Estimating child linguistic experience from historical corpora

Jordan Kodner
University of Pennsylvania
Philadelphia, PA
jkodner@sas.upenn.edu

Abstract Child language acquisition is often identified as one of the primary drivers of language change, but the lack of historical child data presents a challenge for empirically investigating its effect. In this work, I observe the relationship between lexicons extracted from modern child-directed speech and those drawn from modern and historical literary corpora in order to better understand when language acquisition can be modeled over historical and non-child corpora as it is over child corpora. The type frequencies of morphophonological and syntactic-semantic patterns occur at similar type frequencies in these corpora among high token frequency items, and furthermore, when a learning algorithm is applied to lexicons sampled from these sources, it consistently achieves the same learning outcomes in each. With appropriate care and pre-processing, modern and historical text corpora are effectively interchangeable with child-directed speech corpora for the purpose of estimating child lexical experience, opening a path for modeling language acquisition where child-directed corpora are not available.

Keywords: child language acquisition, corpus linguistics, historical linguistics

1 Introduction

The advent of child-directed speech (CDS) corpora in recent decades containing years' worth of early linguistic input (CHILDES; MacWhinney 2000) has facilitated significant progress in the field of native language acquisition. That said, no CDS corpora exist for the overwhelming majority of the world's languages, and none that do exist date back before the mid-20th century. Without such corpora, the insights that child language acquisition researchers gain from modern methodologies cannot be extended to most of today's world, let alone to past eras. The contribution of this paper is methodological: I establish that, despite the differences that intuitively exist, CDS and modern and historical non-CDS corpora are fundamentally similar along dimensions relevant for native language acquisition. This stands to facilitate acquisition research for a more diverse range of languages, and, as discussed here, research into child language acquisition in the past.

Four aspects of language learning which are reflected in CDS motivate this work. First, the relative uniformity of language acquisition: learners exhibit remarkable synchronic uniformity despite the variability of the input they receive (Labov 1972). Second, the crucial role of type frequency: convergent results from a wide variety of research programs connect grammar learning to the *number of types* over which linguistic patterns are expressed in the input rather than the attestation of any particular lexical items (Aronoff 1976; MacWhinney 1978; Bybee 1985; Baayen 1993; Elman 1998; Pierrehumbert 2003; Yang 2016). Third, token frequency and availability: the relative age at which learners acquire vocabulary items is correlated with their *token frequencies* (Goodman et al. 2008) in the input. And fourth, small early vocabularies: the typical learner knows only a few hundred to a thousand words by around age three (Hart & Risley 1995; 2003; Szagun et al. 2006). Since children acquire most properties of their native grammars by that age, the bulk of grammar acquisition is undertaken on the basis of relatively few mostly high frequency items rather than large adult-like lexicons.

Lexical variability between CDS corpora reflects the real-world variation in early linguistic experience that leads to precociousness or delays among learners (Maratsos 2000; Yang 2002). It also reflects realistic assumptions about learner knowledge. Since higher token frequency items tend to be acquired earlier, young learner's lexicons may be estimated trimming off the less frequent items from CDS (Nagy & Anderson 1984; Yang 2016). Doing so yields approximations of "typical" children's lexicons which are the right size and consist primarily of high frequency items. It is these properties that make corpora of child directed speech such useful resources for studying grammar learning. If the field can establish whether historical and other non-CDS corpora share these properties as well, researchers can apply models of language acquisition to historical data to work out how, when, and whether the process of native language acquisition effects change.

To that end, I conduct three studies which elaborate on the similarities between modern and historical non-CDS corpora on one hand and CDS on the other for the purpose of modeling productivity. I begin in Section 2 by illustrating the effect that trimming low token frequency items has on CDS and adult