functions, they can be used synergistically. Indeed, children used them in this way, and that helped the creation of some of the most sophisticated scenes.

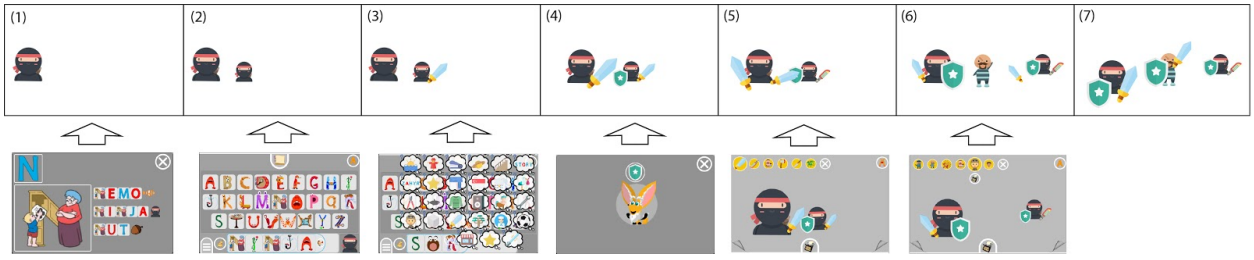


Fig. 6.51. Synergistic usage of multiple word sources

An example of such synergy is the construction of a ninja scene by Ananda. The origins of the scene likely lie in the play of Ananda’s peers, two of whom made ninjas during earlier sessions and were playing with ninja sprites on the same day. Ananda started by going to the letter page for N and picking NINJA from sample words (Fig. 6.51, (1)). Once she built NINJA with scaffolding, she challenged herself to repeat the construction in the open-ended mode (Fig. 6.51, (2)). Two ninjas were placed on the canvas and arranged into different sizes, according to Ananda’s intent to make them a father and a son. The girl said: “They are practicing!”, and to give them something to

practice with, she said: “I’m going to make a sword!” She spelled SOR independently and got SWORD among the guesses of the invented spelling interpreter (Fig. 6.51, (3)). She then decided to give the ninjas shields. She tried to use invented spelling again with the letter keyboard, but since there is no letter corresponding to the sound *[ʃ]*, she couldn’t find how to start the word. Instead, she switched to speech recognition (Fig. 6.51, (4)). With the addition of both swords and shields, the scene acquired some completeness, and Ananda switched to seeking ways of elaborating it. She used the associations to give the son a dagger (Fig. 6.51, (5)). Then a long journey through the association network brought a new theme to her work: she introduced PRISONER as a villain (Fig. 6.51, (6)). She completed the scene by building SWORD and SHIELD again, giving them to the villain and arranging the trio in a dynamic fashion on the page.

6.6. Letters vs. Onomatopoeic Mnemonics

The previous section dealt with the software routines that scaffolded children’s word building. This section assesses a less direct type of scaffolding — supplying children with blocks that are intended to make the process of word construction more straightforward. To remind the reader, an element of SpeechBlocks II design was the inclusion of phoneme blocks. It aimed at allowing children to construct words directly out of phonemes, without having to deal with orthographic complications. To visually represent phonemes, I used the onomatopoeic principle: I developed animated characters, each of which produced the sound of a particular phoneme via some action. These characters are referred to as “sound creatures”. A decision was made to integrate letters into the designs of these characters, so that children could use letters as additional cues. However, because the same phoneme can be expressed by different graphemes in different contexts, the creatures assumed several forms corresponding to different graphemes. The phoneme blocks behaved inversely to the letter blocks: for the latter, the spelling stayed the same, but the sound changed based on the context; for the former, the sound stayed the same, but the spelling changed. Although “sound creatures” were designed for phoneme blocks, they can be used with letter blocks as well.

Experience of using SpeechBlocks II in K-1 classrooms shows that the difference between the letter and phoneme blocks was much less important than I originally thought due to an overwhelming majority of the words being constructed in the direct guidance mode (see section 6.5). In this mode, the target context for each block is known, and both pronunciation and spelling of each block remain as prescribed by the target context, rendering the difference between the two block types irrelevant. Construction of real words in the open-ended mode happened too infrequently to reliably compare the two block types. There wasn’t, however, even anecdotal evidence in favor of the phoneme blocks.

However, the onomatopoeic creatures did show their value for some children. Evidence of this value was gathered from three sources: (1) observations of children’s interactions with the creatures during the study, (2) post-study interviews in which children answered questions about

one or two of the creatures, and (3) a sound-finding mini-game, administered along with the interviews, in which children were asked to find blocks corresponding to sounds on “letters” and “creatures” keyboards. Observations show that many of the players responded to the creatures with significant interest and positive affect, repeated the sounds and the actions of the creatures and sometimes reacted to them in dramatic ways. However, it appears that the creatures were occasionally perceived in a way that obscured their association with their sound, which might be a limitation of the onomatopoeic approach itself. Some observations from the interviews also question how salient the creatures were for most children. In the post-study interview, most children showed an understanding of the principles behind the creatures: that a creature is associated with a particular sound, that the same creature can have multiple forms, etc. Bayesian analysis of the sound-finding mini-game results suggests that a sizable fraction of children can find blocks quicker and more accurately in the creatures mode. However, there was also a sizable fraction of children for whom the opposite pattern held. It is currently unclear which factors placed children in each group.

6.6.1. On Letter vs. Phoneme Blocks

Construction of meaningful words in the open-ended mode happened rarely, so the available evidence regarding advantages and disadvantages of letter and phoneme blocks is anecdotal. We saw that most attempts to build words in the open-ended mode involved words that children were very familiar with and knew them letter-by-letter. In such cases, automatic changes of spelling in the process of word construction were likely confusing. One such case was observed with Ericson, who was spelling his name and appeared slightly confused when one of the blocks changed. The facilitator suggested that he should switch to the letter mode. Once he did that, he was able to complete the word much faster.

The phoneme keyboard might also have been more difficult for children to navigate. With the letter keyboard, we saw that children used such strategies as the alphabet song to locate blocks. These strategies didn’t apply to the phoneme keyboard. Despite my attempt to arrange the keys on it in a perceptually motivated way (as described in section 3.3), I have observed several cases of children struggling to find blocks on it even though they knew which sound they were looking for. However, with the letter keyboard, children occasionally got stuck trying to find blocks for sounds that were not normally represented by a single letter. For instance Ananda struggled to find the *[ʃ]* sound on the letter keyboard during construction of the word SHIELD. The letter keyboard might benefit from including several letter combinations (such as TH, CH and SH) that represent sounds which do not have a corresponding “default” letter. With these three combinations, all consonant phonemes become represented.

The combination of such modified keyboard and invented spelling interpretation might be a sufficient solution to address the issue of orthographic complexity that motivated the phoneme blocks. With these systems in place, children can use grapheme blocks to represent both letters and sounds. This ambiguity of usage would be resolved by the invented spelling interpreter.

6.6.2. Children's Engagement with the Onomatopoeic Creatures

Creatures were well received by children upon their introduction at the beginning of the study. Kids spent significant amounts of time watching their animations on the keyboard and visiting the associated creature pages. About a quarter of the children exhibited a distinct positive affect, e.g. laughter and excited remarks, such as "This is hilarious!" and "Mamma mia!"

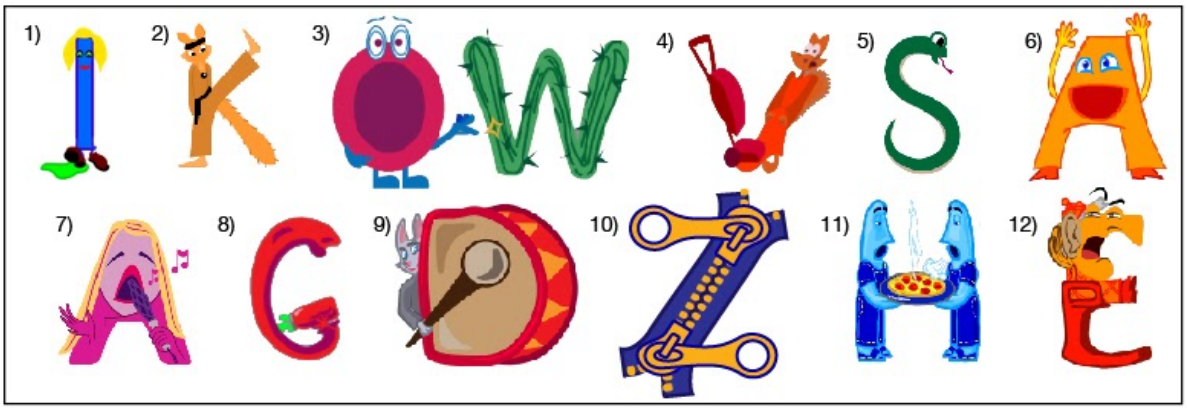


Fig. 6.52. Some creatures that evoked children's reactions

More than half of the children were seen reacting to the animations of the onomatopoeic creatures in various ways. Some of these reactions were related to the sounds of the creatures. For instance, children enjoyed repeating the high-pitched squeak after the *[i]* character (who squeaks after stepping into something icky; Fig. 6.52, (1)). Several children played the animation and repeated the sound many times in a row. It even evoked an exchange between two children: they played the animation simultaneously and squeaked at each other. Similarly, children repeated the sounds after the scaffolding system when they were constructing words. Consonant sounds were sometimes coupled with a vowel: e.g. "duh" or "bah".

In addition to mimicking sounds, children repeated the descriptions of creatures' actions after the system, or came up with their own interpretations of it. For example, in response to the creature for *[k]* (whose name is Katie and who does karate kicks; Fig. 6.52, (2)), they said: "Katie does karate kicks" (exact repetition of the system's prompt), "Katie-karate", "He does karate". In response to the creature for *[aʊ]* (who cries: "Ow!" after touching a cactus; Fig. 6.52, (3)) a child said: "Ouch! His finger is bleeding!" In response to the creature for *[v]* (who got caught in a *vvvvv*-humming vacuum, Fig. 6.52, (4)), a child said: "Nice! He is stuck!" Reactions to creatures' actions were at times dramatic — for instance, children jumped in their seats and pretended to be frightened upon seeing the *[s]*-snake (Fig. 6.52, (5)), or waved back to *[eɪ]* creature ("Abe waves to a friend", Fig. 6.52, (6)). It is possible that interest in creatures and their actions can help children to memorize the link between the creature and its sound.

However, it seems that the creatures were occasionally perceived in ways that obscured this link. One issue was that an action rarely (if ever) had unambiguous onomatopoeic representation. Although the target phoneme was played along with the animation, some children preferred to ignore it and instead came up with their own sound. For instance, children responded to the karate-kicking animation of the *[k]* creature (Fig. 6.52, (2)) by exclaiming "Hiya!" and "Pff! Pff!", and to the singing animation of the *[ɑ]* creature (Fig. 6.52, (7)) by singing "La la la!" Those are, of course, quite natural ways to voice the respective animations: "Hiya!" is a sound often made by karate fighters in movies, and "La la la!" is a common way to portray someone singing. The lack of a "gold standard" action-to-sound association might be an inherent limitation of the onomatopoeic approach.

Furthermore, sometimes the animations themselves were interpreted in unintended ways. For instance, the *[k]* creature's (Fig. 6.52, (2)) action was sometimes interpreted as "He is hitting in the tummy," and "He is kicking his tail." For the *[g]* creature (Greg gulps grape juice, Fig. 6.52, (8)), one interpretation was "He is drinking soda." For the *[i]* creature (Fig. 6.52, (1)), one comment was: "Ew! She stepped into something!" (which evokes another sound: *[jʊ]*).

Finally, children sometimes responded to the creature's appearance or superfluous details of its behavior. For instance, in response to *[i]* animation (Fig. 6.52, (1)), different children said: "Such pointy hair!" and "Pops her hair!" A child said "A rabbit" upon seeing the creature for *[d]* (Fig. 6.52, (9)). "That looks like a monster!", exclaimed a child upon seeing the zipper animation for *[z]* (Fig. 6.52, (10)). In a few cases, children focused on these superficial details, rather than on the target sound, while trying to locate a needed block. For instance, several children were observed trying to use the *[h]* creature ("Henry and Harry are blowing on hot food."; Fig. 6.52, (11)), who had a tray of pizza incorporated in its design, in the process of constructing the word PIZZA. In another example, a child was bewildered how the name of a Disney princess, Elsa, could contain the sound *[ɛ]* ("Eddy struggles to hear: Eh?"; Fig. 6.52, (12)): "It is so ugly; how can it be in her name?"

6.6.3. Children's Understanding of the Creatures

In order for the creatures to work as intended, we expected children (1) to be able to interpret the creature's action and relate it to the creature's sound, (2) to understand that the same creature may appear in various forms, (3) to understand that all of these forms produce the same sound, and (4) to see the letter shapes associated with the creatures. In addition, we hoped that children might form parasocial relationships with the creatures. To examine these assumptions, we had a short interview with each participant at the end of the study, in which we asked several questions regarding sound creatures. We chose one creature as our primary focus: Katie the Karate Kicker, representing sound *[k]*. This choice was motivated by the creature having multiple forms (K, C, and Q) and having a clear and memorable animation that elicited responses from multiple children during the course of the main study. We showed each child the page dedicated to the creature, and asked the following questions: (1) What do you see?, (2) (if children didn't refer to the creature in their previous answer) Who is that?, (3) What is her name?, (4) (pointing to different forms of the

same creature) Is this Katie? And this? Can you point to all the Katies?, (5) (again, pointing to different Katies) What does this Katie say? And this?, (6) What do you think each Katie looks like?, (7) Why do you think she looks like she does?, (8) Why do they look different?, and (9) What is she doing? In a few cases, we also asked the same questions regarding a different character, representing the sound *[s]*. However, we had to abstain from systematically investigating several characters in the interest of children's time. Thus, I must note that there is some risk of the present findings not generalizing to other creatures.

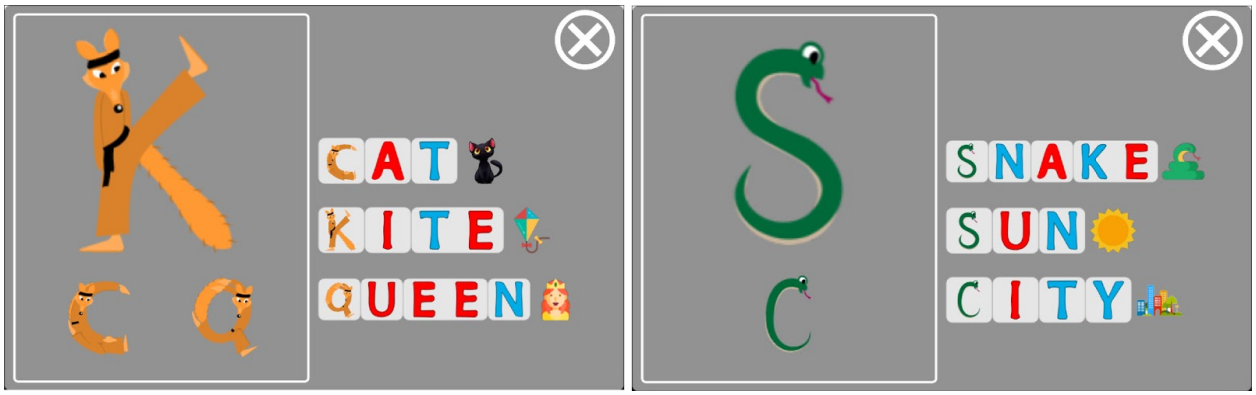


Fig. 6.53. Pages of the Creatures Used in the Interview

When we asked children what they saw on the page, only four referred to the creature. One of them referred to the creature's appearance ("I see a fox."). The remaining three referred to the creature's action (e.g. "I see a karate fox."), with one child also mentioning the resemblance to letters ("A fox doing karate. A fox making a C. A fox making a Q."). The rest of the children responded that they either saw letters (nine kids) or the icons for sample words (twelve kids). This observation puts into question how salient the creatures were for most children. However, it is also possible that some of their responses were guided by what they thought we wanted to hear from them (e.g. that we were asking about letters).

Since most children didn't refer to the creature at all while answering the first question, we followed up by pointing at the titular image of Katie and asking: "Who is that?" Seven children referred to the creature's action; two referred to the aspect of appearance closely connected to the action ("a ninja fox"); two - to the aspect of appearance disconnected from the action ("This is a fox." and "This is a kangaroo."); and one child responded with the creature's name. Combined with the observations from the previous paragraph, we see that children tended to conceptualize the creature in terms of its action. This is a useful property, since the actions are related to the phoneme sounds.

We found that all children except one understood the shared identity behind the different forms of the creature. When we pointed at its different instances, they consistently responded with the same name. When we asked them to find all Kathys, they consistently pointed at those shaped as several different letters, and usually pointed to all *[k]* creatures present on the page (both in the sample box and in the words). The one exceptional child responded that the creatures are "three

foxes." When we asked her whether those were the same or different foxes, she said (after some thought): "Different." Her opinion remained the same even after we tapped on each creature, making them say their name and play their action.

Similarly to how most children recognized the different forms of Katie as instances of the same character, 21 out of the 26 children recognized that they all produce the same sound; moreover, they were able to produce the sound. The 22nd child, Ericson, came close, but produced *[kw]* sound for the Q version of the creature (likely having such words as QUEEN in mind). The remaining children differed in their responses. One of them produced letter names instead of sounds. Ulisses correctly produced the *[k]* sound for each form, but in response to the question whether the sound is the same or different, said "different" (it was somewhat surprising to receive an incorrect answer from Ulisses, whom we knew for his strong literacy skills). Yet another child came up with slightly different sounds, and one child was simply not able to identify the sound even after the recordings were played. Nevertheless, the overall results appear encouraging. They suggest that creatures can be helpful in explaining the many-to-many relationship between the letters and the phonemes.

For those children who didn't mention Katie's action right away, we followed up with a question: "What is Katie doing?". Most described the action correctly. Two exceptions were noted, in which children thought that different forms of Katie were doing different things, e.g. "[She] makes a circle of himself, sticks leg out.". Such interpretations don't help connect the creature with the sound.

All children except one recognized the letters behind the creatures. In fact, as it was mentioned earlier, 9 out of 26 children immediately responded with letter names when we asked them what they saw. Thus, despite the potentially distracting visual details, children were able to perceive letter shapes in the design of the creatures with relative ease.

Children's understanding of the design principle behind the creatures also appears to be manifested, albeit cryptically, in some of the things they said. For instance, when asked "Why do they have different shapes?", two children gave seemingly puzzling responses — "Because they sound the same." and "Because they all start with *[k].*" Their responses would, however, make sense if interpreted as "Because they portray different letters, but sound the same." Another child, when asked the same question, responded: "Because the words. Because you see, over here — A, B, C, D". It is plausible that he meant: "Because they are supposed to look like different letters." Yet another child responded: "Because they turn into different ones.", indicating his knowledge of the variability of creatures' forms. Children's responses to the question "Why does she look like this?" fall into two categories: relating the creature to letters and relating the creature to its action (which is in turn related to its sound). Explanations of the first type sounded like "Because she wants to look like a letter.", "Because she can turn into C.", "Oh, because it is K and Q and C." Explanations of the second type sounded like "Because she does karate kicks.", "Because she does karate, and she wants to look like a letter." These responses seem to point at an understanding that each creature represents a sound via its action, but might take shapes of different letters.

Children didn't exhibit signs of parasocial relationship to the creatures. They perceived Katie as a generic fox, a generic ninja or a generic karate fighter. Only three children recalled the name of the creature, and in some of these cases, they may have heard the name right before saying it from the character's animation. Given that *[k]* was among the creatures that elicited the most reactions from children, it is unlikely that children remembered other creatures' names any more than that. Assuming that knowing a name seems to be an essential part of a social relationship, we should conclude that children generally didn't form parasocial relationships with the creatures. This is perhaps not surprising, considering how little of the character-defining information and relatable traits were introduced in their brief description, and that creatures didn't exhibit such interactive behavior as Mr. Fox (one they *did* relate to parasocially).

6.6.4. Quantitative Assessment of the Onomatopoeic Creatures

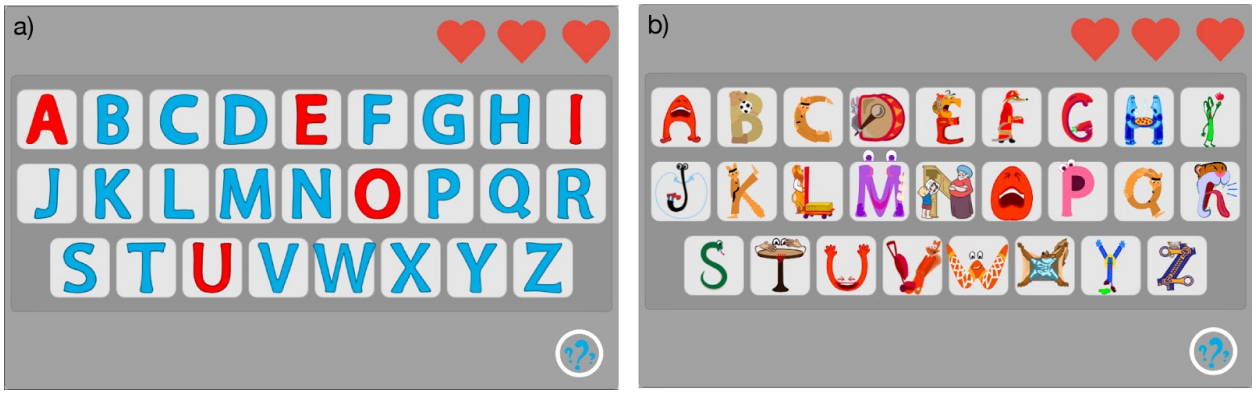


Fig. 6.54. Sound-Matching Mini-Game

The qualitative observations listed in the previous two sections suggest that the sound creatures may have served as memorable symbols for phonemes. To evaluate this assumption quantitatively, at the end of the SpeechBlocks II study, we[33] presented children with a mini-game designed to evaluate speed and accuracy of locating sounds on a keyboard. The keyboard had two modes: letter and creature ones (Fig. 6.54 (a) and (b), respectively) which acted as two conditions. To reduce the number of confounding factors, I chose both keyboards to be in the alphabetic layout. That meant that the sound keyboard didn't display all phonemes available, and the phoneme *[k]* appeared three times (as K, C, and Q). In the restricted context of the game, this was not an issue: the prompts were restricted to the phonemes present on the keyboard, and K, C, and Q were all treated as correct responses for *[k]*. At every turn, the game played a prompt asking the player to look for a particular phoneme, e.g. "Can you find *[k]*?", then registered the number of taps on the keyboard and the time elapsed between playing the prompt and finding the correct sound. I was concerned that children would get stuck on sounds they didn't know, so I limited the maximum number of tries for each sound to three. I used the video game metaphor of "health points" (displayed as hearts in the top right corner of the screen) to convey this to children. When an

[33] Me and my colleague Jim Gray

incorrect option was selected, a discordant tune was played, and one heart went away; when a correct option was selected, a cheerful tune was played. The question button at the bottom right allowed the child to hear the prompt again, if s/he forgot or didn't catch it. There were eight rounds in the game, preceded by two practice rounds, which were intended to help children get familiar with the interface in letter and creature modes. Half of these rounds employed the letter keyboard, and the half employed the creature keyboard. In the following week, the experiment was repeated with conditions switched: a child who had earlier received a particular phoneme in the letter condition now received it in the creature condition, and vice versa. This way, for each phoneme and each child, we had responses in both conditions, allowing for more direct comparison and minimizing the data sparsity issues. Since the two data collection periods were separated by a week, I considered the possible effects of memorization or fatigue negligible. The phonemes selected for the prompts were: *[t]* and *[b]* for the practice rounds and *[r], [w], [k], [z], [m], [s], [d],* and *[f]* for the testing rounds. This choice was dictated by several factors: (a) these phonemes had a pretty straightforward match to a letter; (b) they were all consonants, which are arguably more familiar to children of this age than vowels; (c) the corresponding creatures had animations that appeared well-executed, clear, memorable, and children had demonstrated a noticeable interest in them during the course of the main study. Therefore, these animations offered a better chance to evaluate the potential of the mnemonics without being limited by shortcomings of the particular implementation.

Modeling

Analysis of the data coming from this experiment is somewhat complicated by the fact that data points have multiple dependencies. There are multiple responses coming from each child; there are also multiple responses associated with each phoneme. It is entirely plausible that some children are faster than others, and that some phonemes are more difficult to match than others. Furthermore, children might differ on whether letter or creature mode is easier for them, and the same is true for different phonemes. Therefore, simple significance tests (such as paired t-test), which assume independence between samples (or sample pairs), can yield misleading results in analysis of such data. Fortunately, there is a technique designed specifically for this type of interdependency between data points: Linear Mixed Effects Models (for a nice introduction, see (Barr et al., 2013)). A brief explanation of these models is needed to help the reader interpret the results described below. Mixed Effects Models treat a particular dataset as a sample from the space of hypothetical datasets which could theoretically be collected with various sets of subjects and various test items. They presume that distribution of the response variable (in our case, response time or number of incorrect attempts) depends on a linear combination of predictor variables (in our case, which keyboard was used, plus one-hot variables indicating particular child and particular phoneme). The coefficients in this linear combination are of two types. One type, called *fixed effects*, is the same for all of the hypothetical datasets and represent fundamental regularities that we want to unearth (e.g. the effect of the treatment). Another type of coefficients, called *random effects*, represents quirks of particular dataset; they are called random since they will be different for different datasets and are assumed to come from a random distribution. In our case, there are random effects for each child and for each phoneme. We are interested not as

much in the values of random effects for our particular dataset as in the distribution of these coefficients in general case. Inference algorithms presume that the random effects are normally distributed, and estimate the parameters of these distributions from data. There are two types of random effects: random intercepts and random slopes. Random intercepts are added to the bias constant in the linear combination and affect the response without regard for predictor variables. For instance, they represent how fast each child is, or how difficult each phoneme is. Random slopes appear as coefficients for the predictor variables; they describe peculiarities in response to these variables. For instance, they represent individual preferences towards letters or creatures for every child, or in case of every phoneme. Different researchers provided different recommendations on which random effects should be included in the model. In this work, we use (Barr et al., 2013) recommendation to include all random effects that can be unambiguously inferred from the data. Therefore, we include both random intercepts and random slopes for both children and phonemes.

There are both frequentist and Bayesian approaches to mixed effects modeling. Although the work of statisticians is traditionally associated with frequentist approaches, Bayesian approaches have become very popular in recent years. In this work, I use a Bayesian model, implemented in *brms* (Bayesian Regression Models using Stan, (Bürkner & Others, 2017)) toolkit for R, as my primary inference method, and use a frequentist model, implemented in *lme4* (Linear Mixed Effects models 4, (Bates et al., 2015)) toolkit for the same language, as a check. The choice of the Bayesian model as the primary method is motivated by several factors. Most importantly, aside from the main hypothesis testing, it allows us to do a great range of statistical inferences in a simple and uniform way: by drawing samples from the posterior distribution via the sampling algorithm. This is further supported by Bayesian models treating random effects as coefficients, while frequentist approaches treat them as part of the error term (Bürkner & Others, 2017). These properties are vital for some of the inferences I will make. The second advantage of Bayesian approaches is that, according to some researchers, credible intervals have more intuitive properties than frequentist confidence intervals (Morey et al., 2016). Finally, the frequentist mixed effects algorithms tend not to not converge inon complicated models (Frank, 2018) an issue that I experienced in my analysis as well. Bayesian models typically do not have these convergence issues.

As is always the case with Bayesian modeling, the choice of priors comes into question. By default, the *brms* package uses weakly informative priors (Bürkner & Others, 2017), which tend to not introduce a significant bias into the analysis. For instance, the prior for fixed effects is the improper[34] prior over reals, meaning that any candidate value of the effect is equally preferred by the prior. I chose to resort to default *brms* priors, since I don't have any additional information that would motivate a better choice of prior. I check that my Bayesian model is not too far off by comparing the results with those of the frequentist approach.

A few words should be said about modeling of the response distributions: one for the response times and another one for the numbers of mistakes. In the case of the number of mistakes, the

[34] The prior is called "improper", since it is not a true probability distribution

distribution is clearly not normal. In most cases, the number of mistakes is 0, but occasionally it can be 1, 2, and 3; it is never negative. Therefore, we have a highly skewed discrete distribution. A fitting distribution family for such a situation is geometric. It represents the number of tosses of a weighted coin before it lands on heads. It is a special case of negative binomial distribution, and can be modelled in this way in *lme4*. In *brms*, it can be used directly.

For the distribution of response times, we first need to decide what data points count. We cannot count the data points where the correct answer has not been reached, because we don't know how much time it would take for the child to reach it. Should we count all the remaining data points, or only those in which the correct answer has been reached immediately? I tried both options, and the model yielded quite similar results. In this document, I will report results pertaining to the first option (that includes all data points where the correct answer has been reached). Second, the distribution of times is also skewed (Fig. 6.55, a): typically, children respond quickly, but sometimes it might take them a minute or so to find the correct answer. However, when I looked at the orders of magnitudes (in other words, logarithms) of time spent, I found that this distribution is much closer to normal (Fig. 6.55, b). From the modeling perspective, having response times distributed in this way makes a lot of sense: if our model predicts one response to take 0.1s and another one to take 40s, we can't expect the absolute error for these predictions to be the same. But relative error can be the same, and since logarithms convert multiplication into addition, they allow for modeling relative error in additive terms. Based on these observations, I use models with normal response distributions to fit the logarithm of response times.

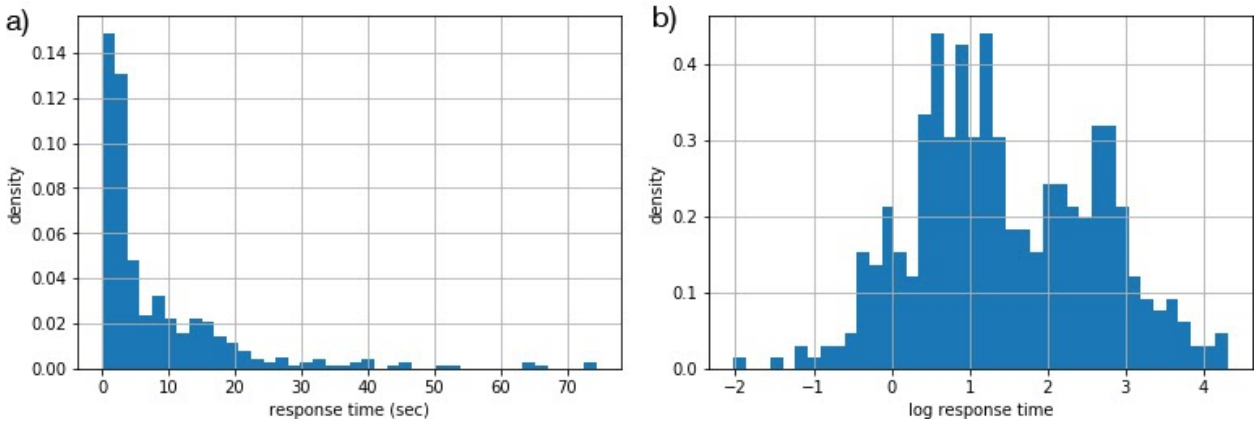


Fig. 6.55. Distribution of Response Times on (a) Linear Scale and (b) Log Scale

Results

With the setup complete, this section will focus on the analysis. First, it can be noted that the overall statistics for the two conditions look similar (Fig. 6.56). For 208 overall tasks in each condition, children made 122 errors in 55 tasks in the letter condition and 110 errors in 58 tasks in the creature condition. The mean response times were 5.97 seconds in the letter condition and 6.76 seconds in the creature condition. The median response times were even closer: 3.1 seconds and 3.17 seconds, respectively.

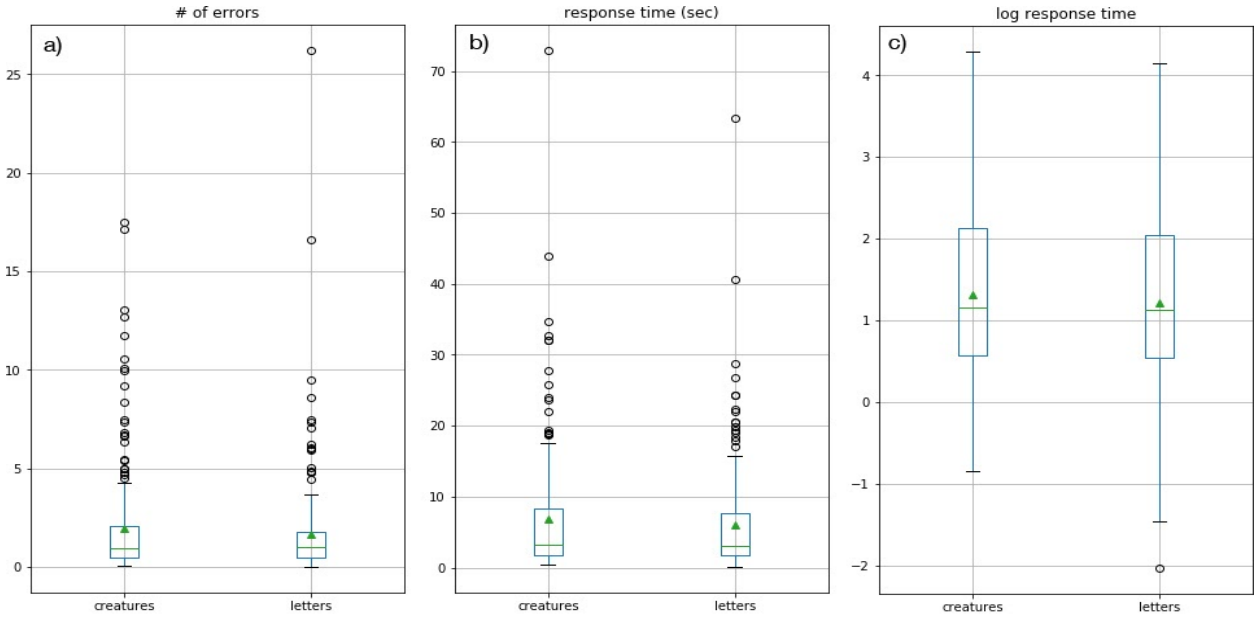


Fig. 6.56. Boxplots for (a) Number of Errors, (b) Response Times and (c) Log-Response Times in Letters and Creatures Condition

With the statistics in the two conditions being so similar, it is not surprising that none of the four models (for the response times and the numbers of errors, frequentist, and Bayesian) showed a significant fixed effect of condition. Table 6.1 provides estimates of this effect for different models, and table 6.2 shows average changes in the response variables associated with these effects. Fig. 6.57 shows the posterior distribution of estimates of the effect of letters on the error rate ratio and response time ratio. I conclude that in the general case, the differences between the two conditions seem to be small enough to preclude reliably establishing them from the data.

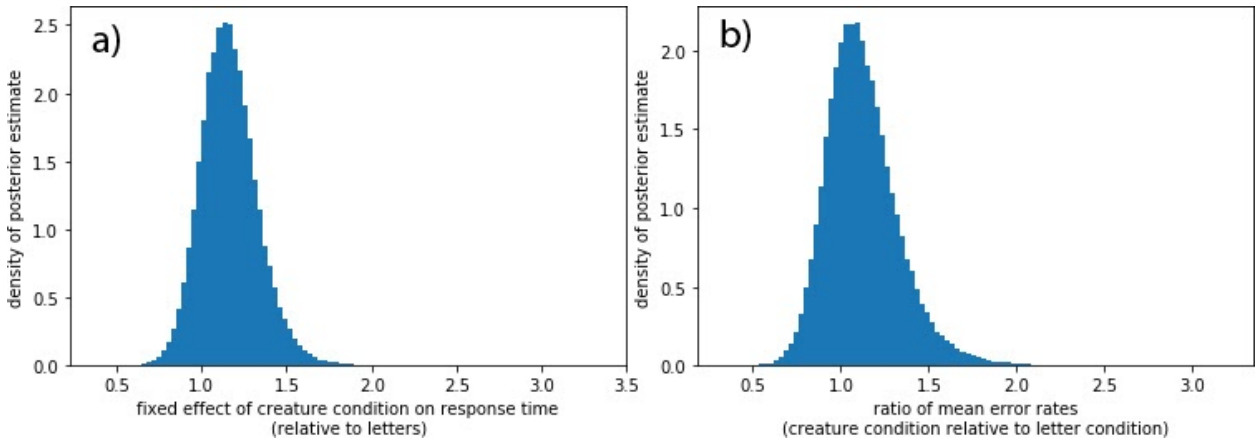


Fig. 6.57. Posterior Estimates of the Effects of the Creature Condition on (a) the Error Rate Ratio and (b) the Response Time Ratio

Table 6.1. Fixed Effects of Letter/Creature Condition for Different Models

Model	Estimation Method	Fixed Effect Estimates (for Letter Condition) (credible interval for brms and confidence interval for lme4)		
		Low-90% Bound	Expectation	High-90% Bound
Resp. times	brms	-0.38	-0.14	0.09
	lme4	-0.35	-0.14	0.07
Error rates	brms	-1.01	-0.38	0.19
	lme4[35]	-0.87	-0.37	0.03

Table 6.2. Effects of Letter/Creature Condition on Response Variable for Different Models

Variable	Estimation Method	Average Effect of Letter Condition (credible interval for brms and confidence interval for lme4)		
		Low-90% Bound	Expectation	High-90% Bound
Resp. times	brms	1.46 times faster	1.15 times faster	1.1 times slower
	lme4	1.41 times faster	1.15 times faster	1.07 times slower
Error rates[36]	brms	1.5 times less errors	1.12 times less errors	1.2 times more errors

Does that mean that children are generally indifferent (in terms of sound-finding efficiency) to whether they see letters or creatures? Or are there "letter-lovers" and "creature-lovers" that cancel out each other's contribution in our data? Examination of the data suggests the second option. First, there are qualitative observations of "letter-loving" behavior. Four children pronounced letter names after hearing the target sound, even when being in the creature mode. Three children (one having been in the previous group) used an alphabet song to facilitate their searches for the correct block — again, even in the creature mode. These observations suggest that they were looking for particular letters. Second, several children dramatically differed in their performance between the two conditions. For instance, Arnold made 7 errors in the letter mode and 0 errors in the creature mode, and Edward's corresponding tallies were 14 vs. 3, while several other children's tallies were 0 vs. 3-4 in favor of letters. For some children, the median ratio of response times for the same phoneme was strongly in favor of creatures (up to 3.2 times), while for others it was strongly in favor of letters (up to 6.7 times). However, as large as these differences are, there is still a possibility that they might have emerged entirely by chance. A more reliable statistical analysis is needed.

[35] The full lme4 model didn't converge, so I had to remove random slopes for phonemes to achieve these results

[36] Nonlinearities in parametrization of the geometric response distribution make it impossible to directly translate effects into differences in error rates. Sampling was used to make this estimate in the case of the Bayesian model.

This is where the advantages of the Bayesian approach come to play. *Brms* estimates the covariance matrices of the probability distributions for random effects, and provides credible intervals for these estimates. We are interested in the credible interval for the standard deviation of random slopes for children. If the lower bound of this interval is zero, then there is no sufficient evidence to believe that children exhibit much variability in their response to the two conditions. But if it is above zero, then such variability likely exists. Since we presume the normal distribution of random slopes, that means that there will be some children for whom their random slope overpowers the fixed slope, and some children for whom it does not. Table 6.3 shows us that this is indeed the case. Therefore, for each of the conditions, there will be children who favor it.

Table 6.3. Standard Deviation of the Distribution of Random Slopes for Different Models

Model	Standard Deviation of Random Slopes for Children		
	Low-95% Bound	Estimate	High-95% Bound
Resp. Times	0.31	0.94	1.71
Errors Number	0.02	0.28	0.63

However, this answer is still not entirely satisfactory to a designer who needs to decide whether to focus the effort on developing and refining the creatures. What if there is only a small fraction of children who perform better in the creature condition? Or, even if this fraction is sufficiently large, what if the difference between the conditions for most of these children is small? In both of these cases, investing effort in creatures would not be a priority. Fortunately, samples drawn from the Bayesian models allow us to estimate the relevant fractions as well. Figs. 6.58 and 6.59 show the posterior estimates of the percent of children who perform (significantly) stronger in each condition, at various levels of difference in performance (e.g. 1.25 times less errors, 1.5 times less errors, etc). Tables 6.4 - 6.7 show the means and the 90% credible intervals for these estimates. In the case of error rates, we can conclude that for each condition, there is a sizable fraction of the population that performs better in it. This can be concluded even when we look at a high (up to 1.5 times) difference in the level of performance. In the case of response speed, for each condition, there likely is a sizable fraction of the population that performs better in it, but we cannot say it with confidence for creatures. Moreover, when we look at high difference levels (1.5 times and higher), we see that even the posterior mean estimate for such a fraction is small. In other words, creatures hold a sizable advantage for a sizable fraction of children in terms of accuracy, but likely not for a sizable fraction in terms of response speed. Letters hold a sizable advantage for sizable fractions of children both in terms of accuracy and response speed. Of course, these extrapolations are only correct if our sample is at least somewhat representative of the general population.

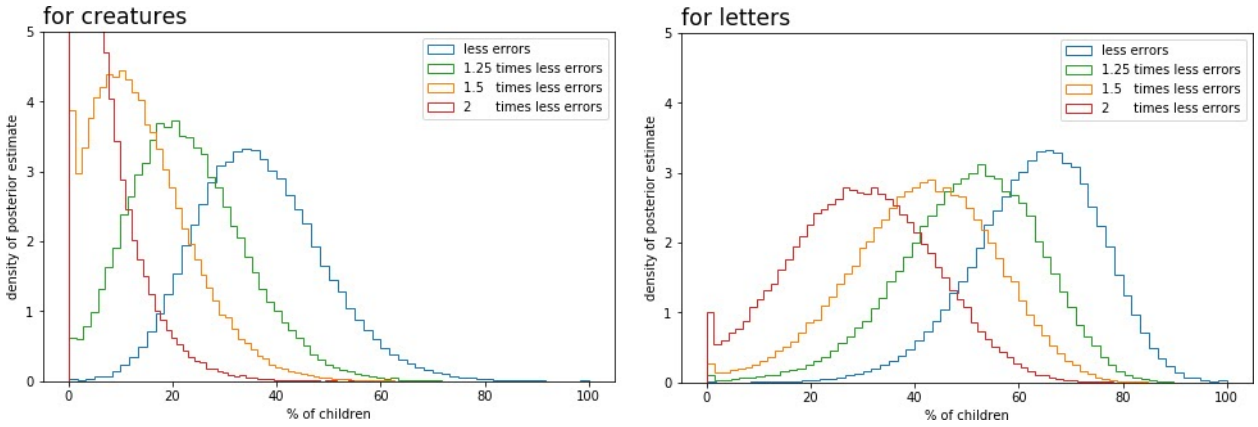


Fig. 6.58. Estimates of the Fraction of Children who Make (Significantly) Less Errors in Each Condition

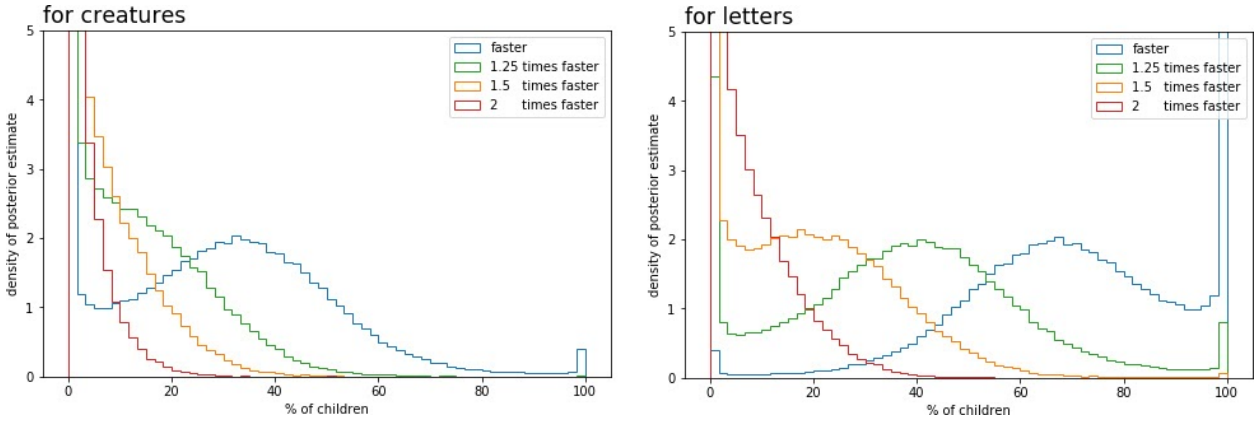


Fig. 6.59. Estimates of the Fraction of Children who Find Sounds (Significantly) Faster in Each Condition

Table 6.4. Estimates of the Fraction of Children who Make (Significantly) Less Errors with Creatures

Error Rate Ratio (in favor of creatures)	Fraction of Children with this Error Rate Ratio		
	Low-90% Bound	Expectation	High-90% Bound
>1	18.1%	36.7%	58.1%
>1.25	6.3%	22.9%	42.7%
>1.5	1.3%	14%	31.8%
>2	0%	5.9%	18.2%

Table 6.5. Estimates of the Fraction of Children who Make (Significantly) Less Errors with Letters

Error Rate Ratio (in favor of letters)	Fraction of Children with this Error Rate Ratio		
	Low-90% Bound	Expectation	High-90% Bound
>1	41.9%	63.3%	81.9%
>1.25	26.8%	50.3%	70.8%
>1.5	16.9%	41.3%	63.1%
>2	6.9%	29.7%	52.3%

Table 6.6. Estimates of the Fraction of Kids who Find Sounds (Significantly) Faster with Creatures

Speed Ratio (in favor of creatures)	Fraction of Children with this Speed Ratio		
	Low-90% Bound	Expectation	High-90% Bound
>1	0%	30.3%	64.3%
>1.25	0%	12.5%	36.5%
>1.5	0%	6%	22.8%
>2	0%	1.9%	9.5%

Table 6.7. Estimates of the Fraction of Children who Find Sounds (Significantly) Faster with Letters

Speed Ratio (in favor of letters)	Fraction of Children with this Speed Ratio		
	Low-90% Bound	Expectation	High-90% Bound
>1	35.7%	69.7%	100%
>1.25	1%	38.1%	76.1%
>1.5	0%	18.7%	46.7%
>2	0%	6%	22.4%

These findings raise a question: which children perform better in each condition? It is natural to assume children with stronger letter-to-sound knowledge would prefer the letter condition, whereas the creature condition might be helpful for children who have weaker letter-to-sound knowledge.

Unfortunately, we didn't measure children's letter-to-sound knowledge in the pre- and post-tests. No indirect evidence[37] supporting this assumption was found in the Bayesian models either.

It is also interesting to look at Bayesian estimates of random effects for particular children[38]. These estimates suggest for whom letters or creatures work better. For both Arnold and Edward, the two children who were observed frequently reacting to the creatures, we notice random slopes favoring the creatures both in the error and in speed models (more pronounced for accuracy). For Edward, the entire 95% credible interval of his random slope in the error model is above zero, which means that we can be confident that he makes fewer errors with creatures. Furthermore, all children who mentioned letters or used the alphabet song during the game have random slopes favoring letters in the error model, and typically in the speed model as well.

6.7. Learning

In this section, we look at the available evidence on whether SpeechBlocks was successful in helping children raise their phonological awareness (PA). The answer to this question is not straightforward. I haven't seen a significant difference in the PA gains between the treatment and control conditions (even though treatment did perform slightly better). However, when I looked at the gains with respect to several potential explanatory variables, I found that children with higher initial PA and higher executive function (EF) seemed to have benefitted from the app, while children with low PA and EF might not have. This picture matches well with the qualitative observations from section 6.2. Indeed, children with high PA and EF tended to use the app for focused and sophisticated play that involved building large amounts of words. On the other hand, children with low PA and EF tended to be distracted by counterproductive behaviors. The exploratory analysis also suggests that boys might have benefitted from the app more than girls. However, this trend is hard to explain, and it is more likely to just be a fluctuation in the data.

To find evidence of learning, I looked at CTOPP gains: the differences between the initial and the final scores. Gains in two kinds of measures can be considered: raw CTOPP scores (which I computed simply as the number of correct answers on the three PA sub-sections of the test) and PA composites (which are computed using scaled scores for each sub-section, adjusted for the child's age). However, although the treatment group performed better on each of these measures, neither difference was statistically significant (Fig. 6.60; t-test was used to compute p-values).

[37] Such indirect evidence would be a correlation between random intercepts and random slopes for children. Such correlation would indicate that children who are doing better with the task in general (likely because of their better letter-sound knowledge) exhibit preference towards letters (or towards creatures). However, no such correlation was found.

[38] Using function *ranef* of *brms* toolkit.

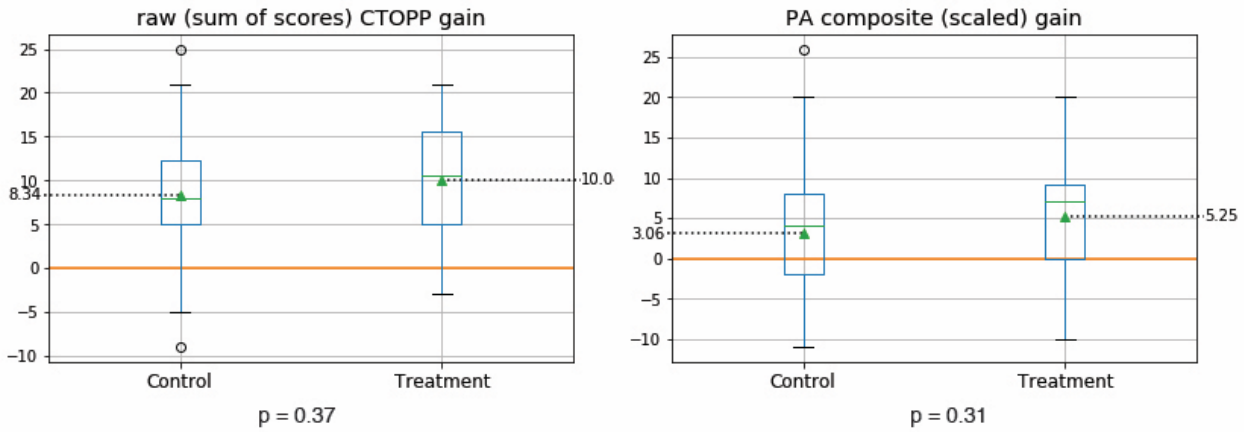


Fig. 6.60. Comparisons of CTOPP Deltas for Sum of Raw Scores and PA Composite (Using Scaled Scores)

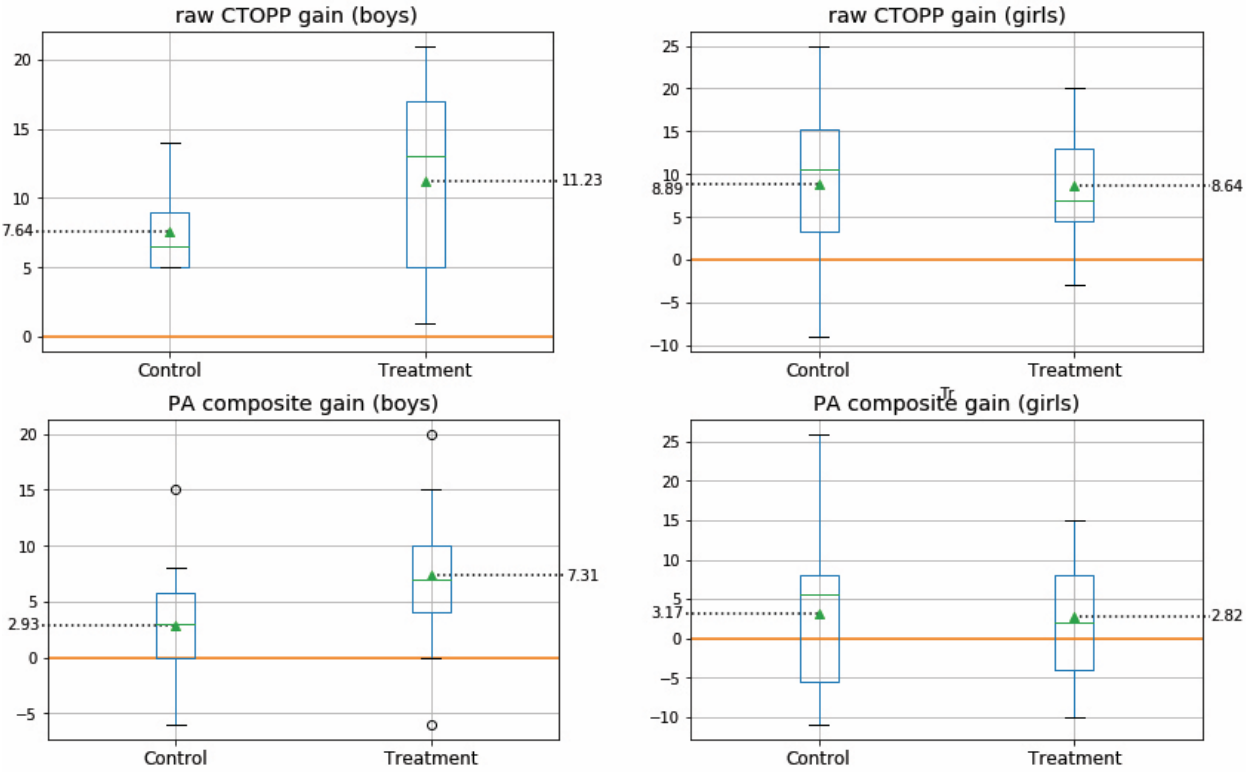


Fig. 6.61. Comparisons of CTOPP Deltas by Gender

However, given the notable differences in how children played with SpeechBlocks, it is plausible that some groups of children might have benefitted from the app, while others might not have. For instance, we saw that focused, literacy-intensive behaviors, such as imaginative play, tended to be associated with high levels of PA and EF, while impulsive, distracted, and unproductive behaviors tended to be associated with low levels of these measures. Therefore, PA and EF may act as mediators determining how much children learn from SpeechBlocks. Furthermore, I decided to

include gender as a potential mediator, because exploratory analysis showed the curious difference in CTOPP gains between boys and girls (Fig. 6.61).

As a result of including these potential mediators, I arrived at the following multiple regression model:

CTOPP gain ~ condition + pre-CTOPP + pre-EF + gender +
*+ condition * pre-CTOPP + condition * pre-EF + condition * gender*

Here, I provide the result for raw (not scaled) CTOPP scores. Raw scores are indicative of children's overall level of phonological awareness, which seems to be a natural factor. In Appendix C, I provide a similar analysis for the PA composite. The variables *pre-CTOPP* and *pre-EF* were normalized (by subtracting the mean and dividing by the standard deviation). Because of the exploratory nature of this analysis, I didn't try to mitigate the potential multiple comparisons issues. Table 6.8 shows the regression result.

Table 6.8. Regression Analysis of Raw CTOPP Deltas

Overall p-value: 0.1. F-statistic: 1.819 on 7 and 48 DF.						
variable	coefficient	p-value	low-95% bound	high-95% bound	low-90% bound	high-90% bound
treatment	-1.63	0.54	-6.9	3.64	-6.02	2.77
pre-CTOPP	**-2.93**	**0.046 ***	-5.81	-0.06	-5.33	-0.53
pre-CTOPP X treatment	**4.66**	**0.021 ***	0.73	8.59	1.38	7.94
pre-EF	-1.95	0.15	-4.65	0.74	-4.2	0.3
pre-EF X treatment	**3.63**	**0.07 ·**	-0.33	7.58	0.32	6.93
gender (m)	**-4.60**	**0.1 ·**	-10.11	0.90	-9.2	-0.012
gender (m) X treatment	**6.93**	**0.08 ·**	-0.84	14.70	0.45	13.41

Indeed, we see significant or nearly significant interactions for pre-CTOPP and pre-EF. Fig. 6.62 illustrates this phenomenon. We see that in the control condition, the lower was the initial CTOPP, the higher was the gain. This is plausible, given that the teachers were likely targeting their curriculum to lower-performing kids to let them catch up. On the other hand, in the SpeechBlocks condition, the higher was the initial CTOPP, the higher was the gain. Similar (albeit less pronounced) pattern can be observed for EF (however, this pattern is not observed for PA

composite — see Appendix D). This aligns well with the aforementioned qualitative observations. Furthermore, in two previous studies, similar patterns were observed: children with low self-regulation weren't able to benefit from early literacy software (Kegel et. al., 2009; Kegel and Bus, 2012). However, I was unable to find observations that would convincingly explain the gender-related trend. It is possible that it was merely a fluctuation in the data; however, it is interesting to note for further research.

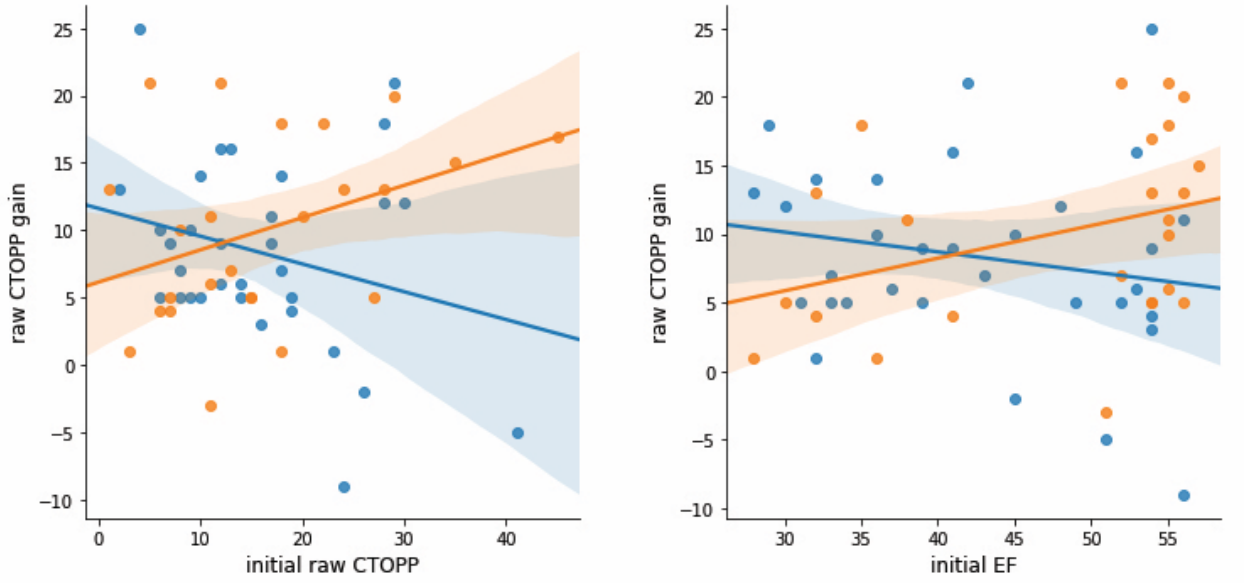


Fig. 6.62. Interaction plots of CTOPP gain for initial CTOPP and EF

The present analysis is exploratory, and its results cannot be considered statistically significant findings. Rather, they should be viewed as interesting patterns for further investigation. Their validation (or invalidation) remains a subject of future research.

Chapter 7. Conclusion

The previous pages detailed the studies of two apps embodying a particular approach to digital technology for early literacy learning. This approach aims at making the learning experience intrinsically motivating for children. In order to do so, it makes the experience **child-driven**, so that learners can connect their play to their lives and interests. It embraces children's natural drive for **expression** and uses it as a vehicle for the learning process. The envisioned role of the machine "intelligence" is not the one of an instructor, but one of a gentle guide. Such a guide **scaffolds** the children's efforts: it does its best to understand their intent and help them to achieve their goals. Taking a step back to look at the picture that emerged from using this approach in practice, I will restate the key learnings, this time backing them with references to the data. After that, I will list a few specific suggestions for designers and educators who are interested in the approach. Finally, I will suggest new directions in which this approach can be taken further.

7.1. What Have we Learned?

Below are the key learnings of this thesis, fleshed out by the data.

1. As anticipated, **the media sparked intrinsic motivation to play, supported the senses of agency and self-efficacy, and allowed children to express themselves in non-trivial ways.**

Intrinsic motivation can be seen in children's requests to play with SpeechBlocks after the end of the first study (section 5.4.1), in them being able to sustain engaged play for half an hour or more (sections 5.4.2 and 6.2.3), in their focused and self-directed play, in their persistent efforts to get their words to sound right through repeated attempts (section 5.2.2), in challenging themselves to build sophisticated words (sections 5.2.2 and 6.2.1), in their excitement in telling others about SpeechBlocks (section 5.4.1), and their delighted remarks (like "This is the coolest iPad I've ever seen!") during their play with the media.

The sense of **self-efficacy** manifested itself in children's proud displays of their work to their peers and to adults (sections 5.5 and 6.4). Their verbal expressions, e.g. "Look what I made!", point to their satisfaction with their ability to build words of their choosing, and their desire for others to notice that ability. Self-efficacy was also reflected in a sense of ownership of their work. For instance, children wanted to keep words that they made in SpeechBlocks I (section 5.3), and were excited to recognize their works made in SpeechBlocks II (section 6.3). The drive for self-efficacy can also be seen in children challenging themselves to build words without scaffolding (section 6.3). For instance, in the SpeechBlocks II study, they were sometimes seen removing the headphones in order to silence the prompts and make words by themselves.

Children's **agency** was manifested in them choosing to work on topics connected to their interests and personal lives. In SpeechBlocks II, they created, vigorously enacted, and passionately

talked about scenes and stories involving superheroes, cartoon characters, jungle animals, families, etc. (section 6.2.2). Their passionate play (see examples in sections 5.2 and 6.2) reveals their investment in what they were making. A curious manifestation of children's agency was their phantasmagoria (section 6.2.2). Children were extremely excited to spell things that they felt a strong personal connection to. Remarkable in that regard were names — their own and people they knew; first names and family names. During deployment of SpeechBlocks I at home, children used their freedom of play to spell various things connected to their lives (places they visited, messages on the walls of their house, events that happened to them, names of actors and singers, etc.), as well as using the app in various creative ways (spelling messages and directions for others, creating lines from their favorite songs and "playing the songs" by tapping on the lines, engaging in wordplay, etc.) Players also enjoyed creating nonsense words (particularly by remixing other words), without being evaluated on the correctness of the words they made (section 5.2.1). Children independently came up with various forms of shared play (sections 5.5 and 6.4). A notable manifestation of learners' agency was creating plans and following them (section 5.3). On the other hand, even simple forms of agency — doing whatever they wanted with the app, however random it was — had strong appeal for some children (section 6.2.3).

Children were able **to express themselves in non-trivial ways.** Particularly notable are the phrases and sentences created during SpeechBlocks I home deployments (section 5.2.2) and the scenes created by imaginative players (section 6.2.2). Phrases in SpeechBlocks I included recounting events that occured in children's lives, describing simple scenes, communicating with others and rendering songs. Occasionally, there were also interesting cases of word play. In SpeechBlocks II, children at times created complex scenes that could display action and involved more than 10 objects that were highly interrelated. Complimenting the scenes, were children's verbal accounts of what was going on in them. They contained rich imaginative details and sometimes multi-step scenarios. Similar features were manifested in children's enactments of stories, which they did by moving objects akin to physical toys. These examples suggest that the present approach may help children to develop their imagination and narrative skills while working on their early literacy skills.

2. There were markedly different ways of using the media. Six play types were observed for SpeechBlocks I (section 5.2):

- **Word Crafting.** This type of play is noticeably more prevalent than in SpeechBlocks II, because of the word-oriented nature of the old SpeechBlocks. While 4 to 5 y.o. children primarily copied words from cards, 6 to 10 y.o. in home conditions created a variety of words related to their lives: places they visited, food they liked, shows, characters, actors and singers they were interested in, and brands they saw. Usage of invented spelling was prominent for older children.

- **Proto-Narrating.** This type of play roughly corresponds to imaginative play in SpeechBlocks II in a sense that children tell stories with or about words and phrases they build. These were not only imaginary stories, but also accounts of events that happened

with the child. For the in-classroom study, which involved younger children, we mainly saw verbal narrations. During in-home studies, we saw collections of words and phrases that looked like stories (e.g. FEELBAD DOCTORSHOT COLD SICK FLU GURMS).

- **Remixing and Rhyming.** This play type is centered on morphing words into other words. A very simple form of remixing is concatenating real words to obtain nonsense words. This was a very common activity in the early period of using SpeechBlocks. It served as a nice first step into the app for many children. A more complex form of remixing transforms real words into other real words and resembles the kind of wordplay found in poetry. Such play is uncommon, but remarkable. Finally, some children in home studies actively explored rhymes.

- **Communicative Play.** The focus of this play type is on making SpeechBlocks speak for the child, typically when the child addresses someone else. Children made greetings and wishes (e.g. HAPPY BIRTHDAY), told others their feelings (e.g. LOVEYOMOM) and opinions (e.g. IDKARE — "I don't care"), issued commands to their pets (e.g. CODYSIT; Cody was a dog), and even teased their siblings. A form of play closely related to communicative play is forming lines of songs in order to make SpeechBlocks "sing" them.

- **Using the App as a Reference.** Children used SpeechBlocks as a source of words that they copied down on a sheet of paper. They also used it to hear the pronunciation of words they didn't know how to read, but could input into the app letter-by-letter.

- **Impulsive Exploration.** This form of play is associated with "probing" the app via erratic actions, without much apparent planning and deliberation.

Types of play observed with SpeechBlocks I roughly correspond to those with SpeechBlocks II, but some differences exist. Some of these differences can be attributed both to different affordances of the media, e.g. SpeechBlocks II allowed for richer narrative behaviors via scene construction, but made remixing much more difficult. Other differences can be attributed to the different environment and age of participants: some forms of SpeechBlocks I play manifested only in home studies with older children. Three main types of play with the app were identified (section 6.2):

- **Word Crafting.** This type of play is characterized by the intrinsic interest in word making as a goal in itself. Children may be attracted to word crafting from the desire to exercise their mastery of spelling, and may challenge themselves to spell words without scaffolding. They can be observed making sophisticated, unusual words. Children enjoy collecting words they made.

- **Imaginative Play.** This type of play uses the sprites resulting from word creation to tell a simple story. It can be done through assembling a static picture, or through enactment by

using the sprites as physical toys, as well as via combination of the two. Verbal narration can compliment within-app play. This form of play is related to proto-narrating in SpeechBlocks I, but much richer. It is demanding, since it requires constructing large numbers of words without veering off-track.

- **Impulsive Exploration.** Similar to SpeechBlocks I, this form of play is driven by short-term rewards (emotional, social, and cognitive) and is characterized by lack of systematicity. Long-term plans may be expressed, but are rarely followed through. Impulsive exploration often appears dynamic and passionate, but chaotic. Impulsive explorers experience difficulties building words; in order to compensate, they find other ways to have fun with the system. However, some of the children who gravitated towards this play type exhibited gradual transition towards more systematic and focused activities.

3. The media encouraged various forms of social play centered on word-making. Such play can potentially serve three functions for learning:

- **Maintaining mutual engagement.** An important role of peers was to be an audience for the player. Children constantly showed and talked about their work (and sometimes other things that excited them in SpeechBlocks) to their peers. The peers often responded with interest and excitement. These emotional exchanges likely fueled children's desire to build words. Children were also entertained by various forms of shared play that they came up with themselves. In home studies with SpeechBlocks, we see some play sessions focused on exchanges between siblings.

- **Providing mutual inspiration.** The simplest form of social interaction was just observing other children. From such observations, players saw what could be done with SpeechBlocks, and were incited to try it themselves. Both word crafters and imaginative players actively borrowed ideas from each other.

- **Enabling learning from each other.** We saw multiple examples of children directly helping each other — sometimes with user interface, but more importantly — with literacy tasks.

4. Real-time, built-in scaffolding for making child-selected words is essential for maintaining meaningful participation of early literacy learners. Without scaffolding, 4 to 5 y.o. children were mainly able to create nonsense words and a few sight words. Making nonsense words was initially fun for them, but started to quickly exhaust itself. While I originally thought that children might be interested in exploring spelling patterns until they gradually make their way to real words, this was not the case of 4 to 5 year-olds. Thus, without scaffolding, there was no natural segue to more sophisticated activities. As a result, in the first SpeechBlocks I pilot, we saw children gradually appearing less and less interested, to the point that they would sometimes leave the station prematurely. The discontinuity between the entry-level and advanced activities was less sharp for older (5 to 10 y.o.) children, a large fraction of whom started to build real words via

invented spelling and tinkering with words until they were able to make them sound right. Nevertheless, expressing oneself in this manner took a great amount of time and effort. This might be one of the causes of quick drop of engagement with SpeechBlocks in home conditions. Furthermore, the value of SpeechBlocks is less for the children at the upper end of this age range, because they typically already have relatively well-developed phonological awareness.

Conversely, children appreciated opportunities to build real words they were interested in. In the first SpeechBlocks I pilot, the facilitators received a steady stream of requests concerning building specific real words. Furthermore, a pronounced rebirth of engagement in play was observed when character cards were introduced and children received an opportunity to independently build words they found interesting. However, by restricting the vocabulary, character cards failed to utilize the open-ended nature of the app. In SpeechBlocks II play, children's interest in building real words was reflected in their choice of activities: an overwhelming majority of the words constructed during the study were real words built using scaffolding. Because making real words is so vital for children's engagement, we found scaffolding of word construction to be an essential element of the approach.

Such scaffolding should respond in real-time to children's goals. We saw how in the process of their play children fluently responded to emerging ideas — whether it is something new that they saw in their scene, something suggested by the association's network or something inspired by their peers. This fluency reflects the Csikszentmihalyi's concept of flow (Csikszentmihalyi, 1997). Flow is associated both with productivity and fulfillment. It is highly desirable to foster that state in children's play. If children are unable to immediately follow their ideas, the flow is disrupted. Therefore, asynchronous solutions to the problem of scaffolding, such as a coach interpreting children's play and sending suggestions to them in her free time, seem to be insufficient.

Our observations from the first SpeechBlocks I study (section 5.6) highlight how labor-intensive and difficult it is for a human to provide real-time scaffolding to multiple children. Even with just four children at a table, a researcher struggled to respond to their simultaneous requests. As a result, the adult's attention became a bottleneck that limited fluency of the children's play and precluded certain scenarios from unfolding. For instance, in SpeechBlocks II play, we saw a rapid exchange of ideas between children at the same table. Such exchange would be much more difficult if they had to wait for a human to scaffold the words they wanted to make. These observations compliment the theoretical arguments why an expressive literacy medium should include built-in scaffolding.

A few interesting discussion points can be made regarding the above-mentioned importance of built-in scaffolding. The first is whether there is some specifics to the literacy domain that made scaffolding so essential — in contrast to domains such as programming, where such constructionist systems as Scratch and ScratchJr flourished. For instance: (1) in the literacy domain, building blocks (letters or phonemes) are by necessity very low-level (e.g. it is very hard to make a system for building words out of syllables, because of how many syllables are in the English language) and (2) the results of children's actions are not visual, but auditory, thus requiring

an attuned hearing to observe. In systems such as ScratchJr (Bers & Resnick, 2015), where blocks are high-level and their effects are easily observable, transition from nonsense strings to meaningful programs might be smoother. Furthermore, while high-level building blocks in ScratchJr introduce a good amount of creativity and expression into even basic programming activities, the low-level process of encoding in writing is generally mechanistic and routine. This is not to say that learning encoding and decoding is entirely devoid of creativity — the phenomenon of invented spelling shows us the opposite. However, most avenues of meaningful expression in writing are associated with the level of the words. Scaffolding allows children to engage in high-level creativity and expression while simplifying the necessary routine.

This observation brings us to the second point: the apparent tension between the child-driven and scaffolded principles. Since my scaffolding design restricts children's actions, one might suspect that it limits children's agency. However, I argue that in SpeechBlocks II scaffolding actually increased children's agency and expressive capabilities by greatly supporting word crafting and imaginative play. A possible way to look at this is to consider agency as a multi-level phenomenon. At a low level, agency is associated with the freedom to assemble whatever sequence of blocks the child wants, as well as perform any other action within the system. The child is always free to do that in the open-ended mode, but scaffolding indeed restricts this type of agency. But by paying this price, children gain a higher-level agency: the ability to build real words and complex scenes. While low-level agency is enjoyable, the enjoyment it brings is limited compared to what high-level agency can bring. This is why we see a gradual shift towards more structured behaviors as children get a "taste" of them and become able to engage in them (section 6.2.3), which appears to be similar to Montessori's notion of "normalization" (Lillard, 1972). This view echoes Brennan's ideas about the mutual support between agency and structure (Brennan, 2013). It also has parallels to the multi-scale theory of complexity (Siegenfeld & Bar-Yam, 2019).

5. Different types of word scaffolding were observed to have different functions that complemented each other:

- **Responding to specific requests.** This function directly addresses the above-mentioned children's need to build specific real words. Children used scaffolding systems of this type to both express their own ideas and quickly borrow ideas from peers. Popularity of speech recognition as a word source (section 6.5.8) can be explained by the fact that it fulfills this function most naturally. This also partially applies to the Word Bank. However, the invented spelling interpreter, which also was designed with this function in mind, wasn't successful in fulfilling it (section 6.5.4). This happened because its demands on children's literacy skills turned out to be higher than what 4 to 5 year-olds possess.

- **Facilitating search for ideas.** This function helps children develop and expand their creations, sometimes bringing them in new directions. A most notable example of this is how children used the association network (section 6.5.6). A secondary example is using Richard Scarry's books along with text recognition (section 6.5.7). Scenes presented in the books inspired children's ideas about their own scenes.

- **Being a fall-back option** when children experience difficulties with more sophisticated technology. All "high-tech" modes of scaffolding had multiple failure modes (as described in section 6.5). These technical issues particularly strongly affected less patient and more impulsive children. A simple and reliable (even if limited) scaffolding option, such as Word Bank, allowed children to resort to it when more sophisticated systems failed them.

6. For 4 to 5 year-olds, **usage of letter vs. phoneme blocks turned out to be a less important factor than it was originally thought**. This was because most words were created in direct guidance mode (section 6.5). In this mode, both pronunciation and spelling of the blocks remain fixed, thus rendering the difference between the two block types irrelevant. There is currently no evidence suggesting the advantage of phoneme blocks. However, this observation is made in the context of classrooms that invested heavily in learning letters — the situation may have been different if children's curriculum was structured differently.

While the phoneme blocks themselves didn't seem advantageous, **the onomatopoeic mnemonics** (designed for them) **turned out to be useful for some children, although not for everyone.** Although we introduced the "sound creatures" to children only briefly during a few introductory demos, most children understood the principles behind them. For instance, they understood that a creature represents a particular sound, and that the same creature can have multiple shapes corresponding to multiple letters which all made the same sound. The creatures evoked significant interest in many children. They showed signs of positive affect, viscerally reacted to the animations, and repeated the sounds of the creatures and their actions. However, it appears that the creatures were sometimes perceived in a way which obscured how they were associated with their sound. This may be a limitation of the onomatopoeic approach. A custom mini-game presented to children at the post-test showed that a significant fraction of children benefited from the mnemonics in terms of speed and accuracy of finding blocks on the keyboard. However, there is also a significant fraction of children for whom conventional, unadorned letters worked better. The factors mediating whether or not onomatopoeic mnemonics were advantageous for children are currently unclear.

7. Initial phonological awareness (PA) **and executive function** (EF) **appear to be moderating factors in how productive children's engagement with the media will be.** It appears that imaginative play, perhaps the richest form of play in SpeechBlocks II, was associated with good initial PA and EF (section 6.2.2). Children with high PA and EF tended to be more focused, were less affected by technology issues, and by the end of the study were often able to play almost completely independently. Conversely, children with low PA and EF exhibited a lot of unproductive behaviors, such as constantly scaling sprites up and down, "taking pictures" of each other via text recognition interface, and performing a lot of random taps and swipes (section 6.2.3). However, some of these children developed more purposeful behaviors as the study progressed. Analysis of PA gains suggests that children with high PA and EF benefitted from playing with SpeechBlocks, while for low-PA-and-EF children that might not have been the case (section 6.7).

This pattern is similar to some earlier findings with a different literacy-oriented digital technology (Kegel et al., 2009).

If this pattern indeed holds[39], it is undesirable, because children with low PA and EF are actually those who need the most help. The observed dynamics even pose a risk that application of expressive literacy media can increase literacy gaps instead of mitigating them, further amplifying the "Matthew effect" in literacy development in which the rich get richer and the poor get poorer (Stanovich, 1986). It also puts under question the concept of using the media in home contexts, as an alternative channel of literacy learning in addition to classrooms. Therefore, these dynamics are an important issue to consider in further research.

Nevertheless, there may be ways to address this issue. First, my current implementation of scaffolding delivered the same type of guidance to all children. This was done to reduce the number of unknowns in the study. However, this is at odds with one of the key aspects of the scaffolding concept — working within the child's Zone of Proximal Development. It is possible that the current dynamic was observed because the app was too hard for children with low PA and EF, but just right for children high on these variables. In the future, it is desirable to conduct experiments with adaptive scaffolding, which functions differently based on the child's level of skill.

Second, it can still be of importance that learners with high PA and EF are able to play with the app nearly autonomously and still seem to derive significant benefits from such play. Because of this, in a classroom, the app might free up teachers' resources to focus on lower-performing learners. When the at-home scenario is considered, some structure can be put in place so that learners meet regularly with teachers or "literacy coaches" who support them and propel them towards the skill level when they are able to autonomously engage with the system.

The last remark points to the remaining importance of adult-provided support. Going back to the original scaffolding paper (Wood et al., 1976), we find that my simple scaffolding system covers only a fraction of scaffolder's duties that these authors outlined. Wood et. al. (1976) identified the following functions performed by a human scaffolder:

- Recruitment: encouraging children to engage in the learning task, rather than free play
- Reduction in degrees of freedom: reducing the scale of the task so that it is manageable for the learner
- Direction maintenance: preventing learners from regressing to other aims
- Marking crucial features

[39] The study didn't focus on quantitative rigour. It is possible that the observed pattern can be attributed to classroom-level effect, or issues with measurement of CTOPP and EF - see section 6.1 for description of potential issues. In addition, this pattern could have emerged if the age range that we chose was too low for the app, allowing only the strongest learners to use it efficiently.

- Frustration management (which can be stated more generally as emotional support of the learner)
- Modeling (which can also be generalized to encompass idea provision)

The built-in scaffolding in SpeechBlocks II fulfills two of these functions. It marks crucial features for children by sounding the target word out and by evoking connections to onomatopoeic mnemonics. It also reduces degrees of freedom by limiting the amount of keys on the keyboard, pre-filling some slots in the target word, and rejecting incorrect block choices. We also saw that some amount of modeling provision was done by children for each other. However, recruitment, direction maintenance, emotional support, and some amount of modeling need to be provided by an adult. These functions are particularly important at the initial stage of children's engagement with the media (when children don't yet know how to meaningfully use them), and also for struggling learners. Fully automating these functions appears hardly possible at the current level of technology. Furthermore, at any level of technological development, it may still be desirable for encouragement, appreciation, and other forms of emotional support to come from human beings. The emotional support and relationship building roles might become the primary ones for a teacher in a classroom equipped with expressive literacy media, while traditional instructional functions might be, to a large extent, taken over by the built-in scaffolding.

7.2. Suggestions for a Designer

Some of the learnings described in the previous section may directly translate into design suggestions. For example, the three observed roles of various scaffolding systems suggest designing three mechanisms to cover these roles. Many other design suggestions can be derived from the body of research on construction-based learning, in which the present work is situated. To avoid repeating them, I recommend articles by Resnick and Silverman (2005); Resnick and Rosenbaum (2013); Resnick's (2017) book (which includes the section "Ten Tips for Developers and Designers"); as well as the foundational book by Papert (1980). The work of Makini (2018), which inspired the design of SpeechBlocks II, may be helpful as well. A designer might also be interested in the specifics of how various technologies performed in the field; for that information, I refer him/her to section 6.5. In this section, I will mention a few additional suggestions which followed from the work, but were too specific to be included among main learnings.

1. Incorporate personally meaningful content — for example, names. Children's extreme interest in names (their own, their friends, and their relatives) is documented in sections 5.2.2 and 6.2.1. Other types of personally meaningful content can include items related to children's hobbies, their favorite actors and characters, places around them, occupations of family members, etc. Such incorporation should not be limited to the possibility of building these words within the system, but should involve the system "knowing" them internally. In this case, the system would be able to (for example) scaffold construction of these words by the child. Including content associated with these words (e.g. photographs of the people, to be used as sprites in scene

construction) can create additional motivation for learners to engage in word building. To input such content, a dedicated interface for parents or teachers/coaches should be considered.

2. Support children's sense of ownership of their work by allowing them to save and exhibit it. Sections 5.3 and 6.3 detail children's notable interest in keeping their work. For example, in the earliest version of SpeechBlocks, where saving words was not yet available, children proceeded to write down their creations in journals. Furthermore, they requested to keep these journals. These sections also detail children's desire to share their creations with others. Giving them easy means to save and exhibit their work can reinforce their motivation to play.

3. Account for impulsive behaviors. Seek to eliminate elements of design that provoke and reinforce such behaviors. When including new features, their advantages should be weighed against their potential distracting effect. In context of this work, such an effect took place with text recognition (section 6.5.7). Redesign interface elements that require carefully coordinated actions, or long sequences of actions: they tend to create obstacles for impulsive explorers. An example of a successful redesign is incorporating voice detection in the speech recognition interface. The problem with the original interface and improvement after the redesign are documented in section 3.3. It may also be useful to provide fall-back options to features that require some coordination and patience to operate. As to this work, such fall-back mode was the Word Bank.

Aside from these suggestions, I also would like to make a few data-derived speculations regarding **the choice of blocks and the keyboard design**. The strategy of scaffolding word construction sound-by-sound (as opposed to letter-by-letter) continues to appear reasonable. Therefore, I recommend that in the scaffolded mode, blocks on the keyboard should correspond to grapheme-phoneme pairs. But since children often used letters as cues to look for sounds, it might be reasonable to put the "default" graphemes for these sounds on the keyboard as well, even if they are not present in the target word. For example, it might make sense to include both F and PH on the keyboard when building PHONE. If the child chooses to use F, the scaffolding system should accept this as an instance of invented spelling.

But what should the keyboard look like in the open-ended mode? Since there was no evidence suggesting an advantage of phoneme blocks, it seems preferable for a designer to proceed with a conventional letter keyboard. Such a keyboard fits better into the contexts of the classroom and broader culture. Issues related to the orthographical challenges of English language could be addressed by supporting invented spelling (although further research is needed on this matter). Section 6.5.9 gives an example of a child struggling to locate a block for *[ʃ]* on the letter keyboard. To avoid such issues, it is likely desirable to incorporate such graphemes as TH and SH into the keyboard in order to represent every consonant phoneme. The keyboard may use mnemonics, but in that case it is preferable to allow the child to switch into conventional letter mode.

To enable smooth transition from scaffolded mode to open-ended mode, it appears desirable to base the layout of the "scaffolded" keyboard on the "open-ended" one - e.g. by disabling and

greying out unused keys when moving to scaffolded mode. Since multi-letter graphemes are often needed, the scaffolded keyboard should include extra space for spawning additional keys for them.

7.3. Suggestions for an Educator

This section gives some suggestions to educators who are interested in using child-driven expressive early literacy media in a classroom. As in the case with a designer, many suggestions for an educator follow from the body of research in which the present work is situated. Rather than repeating them, I would like to refer the reader to these works. A book on emerging literacy edited by Strickland and Morrow (1989) offers a great list of specific techniques to support emerging literacy in the classrooms. Many of these ideas, such as “class newspaper” (page 22), “letter to a guest” (page 24), “me museum” (page 54), and “mailbox” (page 133), can be adapted for use with expressive media. Further, I highly recommend Brennan’s (2013) PhD dissertation, which discusses how an educator can set up a structure that supports, rather than inhibits, the child’s agency. A teacher accustomed to the instructionist model might be inclined to seize initiative and organize children’s activities around teacher-suggested ideas and themes. However, we saw that children are very capable of generating creative ideas of their own, while externally imposed ideas may lead to their disengagement (see “Mad Libs” game in section 5.4.1). Brennan discusses more productive directions for the teacher’s efforts in a child-driven setting. The idea of a teacher being a guide responding to the child’s lead is also essential to the philosophies of Montessori and Reggio Emilia. I recommend Lillard’s (1972) book as an introduction to the former, and Edwards et al.’s, (1998) book for a flavor of the latter. Finally, Resnick’s (2017) book is a great presentation of a child-driven, expressive approach to learning that gives specific suggestions to educators. Below are a few additional suggestions that follow from the experience with the present studies.

1. Arrange children in a way that is conducive for social play. In our studies, the small groups arrangement (with 4-5 students at a table) seemed to be very well suited for that purpose, evoking various types of play described in sections 5.5 and 6.4.

2. Create conditions that help children focus. Expressive play requires concentration. The more distractions the classroom has, the more likely children are to go into the unsophisticated, impulsive mode of play. One of the strongest distractions is noise. Particularly problematic is a scenario when scaffolding systems on multiple devices talk over each other: it becomes difficult for children to distinguish which prompts are directed at them and which ones are for their peers. For these reasons, I recommend **headphones**. In the SpeechBlocks II study, after the introduction of headphones, children became visibly more focused. Contrary to the original concerns, headphones still allowed for flourishing social interactions. A side benefit of headphones is that they can carry a microphone, which greatly increases performance of speech recognition in a noisy environment. In addition to noise, children were often distracted by the sight of their peers doing something odd somewhere else in the classroom. To counter that, a station for expressive play could be located in a secluded corner of the room. Finally, such approaches as Montessori include dedicated exercises that help children develop their capacity for focused work (P. P. Lillard, 1972).

3. Give children sufficient time to play. In several different classrooms, I witnessed teachers advocating for quick changes of activity out of concern about the attention span of 4 to 5 year-olds. However, in the case of child-driven activities, children can remain engaged for remarkably long periods of time. Sections 5.4.2 and 6.2.3 show that children could enthusiastically play with SpeechBlocks for 30 minutes or more in a row. Conversely, I saw that too short play sessions (e.g. 10 minutes) limited children's ability to engage in sophisticated activities. As a result, their behavior appeared more restless, unfocused and chaotic. They also expressed frustration that they couldn't finish the words they planned to make. Therefore, I recommend allocating at least 15-20 minutes for play sessions with expressive media. An even better option could be allowing children to set up their play time on their own, as in Montessori classrooms.

7.4. Future Directions

At the end of five years of working with SpeechBlocks, I wonder if research work ever seems complete to a researcher who conducts it. With each new study, new questions and possible directions emerged, usually more numerous than what existed prior to it. Below, I suggest some future directions that directly follow from the present work and that I find promising or important. They are divided into two categories: academic and design directions.

Academic Directions:

1. Moving from exploration to rigorous quantitative studies. I remind the reader that the present work is exploratory in its nature, and quantitative rigour was not its focus. For instance, while SpeechBlocks II study suggests some quantitative patterns, factors such as small sample size, presence of confounding variables (such as possible effects of classroom and teacher), imperfect administration of pre- and post-tests (doing it group-by-group instead of calling children in random order, which might have introduced effects of the testing environment), modifications of the design mid-way through the studies (to incorporate the learnings) and presence of researchers in the classrooms all might have skewed its results. In addition, the baseline chosen for the study — independent literacy activities with traditional materials — was not particularly strong. It would be interesting to look at other baselines, such as state-of-the-art instructionist digital learning media.

2. Evaluate SpeechBlocks II with an older age range. We saw that children with higher PA were more likely to engage in imaginative play and less likely to regress to unproductive behaviors. But by the age of 5 to 6, the majority of children should be in the zone that was considered high for our sample, but their phonological awareness still has sufficient room to grow. For these reasons, this age range may be more optimal for usage of SpeechBlocks II. Earlier, I mentioned a possibility that the apparent pattern of children with higher PA and EF benefiting more from the app may be related to the app being too difficult for the chosen age range. In connection with that, it would be interesting to see whether this pattern still holds for older children. Furthermore, some qualitatively

new behaviors might emerge. In particular, children of this age range might exhibit more interest in using the invented spelling interpreter.

3. Assess engagement with SpeechBlocks II in home conditions. For SpeechBlocks I, we saw a rapid drop in children's engagement over time. However, I hypothesized that this was because of the "high floor" (difficulty to spell meaningful words) and "low ceiling" (limited expressive capacity) of the medium. In SpeechBlocks II, both of these issues have been to some extent addressed. It would be very interesting to see whether this was sufficient to change the usage dynamics. An encouraging bit of information is that PictureBlocks (Makini, 2018), which inspired SpeechBlocks II, exhibited relatively good retention of children's engagement over time.

4. Investigate potential applications of expressive media for addressing Summer Reading Loss. The primary causes of Summer Reading Loss lie outside of the school environment, and this is where high impact interventions should be directed. Since SpeechBlocks fits well with informal learning environments, the app may offer some help in solving this problem. Possible interventions could use SpeechBlocks in homes or at afterschool programs.

5. Investigate potential applications of expressive media as a tool of Speech-Language Pathologist, for children with delays in phonological awareness development. In conversation with several language and learning professionals, I heard suggestions regarding potential usefulness of SpeechBlocks for helping such children see how word segmentation and blending work, as well as the functioning of various spelling patterns. In such a scenario, an expressive medium would likely be used not independently by a child, as it was designed for, but jointly with a literacy specialist.

6. Investigate which factors affect the usefulness of sound mnemonics. The data from the present work suggests that "sound creatures" helped some children find sounds on the keyboard, but were not useful for others. However, it doesn't identify the factors that determined this difference. For example, could children's letter-to-sound knowledge be one of such factors? It is also interesting to see whether the results were affected by the specifics of the present approach: using onomatopoeic mnemonics instead of rebus principle, and using a child-driven setup instead of a teacher-driven one. Previous works, situated in a teacher-driven context and utilizing the rebus principle, reported learning gains in letter-sound identification for treatment children. We haven't seen it in the case of the current work; the reasons for this would be interesting to investigate.

7. Investigate possible gender differences in using the media. In the quantitative analysis of PA gains in SpeechBlocks II study, we saw a trend suggesting that boys may have benefitted more than girls. This trend was not particularly strong, and analysis of qualitative observations didn't allow me to convincingly explain it. It is possible that the trend was just a fluctuation in the data. Nevertheless, possible gender differences is something to watch for in future studies.

8. Collect a database of invented spellings with associated interpretations. To my knowledge, although there are many examples of invented spelling scattered throughout literature, no centralized database currently exists. Such a database could be very useful in developing tools of automatic interpretation of invented spelling. It could also be useful from a purely research perspective, offering opportunities for computational analysis of invented spelling. As an initial step, about 1100 examples can be extracted from the SpeechBlocks I logs pertaining to home studies. However, they are biased towards more sophisticated spellings, since simple invented spellings were often indistinguishable from nonsense words.

9. Explore the potential of using a synthesizer that can emphasize phonemes on demand. Current scaffolding system relies on emphasizing phonemes in various positions of a word — e.g. "batMMMMan". Unfortunately, no current synthesizer is capable of such emphasis, so SpeechBlocks had to resort to a clumsy bypass by combining the synthesizer output with a voice recording for the target sound. Developing a synthesizer that is capable of emphasizing individual sounds may help children to parse the sound structures of various words.

Design Directions:

1. Explore adaptive scaffolding. As it was mentioned earlier, a major downside of the current scaffolding system is that it doesn't adjust to the current skill level of the child. This limits its capacity to tap into the child's Zone of Proximal Development. The current design of scaffolded mode readily allows for stratifying into difficulty levels. At the lowest difficulty level, it can focus on filling the slot for initial sound only. As the difficulty increases, the system can let the child fill the final sound as well, then proceed to medials — first consonants, then vowels. Complex and unusual letter-to-sound patterns can be targeted last. Simultaneously, the number of blocks on the scaffolded keyboard can increase, until it becomes similar to the full keyboard. Therefore, it is possible to make a smooth ladder from a relatively simple initial experience all the way to the open-ended mode, while remaining true to the principles of the present approach.

2. Try a more conversational approach to scaffolding. For instance, instead of automatically correcting the child, the system might highlight the outcome of wrong choices, e.g.: "This block says RRR, so that would be a RATMAN instead of a BATMAN." Such a system would allow the child to either correct the mistake, or not and instead explore fun words that emerge. That may lead to a more lighthearted, playful, and exploratory way of interaction. Importantly, it may bring some nonsense-word-related humour into scaffolded word building. In the first SpeechBlocks study, we saw the beneficial effects of such humour on children's engagement (section 5.4.1). Furthermore, a conversational approach can help draw children's attention to their mistakes and ways to correct them, an importance of which in literacy learning software was shown by Kegel & Bus (2012). A conversational system might employ a virtual character. We saw the potential of such an approach in children relating to a simple character like Mr. Fox (section 6.5.8).

3. Explore further ways to support invented spelling. While incorporating the invented spelling interpreter into SpeechBlocks II wasn't a success, supporting invented spelling still seems

promising, since it provides unique opportunities for tapping into the child's ZPD. One way to improve the system's performance in interpreting invented spelling is taking into account the context of the scene that is being constructed and the words that the child recently said. Section 6.5.4 provides an example from an actual child's play where such a capacity would have been helpful.

4. Explore how to combine ease of remixing words with ease of sequential word construction. Play mechanics of SpeechBlocks I allowed for easy remixing of words. It led to a distinct play type, referred to as Remixing and Rhyming, which was particularly important in the early period of children's interaction with SpeechBlocks. The Word Box mechanics of SpeechBlocks II makes such interactions difficult, but simplifies construction of words block-by-block. Can the advantages of both designs be combined?

5. Explore object recognition as a word source. In section 6.5.7, I mentioned how it may be better suited to the needs of young children than text recognition. Tinkering with off-the-shelf object recognition technology suggests that it may already be useful for that purpose.

6. Examine reading-oriented applications of text recognition. We saw that text recognition prompted children to explore environmental texts and books. However, its usage was often disconnected from expressive activities. Perhaps text recognition could be a more natural fit to learning designs that are oriented not on production, but on consumption of texts.

7. Explore additional ways to facilitate search for ideas. From the experience with the association network, we see that children derive value from such facilitation. However, the association network doesn't help to "break the blank canvas". The relevance of its suggestions to children's creations is also not perfect. One interesting direction would be to incorporate some inspirational examples within the medium — e.g. a digital version of the Richard Scarry's scenes — and make it possible for children to source words from them. A more sophisticated approach could involve analyzing children's scenes to make relevant suggestions. It might even be beneficial for the system to involve elements of computational co-creativity — e.g. to be a partner to the child in constructing scenes.

8. Explore possible synergies between expressive and instructional approaches. For instance, television might be a good medium to present the onomatopoeic creatures as rich, lifelike personas. That might help children to establish a better connection with them, and consequently simplify their usage of SpeechBlocks. Existing programs, such as Lively Letters or Alphablocks, might be well-suited for this role. It is also interesting whether a gamified instructional experience might serve as an entry point for children with low PA and EF. As they gain skills with such a game, they may become better prepared to successfully engage with the challenging open-ended design of SpeechBlocks.

7.5. Final Thoughts

I would like to finish this work with some personal thoughts. Throughout the process of writing, I was inspired and supported by the works of two great masters, creating in different media and in different cultures, yet connected through their art: Ursula K. Le Guin and Hayao Miyazaki. They helped me think about what role creative expression plays in being human. It is a mysterious facet of our lives that is not always easy to quantify. In its elusive subtleness, it often gets neglected in favor of matters that are easier to put our finger on, such as material comfort. It is towards these economic ends that the "industrial model of schooling" has been developed (Holt, 1989; A. S. Lillard, 2016). In a world where so many lack basic necessities, we cannot ignore people's material needs. Yet we should not pursue them so obsessively that they define us. I am often concerned that in the endless pursuit of material growth, global civilization will soon hit a limit, beyond which lie great perils for our collective well-being. This is something that some MIT researchers anticipated as early as in the 1970s (Meadows et al., 1972). If this occurs, major cultural shifts would be inevitable. We as a species would have to stop seeing ourselves as producers and consumers and think anew who we really are. Perhaps a different form of education, one that prioritizes all-round unfolding of human potential, can play a key role in this process of self-discovery. I am reminded of Maria Montessori, who asserted that humans possess "an intense drive for self-actualization." She saw human beings not as economic units, but as "spiritual seeds" to be cultivated and nurtured. It is up to us to sprout.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., Devin, M., Ghemawat, S., Irving, G., Isard, M., & Others. (2016). Tensorflow: A system for large-scale machine learning. *12th Symposium on Operating Systems Design and Implementation*, 265–283.

Ali, S. A. (2019). *Designing Child-Robot Interaction for Facilitating Creative Learning.* (Master's Thesis, Massachusetts Institute of Technology)

Allington, R. L., & McGill-Franzen, A. (2003). The impact of summer setback on the reading achievement gap. *Phi Delta Kappan*, *85*(1), 68-75.

Anderson, D. (2006). *A reciprocal determinism analysis of the relationship between naturalistic media usage and the development of creative-thinking skills among college students.* (Doctoral dissertation, Northern Illinois University.)

Anderson, J. R., Corbett, A. T., Koedinger, K. R., & Pelletier, R. (1995). Cognitive Tutors: Lessons Learned. *Journal of the Learning Sciences*, *4*(2), 167–207.

Anderson, J. R., & Gluck, K. (2001). What role do cognitive architectures play in intelligent tutoring systems. *Cognition & Instruction: Twenty-Five Years of Progress*, 227–262.

Baldwin, T., & Tanaka, H. (2000). A comparative study of unsupervised grapheme-phoneme alignment methods. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *22*.

Barab, S., & Squire, K. (2004). Design-Based Research: Putting a Stake in the Ground. *Journal of the Learning Sciences*, *13*(1), 1–14.

Baratta-Lorton, R. (1985). *Dekodiphukan.*

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, *68*(3).

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). *lme4: Linear mixed-effects models using Eigen and S4. R package version 1.1--7. 2014*.

Beck, I. L., & Beck, M. E. (2013). *Making sense of phonics: The hows and whys.* Guilford Publications.

Bers, M. U., & Resnick, M. (2015). *The Official ScratchJr Book: Help Your Kids Learn to Code*. No Starch Press.

Bickford, T. (2017). *Schooling New Media: Music, Language, and Technology in Children's Culture*. Oxford University Press.

Bissex, G. L. (1980). *Gnys at Wrk: A Child Learns to Write and Read*. Harvard University Press.

Black, A. W., Lenzo, K., & Pagel, V. (1998). *Issues in building general letter to sound rules*. In *The Third ESCA/COCOSDA Workshop (ETRW) on Speech Synthesis.*

Bodrova, E., & Leong, D. J. (2001). Tools of the Mind: A Case Study of Implementing the Vygotskian Approach in American Early Childhood and Primary Classrooms. *Innodata Monographs*, *7*.

Bodrova, E., & Leong, D. J. (2006). *Tools of the Mind: Vygotskian Approach to Early Education*. Pearson Australia Pty Limited.

Brennan, K. A. (2013). *Best of both worlds : issues of structure and agency in computational creation, in and out of school* (Doctoral dissertation, Massachusetts Institute of Technology)

Bürkner, P.-C., & Others. (2017). brms: An R package for Bayesian multilevel models using Stan. *Journal of Statistical Software*, *80*(1), 1–28.

Bus, A. G., Takacs, Z. K., & Kegel, C. A. T. (2015). Affordances and limitations of electronic storybooks for young children's emergent literacy. *Developmental Review: DR*, *35*, 79–97.

Castles, A., Rastle, K., & Nation, K. (2018). Ending the Reading Wars: Reading Acquisition From Novice to Expert. *Psychological Science in the Public Interest: A Journal of the American*

Psychological Society, *19*(1), 5–51.

Chiong, C., & Shuler, C. (2010). Learning: Is there an app for that. In *Investigations of young children's usage and learning with mobile devices and apps.* New York: The Joan Ganz Cooney Center at Sesame Workshop (pp. 13-20).

Clouder, C., & Rawson, M. (1998). *Waldorf education*. Rudolph Steiner Press.

Cooper, H., Nye, B., Charlton, K., Lindsay, J., & Greathouse, S. (1996). The effects of summer vacation on achievement test scores: A narrative and meta-analytic review. *Review of educational research*, *66*(3), 227-268.

Cossu, G., Gugliotta, M., & Marshall, J. C. (1995). Acquisition of reading and written spelling in a transparent orthography: Two non parallel processes? *Reading and Writing*, *7*(1), 9–22.

Creutz, M., & Lagus, K. (2002). Unsupervised Discovery of Morphemes. *Proceedings of the ACL-02 Workshop on Morphological and Phonological Learning - Volume 6*, 21–30.

Csikszentmihalyi, M. (1997). *Flow and the psychology of discovery and invention.* HarperPerennial, New York, 39.

Cunningham, A. E., & Stanovich, K. E. (1998). What reading does for the mind. *American educator*, 22, 8-17.

Damper, R. I., Marchand, Y., Marseters, J.-D., & Bazin, A. (2004). Aligning letters and phonemes for speech synthesis. *Fifth ISCA Workshop on Speech Synthesis*.

Davidse, N. J., de Jong, M. T., Bus, A. G., Huijbregts, S. C. J., & Swaab, H. (2011). Cognitive and environmental predictors of early literacy skills. *Reading and Writing*, *24*(4), 395–412.

Davis, N. M., Popova, Y., Sysoev, I., Hsiao, C. P., Zhang, D., & Magerko, B. (2014). Building Artistic Computer Colleagues with an Enactive Model of Creativity. In *ICCC* (pp. 38-45).

Davis, N. M. (2017). *Creative sense-making: A cognitive framework for quantifying interaction dynamics in co-creation* (Doctoral dissertation, Georgia Institute of Technology).

De Graaff, S., Verhoeven, L., Bosman, A. M., & Hasselman, F. (2007). Integrated pictorial mnemonics and stimulus fading: Teaching kindergartners letter sounds. *British Journal of Educational Psychology*, *77*(3), 519-539.

De Marcken, C. (1996). Unsupervised Language Acquisition. (Doctoral dissertation, Massachusetts Institute of Technology)

DiLorenzo, K. E., Rody, C. A., Bucholz, J. L., & Brady, M. P. (2011). Teaching letter–sound connections with picture mnemonics: Itchy's alphabet and early decoding. *Preventing School Failure: Alternative Education for Children and Youth*, *55*(1), 28-34.

Edwards, C. P., Gandini, L., & Forman, G. E. (1998). *The Hundred Languages of Children: The Reggio Emilia Approach - Advanced Reflections*. Greenwood Publishing Group.

Ehri, L. C., Deffner, N. D., & Wilce, L. S. (1984). Pictorial mnemonics for phonics. *Journal of educational psychology*, *76*(5), 880.

Ehri, L. C. (2005). Learning to Read Words: Theory, Findings, and Issues. *Scientific Studies of Reading*, *9*(2), 167–188.

Falbel, A. (1985). A Second Look at Writing to Read: A Teaching System for Schools Becomes A Medium for Learning in the Home. An MIT Media Laboratory internal report.

Fischler, M. A., & Bolles, R. C. (1981). Random Sample Consensus: A Paradigm for Model Fitting with Applications to Image Analysis and Automated Cartography. *Communications of the ACM*, *24*(6), 381–395.

Fisch, S. M., & Truglio, R. T. (2014). *G Is for Growing: Thirty Years of Research on Children and Sesame Street*. Routledge.

Franc, B., & Subotic, V. (2015). Differences in phonological awareness of five-year-olds from Montessori and regular program preschool institutions. *Researching Paradigms of Childhood and Education Conference Book of Selected Papers (2nd Symposium: Child Language and*

Culture), 12–20.

Frank, M. (2018, February 26). *Mixed effects models: Is it time to go Bayesian by default?* URL: http://babieslearninglanguage.blogspot.com/2018/02/mixed-effects-models-is-it-time-to-go.html Retrieved: April 30th, 2020.

Futrell, R., Albright, A., Graff, P., & O'Donnell, T. J. (2017). A Generative Model of Phonotactics. *Transactions of the Association for Computational Linguistics*, *5*, 73–86.

Gee, J. P. (2007). *What Video Games Have to Teach Us About Learning and Literacy.* Palgrave Macmillan; 2nd edition

Georgiou, G. K., Parrila, R., & Papadopoulos, T. C. (2008). Predictors of word decoding and reading fluency across languages varying in orthographic consistency. *Journal of Educational Psychology*, *100*(3), 566.

Geudens, A., & Sandra, D. (2003). Beyond implicit phonological knowledge: No support for an onset--rime structure in children's explicit phonological awareness. *Journal of Memory and Language*, *49*(2), 157–182.

Goodwyn, A. (2014). Reading is now "cool": a study of English teachers' perspectives on e-reading devices as a challenge and an opportunity. *Educational Review*, *66*(3), 263–275.

Gordon, G., & Breazeal, C. (2015). Bayesian active learning-based robot tutor for children's word-reading skills. *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Guernsey, L., & Levine, M. H. (2015). *Tap, Click, Read: Growing Readers in a World of Screens*. John Wiley & Sons.

Hayes, D. P., & Grether, J. (1983). The school year and vacations: When do students learn?. *Cornell Journal of Social Relations.*

Hershman, A., Nazare, J., Sysoev, I., Fratamico, L., Buitrago, J., Soltangheis, M., Makini, S., Chu, E., & Roy, D. (2018). Family learning coach: engaging families in children's early literacy

learning with computer-supported tools. In *Proceedings of the 26th International Conference on Computers in Education.* Philippines: Asia-Pacific Society for Computers in Education (pp. 637-646)

Hill, F., Bordes, A., Chopra, S., & Weston, J. (2015). The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations. *arXiv preprint.* http://arxiv.org/abs/1511.02301

Hinton, G. E., Srivastava, N., Krizhevsky, A., Sutskever, I., & Salakhutdinov, R. R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. *arXiv preprint.* http://arxiv.org/abs/1207.0580

Hirsh-Pasek, K., Golinkoff, R. M., Berk, L. E., & Singer, D. (2009). *A mandate for playful learning in preschool: Applying the scientific evidence*. Oxford University Press.

Hirsh-Pasek, K., Zosh, J. M., Golinkoff, R. M., Gray, J. H., Robb, M. B., & Kaufman, J. (2015). Putting Education in "Educational" Apps: Lessons From the Science of Learning. *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, *16*(1), 3–34.

Hochreiter, S., & Schmidhuber, J. (1997). Long short-term memory. *Neural Computation*, *9*(8), 1735–1780.

Holt, J. C. (1989). *Learning all the time*. Addison-Wesley New York.

Hoover, W. A., & Gough, P. B. (1990). The simple view of reading. *Reading and writing*, *2*(2), 127-160.

Huizinga, J. (1949). *Homo Ludens: A Study of the Play-Element in Culture.* London: Routledge & Kegan Paul.

Hutter, M. (2003). A gentle introduction to the universal algorithmic agent AIXI. In *Artificial General Intelligence*. Springer.

Ingraham, C. (2018). Leisure reading in the US is at an All-Time low. *The Washington Post.* URL:

https://www.washingtonpost.com/news/wonk/wp/2018/06/29/leisure-reading-in-the-u-s-is-at-an-all-time-low/ Retrieved: April 30th, 2020.

Kelly, A., & Safford, K. (2009). Does teaching complex sentences have to be complicated? Lessons from children's online writing. *Literacy*, *43*(3), 118-122.

Jacovina, M. E., & McNamara, D. S. (2017). Intelligent Tutoring Systems for Literacy: Existing Technologies and Continuing Challenges. *Grantee Submission*. https://eric.ed.gov/?id=ED577131

Jones, A., & Castellano, G. (2018). Adaptive robotic tutors that support self-regulated learning : A longer-term investigation with primary school children. *International Journal of Social Robotics*, *10*(3), 357–370.

Jones, J. K. (1968). Comparing ITA with colour story reading. *Educational Research*, *10*(3), 226–234.

Kegel, C. A. T., van der Kooy-Hofland, V. A. C., & Bus, A. G. (2009). Improving early phoneme skills with a computer program: Differential effects of regulatory skills. *Learning and Individual Differences*, *19*(4), 549–554.

Kegel, C. A. T., & Bus, A. G. (2012). Online tutoring as a pivotal quality of web-based early literacy programs. *Journal of Educational Psychology*, *104*(1), 182.

Kingma, D. P., & Ba, J. (2014). Adam: A Method for Stochastic Optimization. In *arXiv [cs.LG]*. arXiv. http://arxiv.org/abs/1412.6980

Kothari, B., & Bandyopadhyay, T. (2014). Same language subtitling of Bollywood film songs on TV: Effects on literacy. *Information Technologies & International Development*, 10(4), pp-31.

Kory-Westlund, J. (2019). *Relational AI: Creating Long-Term Interpersonal Interaction, Rapport, and Relationships with Social Robots* (Doctoral dissertation, Massachusetts Institute of Technology)

Kory-Westlund, J. M., & Breazeal, C. (2019). Exploring the Effects of a Social Robot's Speech

Entrainment and Backstory on Young Children's Emotion, Rapport, Relationship, and Learning. *Frontiers in Robotics and AI*, *6*, 54.

Lawrence, S. G. C., & Kaye, G. (1986). Alignment of phonemes with their corresponding orthography. *Computer Speech & Language*, *1*(2), 153–165.

Lawson, G. M., Hook, C. J., & Farah, M. J. (2018). A meta-analysis of the relationship between socioeconomic status and executive function performance among children. *Developmental Science*, *21*(2), e12529.

Lillard, A. S. (2016). *Montessori: The Science Behind the Genius*. Oxford University Press.

Lillard, P. P. (1972). *Montessori: A Modern Approach*. Schocken Books.

Ling, C. X., & Wang, H. (1997). Alignment algorithms for learning to read aloud. *IJCAI (2)*, 874–879.

Lukeš, D., & Litsas, C. (2015). Building a phonics engine for automated text guidance. *2015 6th International Conference on Information, Intelligence, Systems and Applications (IISA)*, 1–6.

Luk, R. W. P., & Damper, R. I. (1992). Inference of letter-phoneme correspondences by delimiting and dynamic time warping techniques. *[Proceedings] ICASSP-92: 1992 IEEE International Conference on Acoustics, Speech, and Signal Processing*, *2*, 61–64 vol.2.

Makini, S. P. (2018). *PictureBlocks : constructing and deconstructing picture-driven literacy development* (Master thesis, Massachusetts Institute of Technology)

Meadows, D. H., Meadows, D. L., Randers, J., & Behrens, W. W. (1972). The limits to growth. *New York*, *102*, 27.

Meline, C. W. (1976). Does the medium matter? *Journal of Communication*.

Mesgarani, N., David, S. V., Fritz, J. B., & Shamma, S. A. (2008). Phoneme representation and classification in primary auditory cortex. *The Journal of the Acoustical Society of America*, *123*(2), 899–909.

Mesgarani, N., Cheung, C., Johnson, K., & Chang, E. F. (2014). Phonetic feature encoding in

human superior temporal gyrus. *Science*, *343*(6174), 1006–1010.

Miller, B. L., Goldberg, D. E., & Others. (1995). Genetic algorithms, tournament selection, and the effects of noise. *Complex Systems*, *9*(3), 193–212.

Moats, L. C. (1999). *Teaching reading is rocket science: What expert teachers of reading should know and be able to do.* American Federation of Teachers.

Moore, O. K. (1966). Autotelic responsive environments and exceptional children. In *Experience, Structure & Adaptability.* Springer, 169–216.

Morey, R. D., Hoekstra, R., Rouder, J. N., Lee, M. D., & Wagenmakers, E.-J. (2016). The fallacy of placing confidence in confidence intervals. *Psychonomic Bulletin & Review*, *23*(1), 103–123.

Nazare, J., Hershman, A., Sysoev, I., Fratamico, L., Buitrago, J., Soltangheis, M., Makini, S. P., Chu, E., & Roy, D. (2018). In *Proceedings of the 13th International Conference of Learning Sciences.* London, United Kingdom, pp. 1409-1410.

Nicolopoulou, A., Barbosa de Sa, A., Ilgaz, H., & Brockmeyer, C. (2009). Using the transformative power of play to educate hearts and minds: From Vygotsky to Vivian Paley and beyond. *Mind, culture, and activity*, *17*(1), 42-58.

Novak, J. R., Yang, D., Minematsu, N., & Hirose, K. (2011). *Phonetisaurus: A WFST-driven phoneticizer.* URL: *https://github.com/AdolfVonKleist/Phonetisaurus*.

Ouellette, G., Sénéchal, M., & Haley, A. (2013). Guiding Children's Invented Spellings: A Gateway Into Literacy Learning. In *The Journal of Experimental Education* (Vol. 81, Issue 2, pp. 261–279).

Papert, S. (1980). *Mindstorms: Children, computers, and powerful ideas*. New York: Basic Book

Pratham (2019). *Annual Status of Education Report (Rural) 2018, Provisional*.

Rao, K., Peng, F., Sak, H., & Beaufays, F. (2015). Grapheme-to-phoneme conversion using Long Short-Term Memory recurrent neural networks. *2015 IEEE International Conference on*

Acoustics, Speech and Signal Processing (ICASSP), 4225–4229.

Read, C. (1971). Pre-School Children's Knowledge of English Phonology. *Harvard Educational Review*, *41*(1), 1–34.

Reardon, S. F. (2011). The widening academic achievement gap between the rich and the poor: New evidence and possible explanations. *Whither opportunity*, *1*(1), 91-116.

Reinhart, A. (2015). *Statistics Done Wrong: The Woefully Complete Guide*. No Starch Press.

Resnick, M. (2017). *Lifelong Kindergarten: Cultivating Creativity Through Projects, Passion, Peers, and Play*. MIT Press.

Resnick, M., & Rosenbaum, E. (2013). Designing for tinkerability. *Design, Make, Play: Growing the next Generation of STEM Innovators*, 163–181.

Resnick, M., & Silverman, B. (2005). Some reflections on designing construction kits for kids. *Proceedings of the 2005 Conference on Interaction Design and Children*, 117–122.

Rhyner, P. M. (2009). *Emergent Literacy and Language Development: Promoting Learning in Early Childhood*. Guilford Press.

Richgels, D. J. (2001). Invented spelling, phonemic awareness, and reading and writing instruction. *Handbook of Early Literacy Research*, *1*, 142–155.

Ritchie, S. J., Luciano, M., Hansell, N. K., Wright, M. J., & Bates, T. C. (2013). The relationship of reading ability to creativity: Positive, not negative associations. *Learning and Individual Differences*, *26*, 171-176.

Ritchie, S. J., Bates, T. C., & Plomin, R. (2015). Does learning to read improve intelligence? A longitudinal multivariate analysis in identical twins from age 7 to 16. *Child development*, *86*(1), 23-36.

Roberts, T. A., & Sadler, C. D. (2019). Letter sound characters and imaginary narratives: Can they enhance motivation and letter sound learning?. *Early Childhood Research Quarterly*, *46*,

97-111.

Roser, M., & Ortiz-Ospina, E. (2016). Literacy. *Our World in Data*. URL: https://ourworldindata.org/literacy?_ke=eyJrbF9lbWFpbCI6ICJweGdqZ2htaHdmb0BtaWxsZWRtYWlsLmNvbSIsICJrbF9jb21wYW55X2lkIjogIm5GRWUzUiJ9 Retrieved: Oct. 30th, 2020.

Roy, D. K., & Pentland, A. P. (2002). Learning words from sights and sounds: a computational model. *Cognitive Science*, *26*(1), 113–146.

Rubin, K. H., Fein, G. G., & Vandenberg, B. (1983) Play. *Handbook of child psychology*, *4*, 693-774.

Sandel, L. (1998). *The Nature of Traditional Orthography and the Initial Teaching Alphabet. Review of Historical Research: Summary# 4.*

Shmidman, A., & Ehri, L. (2010). Embedded picture mnemonics to learn letters. *Scientific studies of reading*, 14(2), 159-182.

Segers, E., Damhuis, C. M. P., van de Sande, E., & Verhoeven, L. (2016). Role of executive functioning and home environment in early reading development. *Learning and Individual Differences*, *49*, 251–259.

Sejnowski, T. J., & Rosenberg, C. R. (1987). NET talk: A parallel network that learns to read aloud. *Complex Systems*, *1*, 145–168.

Seymour, P. H., Aro, M., Erskine, J. M., in collaboration with COST Action A8 Network. (2003). Foundation literacy acquisition in European orthographies. *British Journal of psychology*, *94*(2), 143-174.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, *27*(3), 379–423.

Shi, J., & Tomasi, C. (1993). *Good features to track*. In 1994 Proceedings of IEEE conference on computer vision and pattern recognition (pp. 593-600). IEEE.

Shuler, C. (2009). *Pockets of potential: Using mobile technologies to promote children's learning.* Joan Ganz Cooney Center at Sesame Workshop.

Siegenfeld, A. F., & Bar-Yam, Y. (2019). An Introduction to Complex Systems Science and its Applications. In *arXiv [physics.soc-ph]*. arXiv. http://arxiv.org/abs/1912.05088

Smith R. A. (2003). *LeapFrog: The Letter Factory.* [Animated Film]

Soltangheis, M. (2017). *From children's play to intentions: a play analytics framework for constructionist learning apps.* (Master's thesis, Massachusetts Institute of Technology).

Strickland, D. S., & Morrow, L. M. (1989). *Emerging literacy: young children learn to read and write*. International Reading Association.

Stroop, J. R. (1935). Studies of interference in serial verbal reactions. *Journal of Experimental Psychology*, *18*(6), 643.

Sutton-Smith, B. (2009). *The Ambiguity of Play*. Harvard University Press.

Sysoev, I., Hershman, A., Fine, S., Traweek, C., & Roy, D. (2017). SpeechBlocks: A Constructionist Early Literacy App. In *Proceedings of the 2017 Conference on Interaction Design and Children*, 248–257.

Taylor, B. Y. K., & Silver, L. (2019). Smartphone Ownership Is Growing Rapidly Around the World, but Not Always Equally. *URL:* https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/. Retrieved: April 30, 2020.

Tenenbaum, J. B., Kemp, C., Griffiths, T. L., & Goodman, N. D. (2011). How to grow a mind: statistics, structure, and abstraction. *Science*, *331*(6022), 1279–1285.

Torgashov, P. (2014, June 8). *Contour analysis for image recognition in C#*. URL: https://www.codeproject.com/Articles/196168/Contour-Analysis-for-Image-Recognition-in-C Retrieved April 30th, 2020.

Toshniwal, S., & Livescu, K. (2016). Jointly learning to align and convert graphemes to phonemes with neural attention models. *2016 IEEE Spoken Language Technology Workshop (SLT)*, 76–82.

Tsur, M., & Rusk, N. (2018, February). Scratch microworlds: designing project-based introductions to coding. In *Proceedings of the 49th ACM Technical Symposium on Computer Science Education* (pp. 894-899).

Vaala, S., Ly, A., & Levine, M. H. (2015). *Getting a Read on the App Stores: A Market Scan and Analysis of Children's Literacy Apps. Full Report.* Joan Ganz Cooney Center at Sesame Workshop.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data: Springer Series in Statistics (Springer Series in Statistics)*. Springer-Verlag.

Vasicek, O. (1976). A Test for Normality Based on Sample Entropy. *Journal of the Royal Statistical Society. Series B, Statistical Methodology*, *38*(1), 54–59.

Vasilyeva, M., & Waterfall, H. (2011). Variability in language development: Relation to socioeconomic status and environmental input. *Handbook of early literacy research*, *3*, 36-48.

Vernon-Feagans, L., Hammer, C. S., Miccio, A., & Manlove, E. (2001). Early language and literacy skills in low-income African American and Hispanic children. *Handbook of early literacy research*, *1*, 192-210.

Vintsyuk, T. K. (1968). Speech discrimination by dynamic programming. *Cybernetics and Systems Analysis*, *4*(1), 52–57.

Vygotsky, L. (1978). Interaction between learning and development. *Readings on the Development of Children*, *23*(3), 34–41.

Wagner, R. A., & Fischer, M. J. (1974). The String-to-String Correction Problem. *Journal of the ACM*, *21*(1), 168–173.

Weide, R. L. (1998). The CMU pronouncing dictionary. *URL: http://www.speech.cs.cmu.edu/cgi-bin/cmudict* Retrieved: April 30th, 2020.

Wolf, M. (2008). *Proust and the squid: The story and science of the reading brain*. Harper Perennial New York.

Wood, D., Bruner, J. S., & Ross, G. (1976). The role of tutoring in problem solving. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *17*(2), 89–100.

Wood, D., & Wood, H. (1996). Vygotsky, Tutoring and Learning. *Oxford Review of Education*, *22*(1), 5–16.

Wright, A., & Diamond, A. (2014). An effect of inhibitory load in children while keeping working memory load constant. *Frontiers in Psychology*, *5*, 213.

Appendix A. Glossary

This appendix provides a reference for some frequently used terms.

Atom (of writing). In context of this work: a "building block" with aligned pronunciation and spelling, out of which words are constructed. Typically, it consists of a phoneme and a grapheme that encodes it.

CTOPP-2. Comprehensive Test of Phonological Processing. A commonly used test to measure phonological and related skills of people aged 4 to 24, including phonological awareness. The phonological awareness component includes three subtests. The first two are **elision** (removing a sound from a word) and **blending** (combining sounds). The third one differs depending on the age. It is **sound matching** (identifying initial or final sound of a word) for the children 4 to 6, or **phoneme isolation** (identifying initial, 2nd, 3rd, etc. sounds in a word) for older ones. The scores for each subtest can be converted to **scaled scores**, adjusted with respect to the child's age. The scaled scores can be aggregated into the **phonological awareness composite**, which is constructed to be normally distributed around the mean of 100 for the population that was used in the development of the test. The score below 100 means that the child is behind what is normal for her age, and the score above 100 means that the child is ahead.

Executive function (EF). One conceptualization of self-regulating skills. It refers to the child's capacity to maintain focused attention, inhibit impulses and switch between tasks. One way to measure executive function is via **Hearts and Flowers** computerized test. It presents the subject with a series of stimuli - either hearts or flowers - located either on the left or the right side of the screen. When seeing a heart, the subject needs to press a button on the same side of the screen, and when seeing a flower - on the opposite side.

Grapheme. In this document: a letter or a short sequence of letters encoding a phoneme or a short sequence of phonemes (e.g. X can encode two phonemes: *[k;s]*). Some papers use the word "grapheme" to denote letters. In this document, such sequences as PH and SH are also considered graphemes.

Instant hit. A case when a child immediately places the correct block into a slot in the scaffolded mode.

Instant miss. A case when a child immediately places an incorrect block into a slot in the scaffolded mode, without trying to locate the correct block by tapping on various blocks on the keyboard.

Invented spelling. A phenomenon in early literacy development when a child tries to infer spelling of a word using her developing phonological knowledge.

Knowledge of letter-sound-pattern correspondence. Knowledge of which patterns of letters tend to encode which patterns of phonemes. Note that this is a more sophisticated skill than simply knowledge of letter-sound correspondence, which deals with sounds of individual letters, and can involve knowledge of larger patterns, such as common morphemes.

Phonological awareness (PA). The ability to recognize the sound structure of words. PA typically progresses from the ability to identify large parts of the word (such as syllables) to the ability to identify phonemes (the atomic units of spoken language). PA in itself isn't related to written text, but it is crucial for early literacy development, particularly in English.

Phonics. A method of literacy instruction based on explicit teaching of phoneme-to-grapheme correspondence.

Scaffolding. A method in education based on providing support to the student in context of a particular project, in order to bring its difficulty into the child's Zone of Proximal Development.

Sound creature. A name we gave to the onomatopoeic mnemonic characters, designed to represent various phonemes.

Sprite. An image out of which a composition can be arranged.

Zone of Proximal Development (ZPD). A range of tasks that the child can't yet do independently, but can do with support. According to Vygotsky, child's learning and development primarily occur within ZPD.

Appendix B. “Sound Creatures” Catalog

Vowels

[ɑ] (as in f**a**ther)

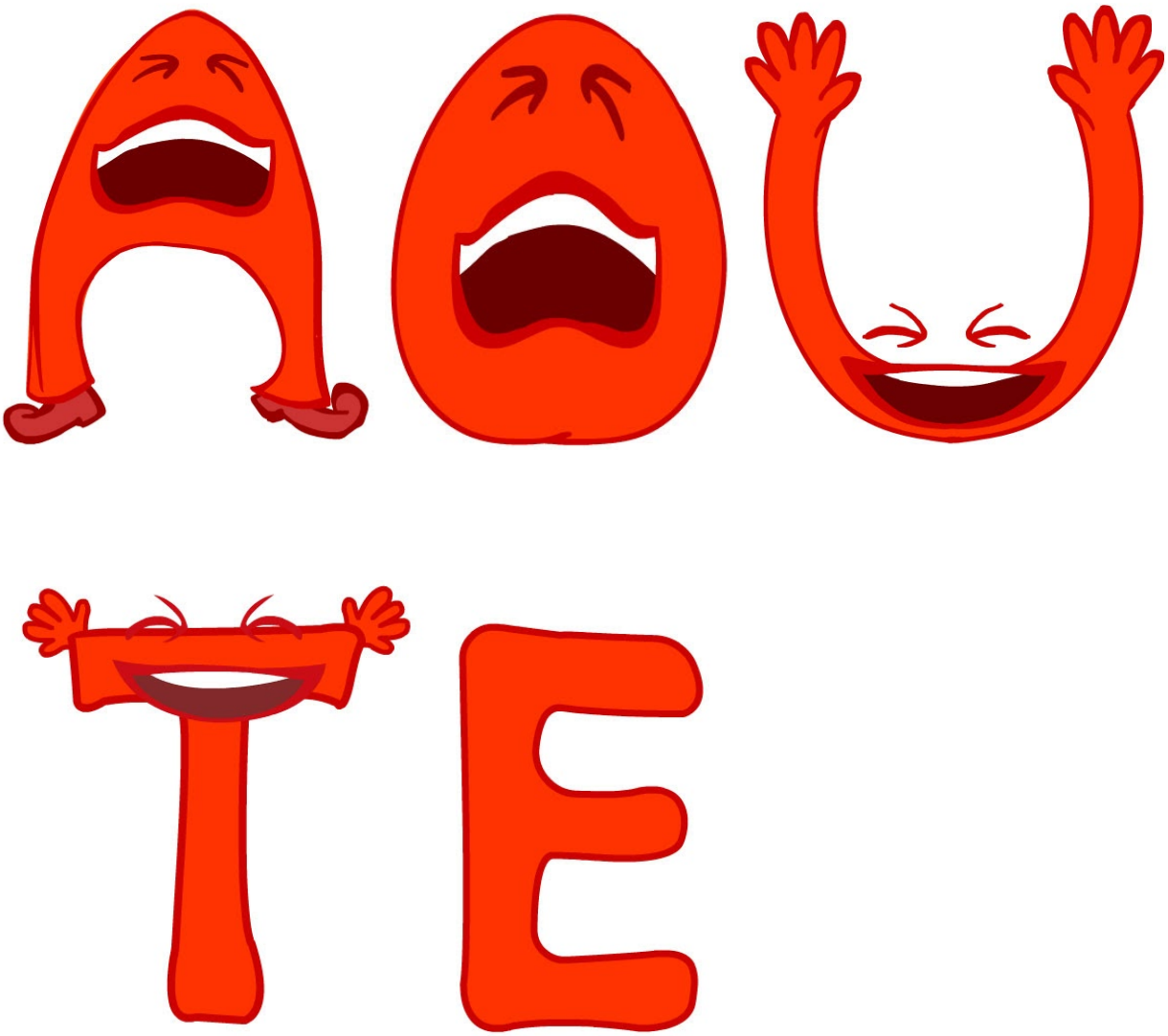
Name: Alita.
Legend: “Alita sings: [ɑ]-[ɑ]-[ɑ]”
Graphemes: A (f**a**ther), O (f**o**x, p**o**t), E (g**e**nre)
Concept: Abraham Tena. **Design:** Ivan Sysoev. **Animation:** R Ryan Hayes.

[æ] (as in c**a**t)

Name: Andy.
Legend: “Andy dislikes his food: [æ]”
Graphemes: A (c**a**t)
Concept, design and animation: Ivan Sysoev.

[ʌ] (as in tr**u**ck)

Name: Alex.
Legend: “Alex laughs: [ʌ]! [ʌ]! [ʌ]!” (a villainy kind of laughter)
Animation Description: Laughs hysterically while waving his hands or rocking from side to side
Graphemes: A (dat**a**), O (c**o**me), U (tr**u**ck), TE (lis**te**n)
Inspiration: Dekodiphukan. **Design:** Ivan Sysoev.
Animation: Ivan Sysoev (A, O, U), Allan & Danny Gelman (TE)



[ɛ] (as in v**e**ry)

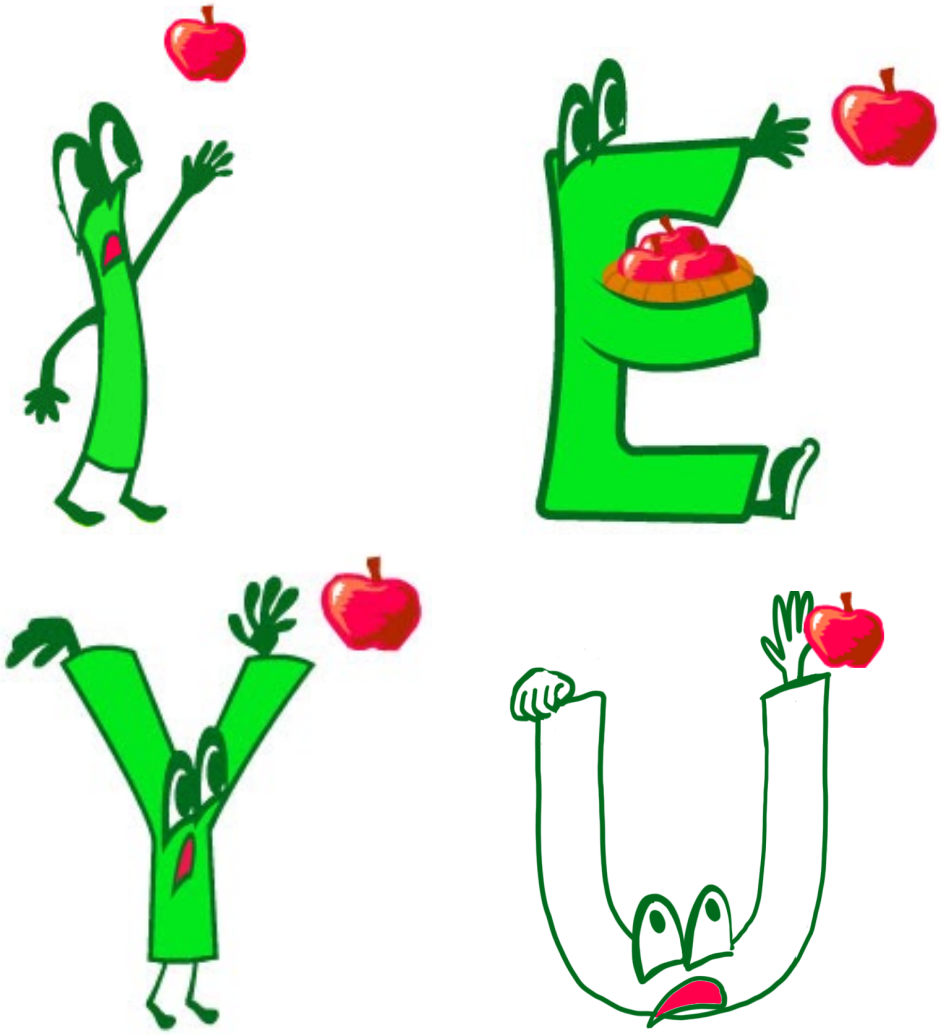
Name: Eddy.
Legend: “Eddy struggles to hear: eh?”
Graphemes: E (v**e**ry), A (c**a**re), U (b**u**ry)
Inspiration: Dekodiphukan. **Design and animation:** Ivan Sysoev.

[i] (as in squ**e**ak)

Name: Eve.
Legend: “Eve squeaks: eeeeeeeee!!!” (after stepping into / touching something icky)
Graphemes: E (w**e**), I (mach**i**ne), Y (ver**y**)
Inspiration: Dekodiphukan. **Design and animation:** Ivan Sysoev.

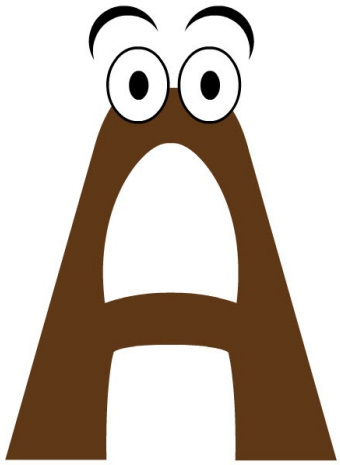
[ɪ] (as in h**i**t)

Name: Ines.
Legend: “Ines tries to reach: ih! Ih!” (while trying to pick a fruit)
Graphemes: I (h**i**t), E (**E**nglish), Y (s**y**stem), U (b**u**siness)
Concept and design: Ivan Sysoev. **Animation:** R Ryan Hayes (E, Y), Allan & Danny Gelman (I)
Note: The version for U is incomplete; showing a sketch here.



[ɔ] (as in l**o**ng)

Name: Olaf.
Legend: “Olaf is surprised: ooooooooh!”
Graphemes: O (l**o**ng), A (**a**lso)
Inspiration: Dekodiphukan. **Design:** Ivan Sysoev. **Animation:** Allan & Danny Gelman



[oʊ] (as in n**o**)

Name: Owen.
Legend: “Owen sees that Oh[oʊ]! He’s late!”
Graphemes: O (n**o**)
Concept and design: Ivan Sysoev. **Animation:** Allan & Danny Gelman

[u] (as in t**oo**)

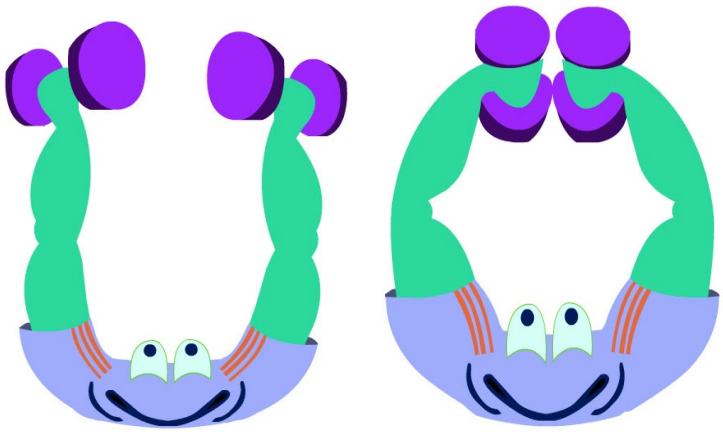
Name: Uno.
Legend: “Uno hoots: Oo! Oo!”
Graphemes: OO (t**oo**), O (wh**o**), U (fl**u**id), EW (cr**ew**)
Concept and design: Ivan Sysoev. **Animation:** Lingxi Li (U), Allan & Danny Gelman (O, OO, EW)





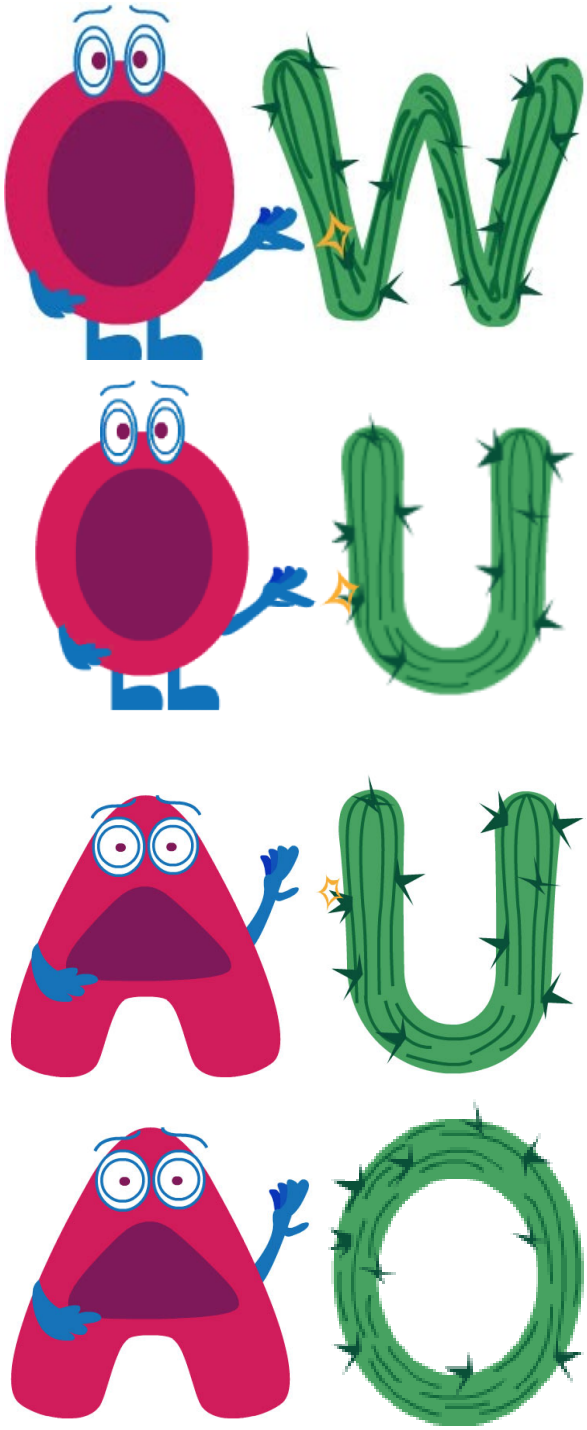
[ʊ] (as in p**u**t)

Name: Ulrich.
Legend: “Ulrich lifts weights: Uh! Uh!”
Graphemes: U (p**u**t), O (w**o**lf)
Concept and design: Ivan Sysoev. **Animation:** Allan & Danny Gelman



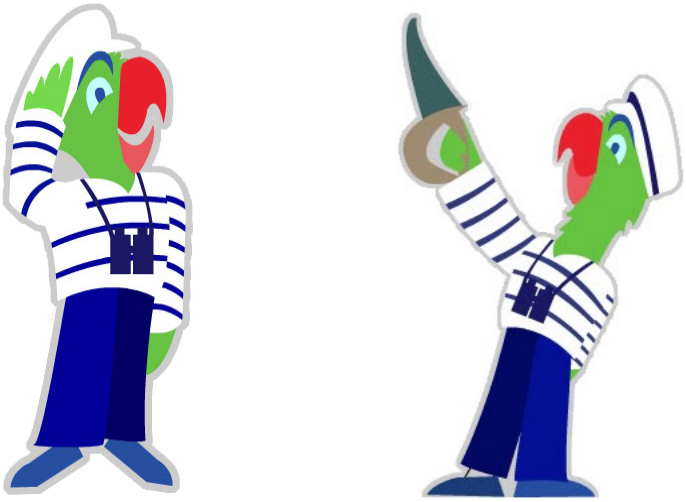
[aʊ] (as in c**ow**)

Name: Owli.
Legend: “Owli touches a cactus: Ow!”
Graphemes: OW (c**ow**), OU (**ou**t), AU (M**au**i), AO (T**ao**ism)
Concept and design: Ivan Sysoev. **Animation:** Allan & Danny Gelman
Note: The name is made-up. No real name starting with [aʊ] was found.



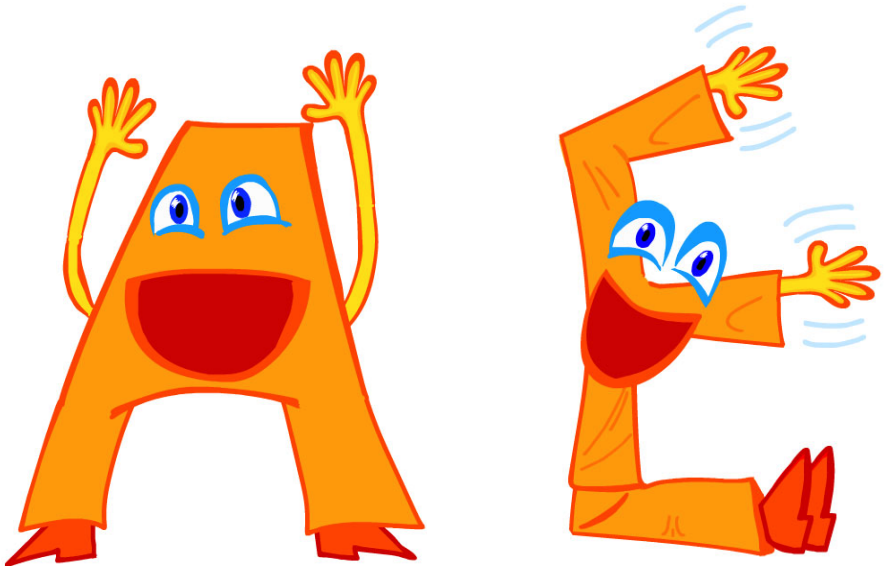
[aɪ] (as in h**i**)

Name: Isaac.
Legend: “Isaac says: aye aye, captain!” (“Aye” sounds like [aɪ])
Graphemes: I (h**i**), Y (m**y**)
Concept and design: Ivan Sysoev. **Animation:** Allan & Danny Gelman



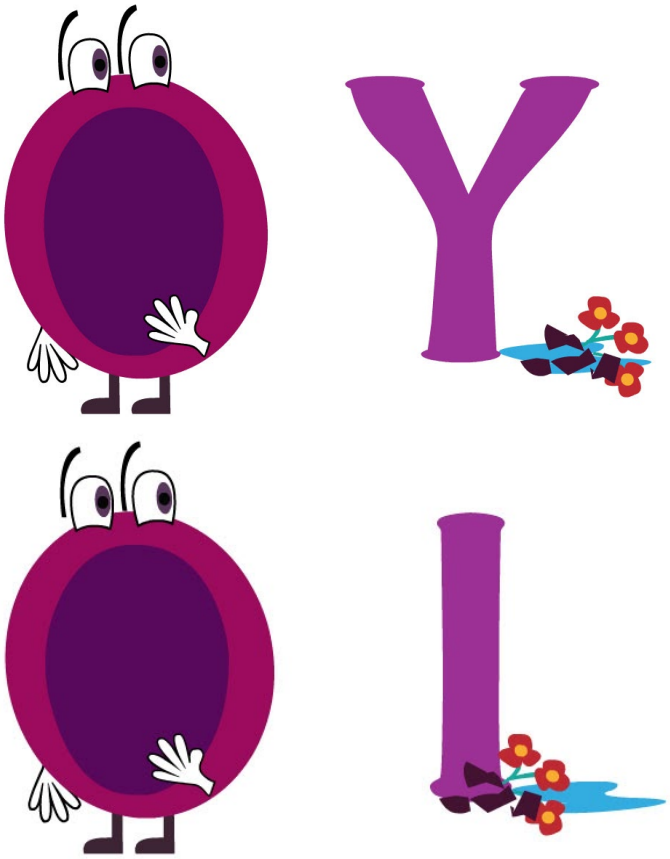
[eɪ] (as in n**a**me)

Name: Abe.
Legend: “Abe waves to a friend: Ey!”
Graphemes: A (n**a**me), E (caf**e**)
Concept, design and animation: Ivan Sysoev.



[ɔɪ] (as in b**oy**)

Name: Oin.
Legend: “Oin drops a vase: Oy!”
Graphemes: OY (b**oy**), OI (p**oi**nt)
Concept and design: Ivan Sysoev. **Animation:** Allan & Danny Gelman

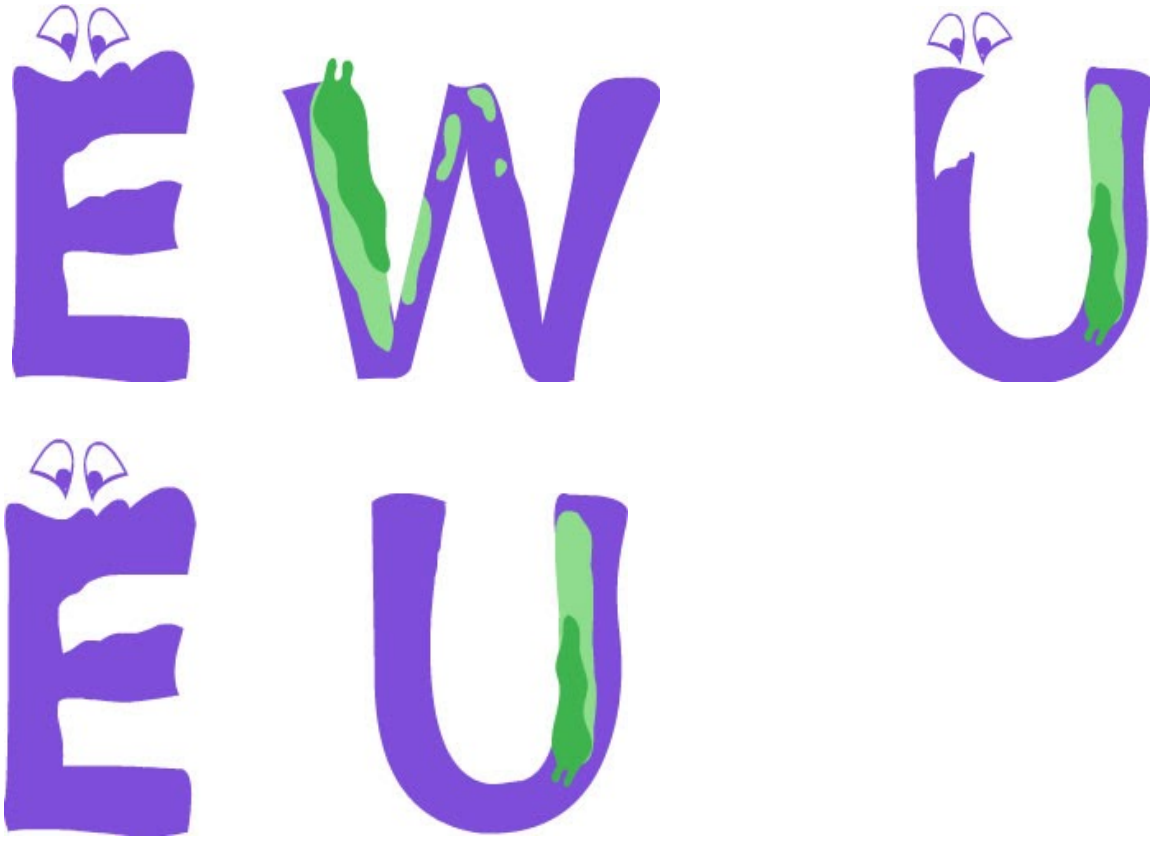


r-colored vowel [ɝ] (as in f**ur**)

Name: Ernie.
Legend: “Ernie ribbets: Er! Er!”
Graphemes: any vowel + R (f**ur**, c**ir**cle, w**or**k, etc.)
Concept and design: Ivan Sysoev.
Animation: Allan & Danny Gelman
Note: All r-colored vowels in CMU pronouncing dictionary are coded by the same symbol. R-colored vowels are specific for North American English. For learning purposes, it might be more natural to ignore r-coloring and decouple the vowel from R. This would alleviate the problem of children typically confusing this creature with the one for [r].

[ju] (as in n**ew**)

Name: Eugene.
Legend: “Eugene is disgusted: Ew!”
Graphemes: U (**u**se), EU (**Eu**rope), EW (n**ew**)
Concept and design: Ivan Sysoev. **Animation:** Allan & Danny Gelman



Consonants

[b] (as in **b**all)

Name: Billy.
Legend: “Billy bounces a ball: b! b! b!”
Graphemes: B (**b**all)
Concept and design: Ivan Sysoev.
Animation: R Ryan Hayes

[d] (as in **d**rum)

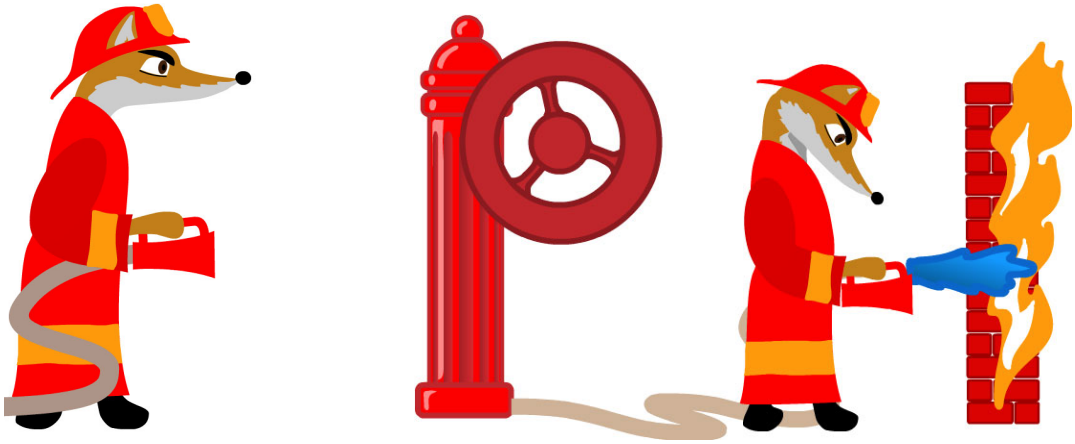
Name: Dan.
Legend: “Dan drums: d! d! d!”
Graphemes: D (**d**ruml)
Concept, design and animation: Ivan Sysoev.
Note: The choice of animal is not ideal: children were seen using the rebus principle and thinking that this is a RABBIT for [r].

[g] (as in **g**ulp)

Name: Greg.
Legend: “Greg gulps grape juice: g! g! g!”
Graphemes: G (**g**ulp)
Inspiration: Dekodiphukan
Design and animation: Ivan Sysoev.

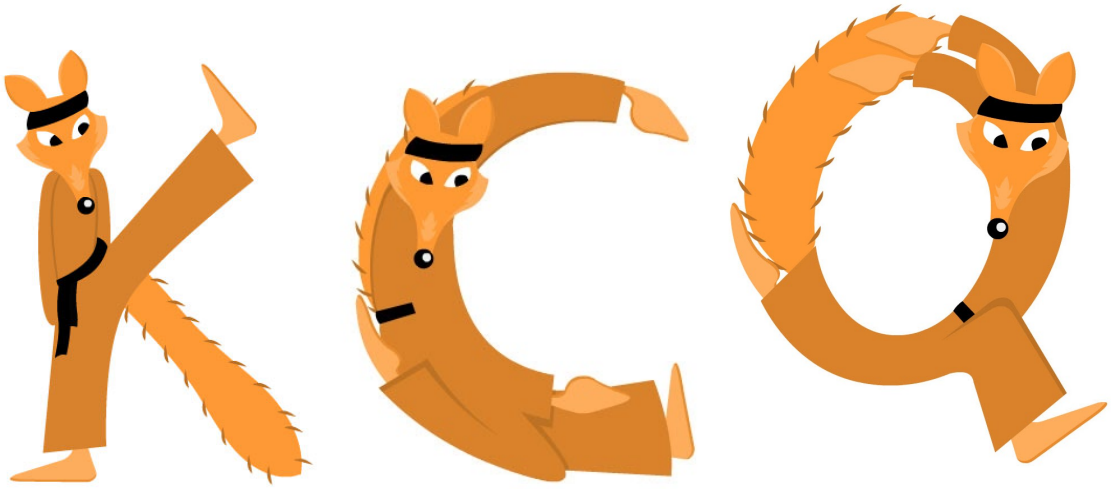
[f] (as in **f**ox)

Name: Fred.
Legend: “Fred fights fire: ffffff!” (the sound of water rushing out of the hose)
Graphemes: F (**f**ox), PH (**ph**one)
Inspiration: Dekodiphukan
Design: Ivan Sysoev. **Animation:** Ivan Sysoev, **later modified by** R Ryan Hayes



[k] (as in **k**ite)

Name: Kathy.
Legend: “Kathy does karate kicks: k! k! k!”
Graphemes: K (**k**ite), C (**c**at), Q (**q**ueen)
Inspiration: Leapfrog
Design: Ivan Sysoev. **Animation:** R Ryan Hayes
Note: Kathy is a **c**oyotte.

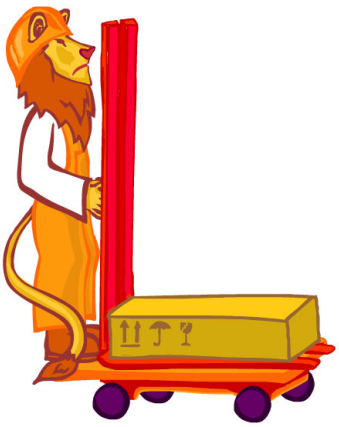


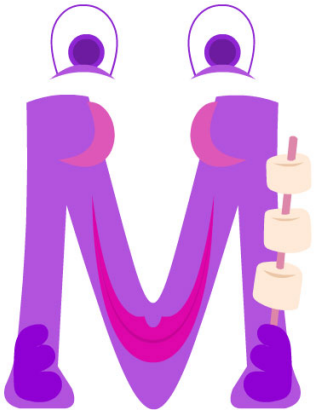
[h] (as in **h**ouse)

Names: Henry and Harry.
Legend: “Henry and Harry are blowing on hot food: hhhh! hhhh!”
Graphemes: H (**h**ouse)
Inspiration: Dekodiphukan
Design and animation: Ivan Sysoev.
Note: The pizza that the characters are holding was a distraction for many children, making them think that the associated sound must be somehow connected to pizza (e.g. [p]).

[l] (as in **l**ion)

Name: Leo.
Legend: “Leo uses a forklift: LLL! LLL!”
Graphemes: L (**l**ion)
Concept: Anneli Hershman
Design and Animation: Ivan Sysoev.



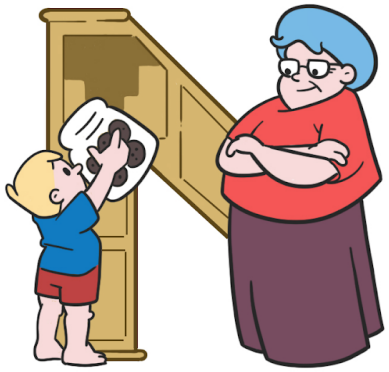


[m] (as in **m**ouse)

Name: Mary.
Legend: “Mary munches on a marshmallow: mmmmmmm!!” (enjoys the marshmallow)
Graphemes: M (**m**ouse)
Inspiration: Dekodiphukan
Design: Ivan Sysoev.
Animation: R. Ryan Hayes

[n] (as in **n**o)

Name: Nana.
Legend: “Nana says: No-no-no!” (to a boy who is trying to get sweets without permission)
Graphemes: N (**n**o)
Concept, design and animation: Abraham Tena.



[p] (as in **p**op)

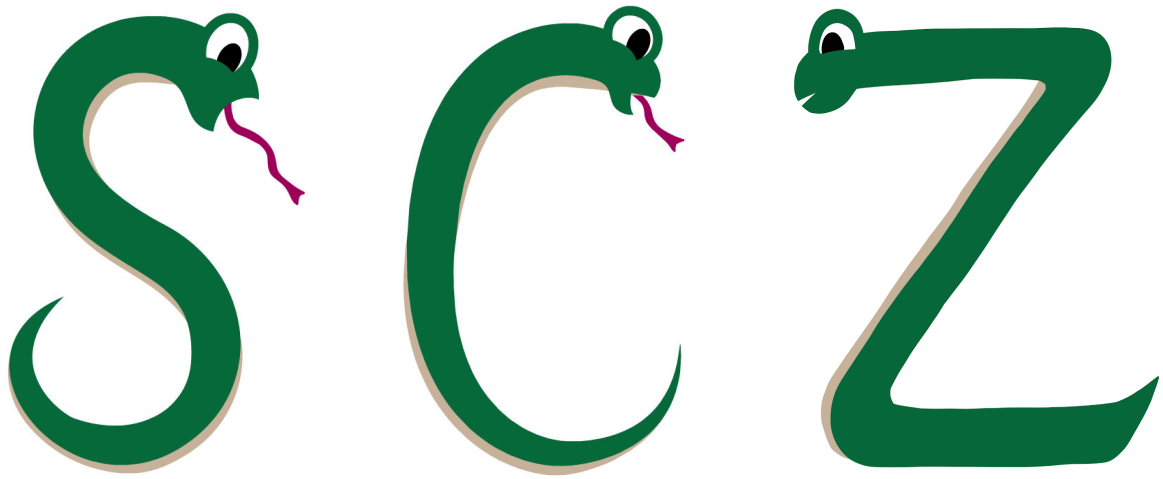
Name: Paul.
Legend: “Paul pops his bubblegum: p! p!”
Graphemes: P (**p**op)
Inspiration: Leapfrog.
Design and Animation: R. Ryan Hayes

[r] (as in **r**oar)

Name: Rex.
Legend: “Rex roars: RRRRR!”
Graphemes: R (**r**oar)
Concept: Ivan Sysoev.
Design and animation: R. Ryan Hayes
Note: Rex was originally a tiger (which ends with [r]). However, it turned out that children think that tiger says “Gurr”, while roaring is culturally associated with lions.

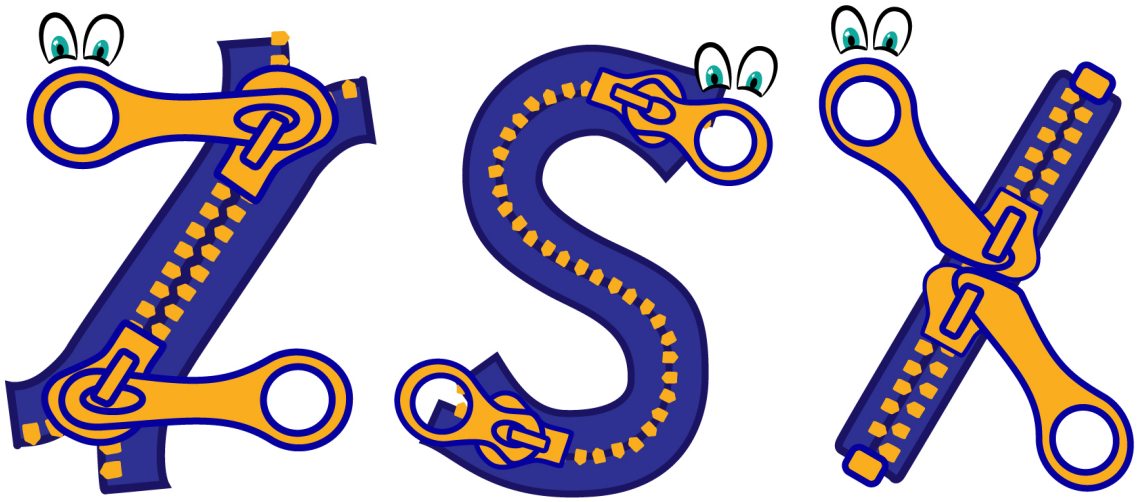
[s] (as in **s**nake)

Name: Sally.
Legend: “Sally hisses: sssss!”
Graphemes: S (**s**nake), C (**c**ity), Z (Swit**z**erland)
Concept and design: Ivan Sysoev. **Animation:** Ivan Sysoev (S, C), Jesso Wang (Z)



[z] (as in **z**ipper)

Name: Zack.
Legend: “Zack zips: zzzz! zzzz!”
Graphemes: Z (**z**ipper), S (u**s**e), Z (**X**erox)
Concept and design: Ivan Sysoev. **Animation:** Ivan Sysoev (Z), Jesso Wang (S, X)



[v] (as in **v**acuum)

Name: Vinny.
Legend: “Vinny got caught in a vacuum: vvvvv!” (the sound of the vacuum)
Graphemes: V (**v**acuum)
Concept and design: Ivan Sysoev.
Animation: Allan & Danny Gelman

[t] (as in **t**ap)

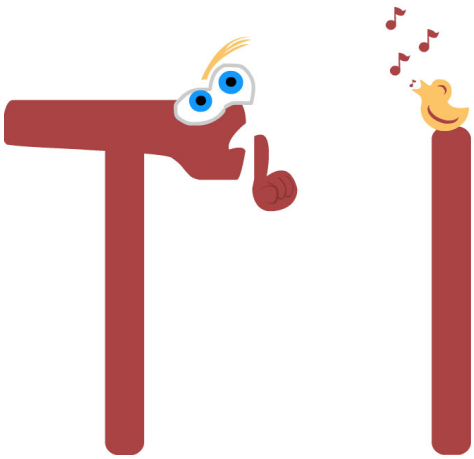
Name: Tommy.
Legend: “Tommy taps: t-t-t-t!“
Graphemes: T (**t**ap)
Inspiration: Lively Letters
Design and animation: Ivan Sysoev.

[ks] (as in fo**x**)

Name: Xenia.
Legend: “Xenia takes an X-ray: ks! ks!” (the sound of the camera)
Graphemes: X (fo**x**)
Inspiration: Dekodiphukan
Design and animation: Ivan Sysoev

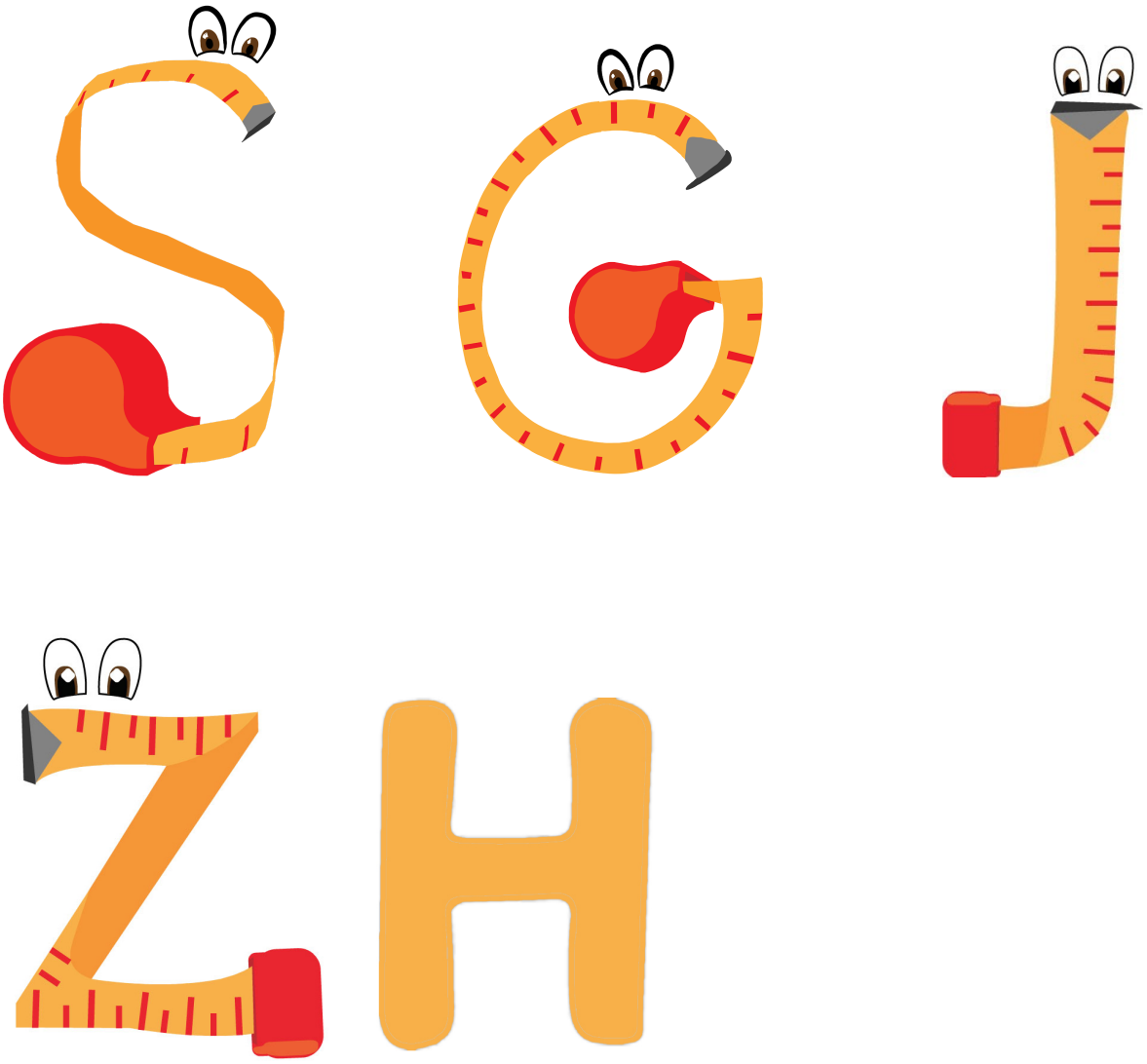
[ʃ] (as in **sh**ark)

Name: Sharon.
Legend: “Sharon shushes: shhhh!”
Graphemes: SH (**sh**ark) or SI (expan**si**on) or S (**s**ure), TI (na**ti**on), CI (so**ci**al) or CH (**ch**icago)
Inspiration: Reading Genie
Design: Ivan Sysoev **Animation:** Allan & Danny Gelman



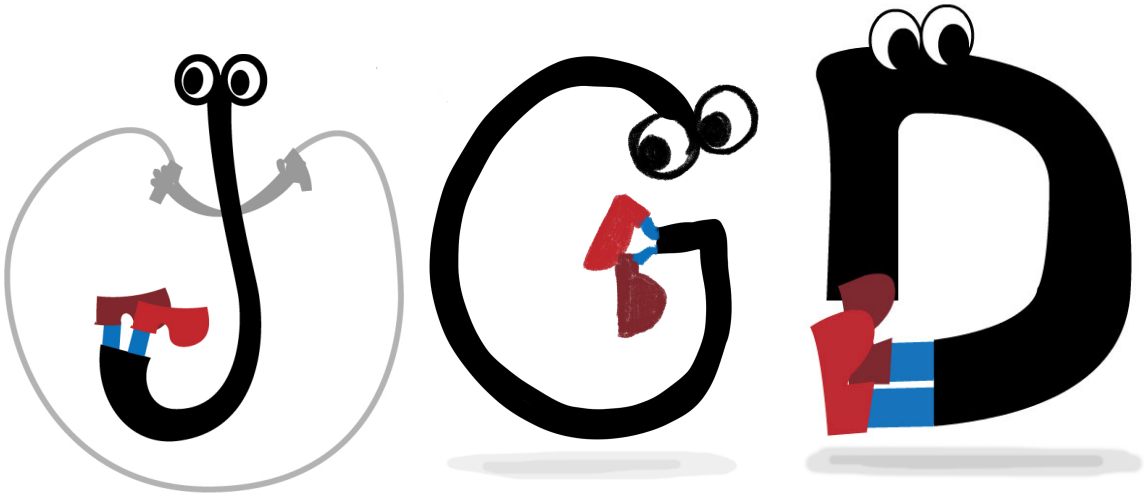
[ʒ] (as in mea**s**ure)

Name: Jacques.
Legend: “Jacques measures: zhhh! zhhh!” (the sound of measuring tape pulled out of its case)
Graphemes: S (trea**s**ure), G (re**g**ime), J (**J**acques), ZH (**Zh**ao)
Concept and design: Ivan Sysoev
Animation: Jesso Wang (Z, G), Allan & Danny Gelman (J, ZH)



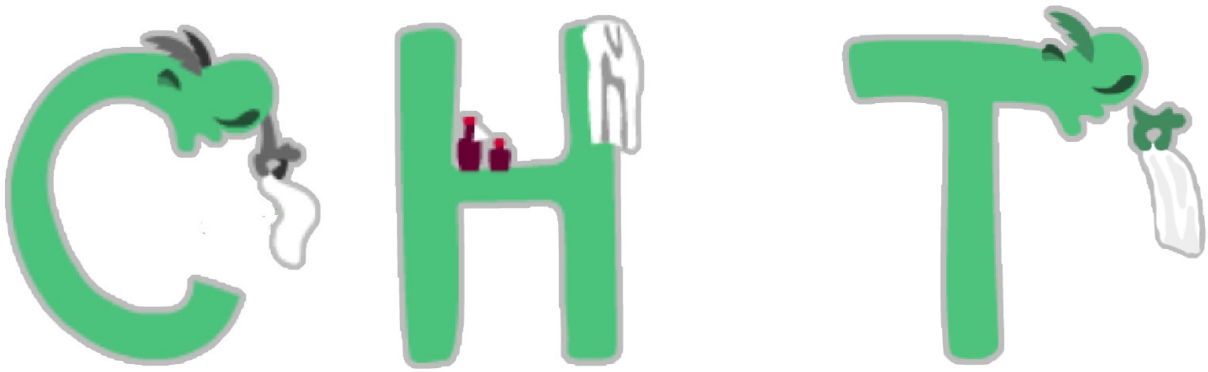
[dʒ] (as in **j**ump)

Name: Jim.
Legend: “Jim jumps: [dʒ]-[dʒ]-[dʒ]-[dʒ]!“
Graphemes: J (**j**ump), G (**g**eneral), D (e**d**ucation)
Inspiration: Leapfrog
Design: Ivan Sysoev. **Animation:** Jesso Wang (J, G), Allan & Danny Gelman (D)



[tʃ] (as in **ch**icken)

Name: Chuck.
Legend: “Chuck sneezes: [tʃ]! [tʃ]!“
Graphemes: CH (**ch**icken) or C (**c**ello), T (na**t**ure)
Concept and design: Ivan Sysoev. **Animation:** Allan & Danny Gelman.

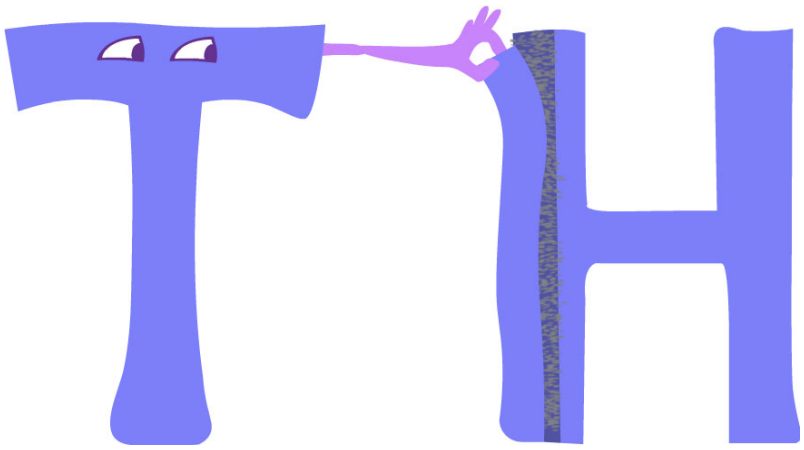


[θ] (as in **th**ing)

Name: Theo.
Legend: “Theo sneaks: [θ]! [θ]!“ (the sound of soft footsteps)
Graphemes: TH (**th**ing)
Inspiration: Reading Genie
Design: Ivan Sysoev. **Animation:** Allan & Danny Gelman.

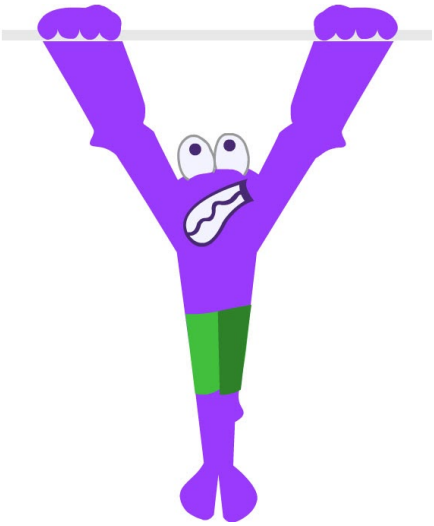
[ð] (as in **th**is)

Name: Thayn.
Legend: “Thayn peels velcro: [ð]! [ð]!“
Graphemes: TH (**th**is)
Inspiration: Reading Genie
Design: Ivan Sysoev. **Animation:** Allan & Danny Gelman.
Note: The Thayn name is made up. I was unable to find a real name starting with [ð].



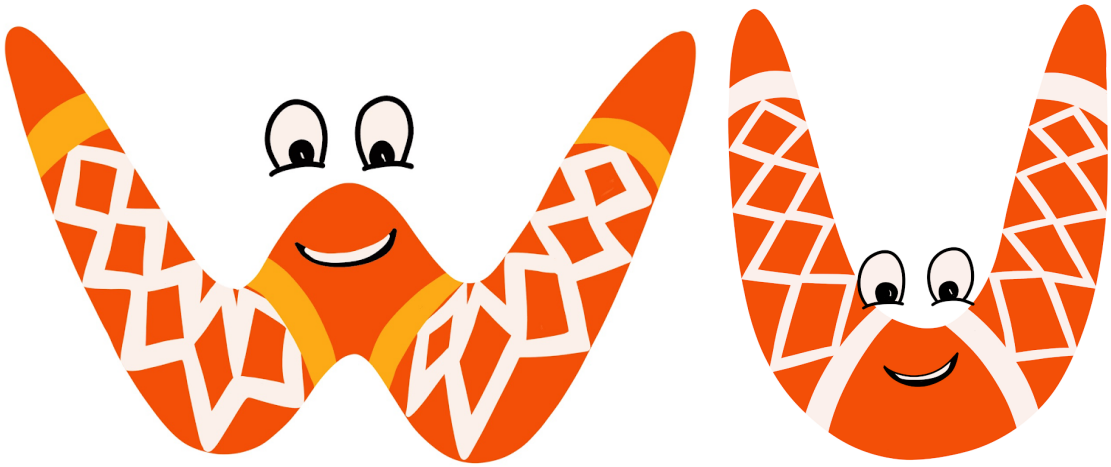
[j] (as in **y**oga)

Name: Yorick.
Legend: “Yorick struggles with pullups: [j]! [j]!“ (a struggling sound)
Graphemes: Y (**y**oga)
Concept and design: Ivan Sysoev. **Animation:** Allan & Danny Gelman.



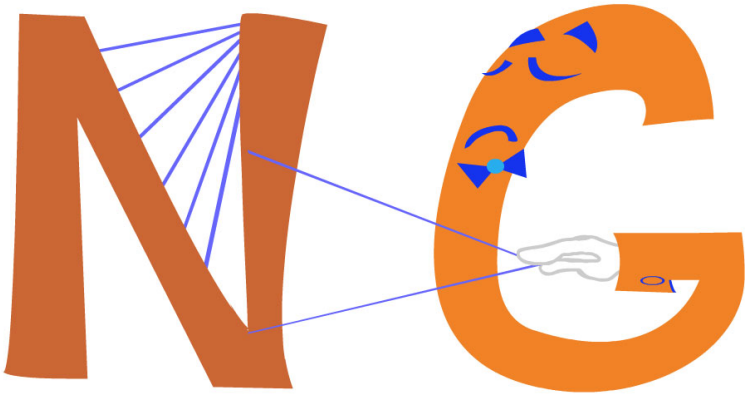
[w] (as in **w**ave)

Name: Willie.
Legend: “Willie the boomerang spins: w-w-w-w!” (the sound of boomerang flying through the air)
Graphemes: W (**w**ater), U (q**u**een)
Concept and design: Ivan Sysoev **Animation:** Jesso Wang



[ŋ] (as in ki**ng**)

Name: Ngazi.
Legend: “Ngazi plucks a string: ng! ng!” (the sound produced by the string)
Graphemes: NG (ki**ng**)
Concept, design and animation: Ivan Sysoev
Note: Ngazi is a made-up name. An alternative is Ting - a real name that ends with [ŋ].



Appendix C. Tools

Several custom tools were implemented to facilitate analysis of SpeechBlocks data.

Play Observatory

To observe and annotate activities from the first SpeechBlocks I pilot in context of what was happening in the classroom, I implemented a tool called Play Observatory (Fig. A1). It plays synchronized observation video(s) plus a simulation of children's screens reconstructed from the log files. The tool includes shortcuts for quick and convenient coding of both point events and intervals on the timeline of the session.

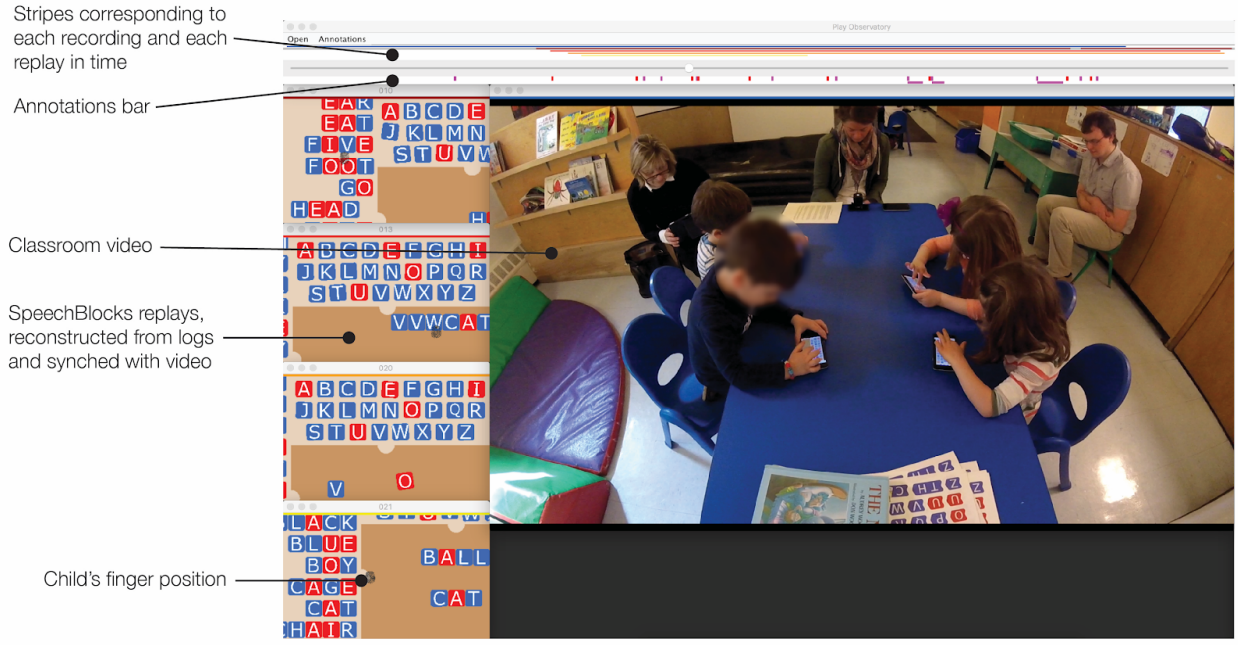


Fig. A1. Play Observatory

PlayTrees

PlayTrees is a visualization designed to have a quick, bird-eye view at a SpeechBlocks I session. They were used both by researchers and by literacy coaches. They were co-designed and co-developed with a fellow student, Mina Soltangheis (Soltangheis, 2017). PlayTrees depicts the process of construction, deconstruction and remixing of words. Time on PlayTrees diagrams flows from top to bottom. Each node represents a word, and edges represent ancestral relations between words. Therefore, when two edges on a diagram merge, it represents putting two words together, and when a split occurs, it corresponds to pulling a word apart. Thus, PlayTrees are

technically directed acyclic graphs, not necessarily trees, but they are called “trees” for simplicity. Several variations of PlayTrees were developed for different purposes. Fig. A2 shows one such variation. In addition to visualizing PlayTrees, an annotator was built to label their nodes.

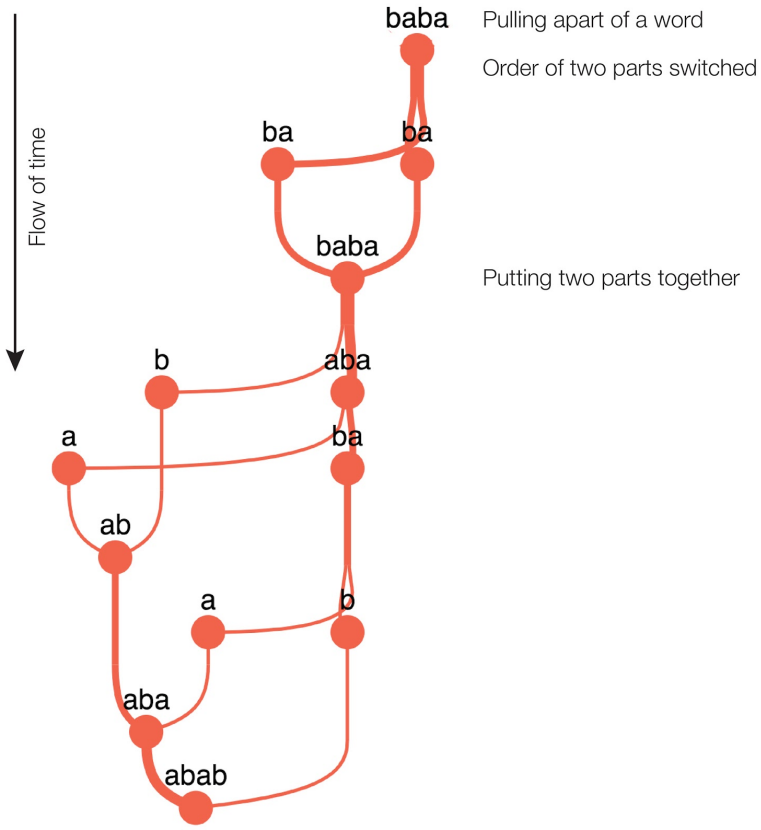


Fig. A2. Play Trees

Play Frames

In order to analyze activities with visuals in SpeechBlocks II, I develop a tool that reconstructs construction and manipulation of scenes step-by-step from the log files.

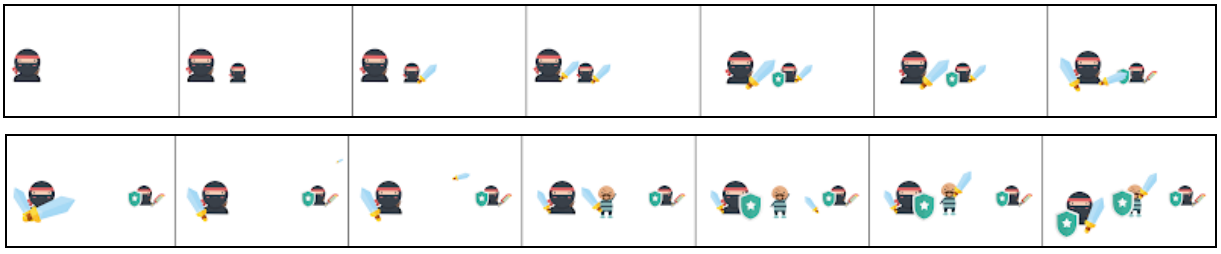


Fig. A3. Play Frames

Appendix D. Additional Statistics

In section 6.7, I show a multiple regression analysis of CTOPP gains. This analysis was done for raw CTOPP (sum of the three phonological awareness scores). If I use CTOPP composite (based on scaled, age-adjusted scores), the results are slightly different: the interaction for executive function is not significant anymore, while the gender-related interaction is more prominent.

Overall p-value: 0.06. F-statistic: 2.118 on 7 and 48 DF						
variable	coefficient	p-value	low-95% bound	high-95% bound	low-90% bound	high-90% bound
treatment	-3.18	0.31	-9.35	3	-8.33	1.98
pre-comp. CTOPP	**-5.22**	**0.004 ****	-8.12	-1.72	-8.14	-2.3
pre-comp. CTOPP X treatment	**5.46**	**0.02 ***	0.81	10.1	1.58	9.33
pre-EF	-1.15	0.46	-4.26	1.95	-3.74	1.44
pre-EF X treatment	2.47	0.28	-2.12	7.07	-1.35	6.3
gender (m)	-4.9	0.13	-11.35	1.54	-10.28	0.47
gender (m) X treatment	**8.89**	**0.05 ·**	-0.14	17.91	1.36	16.41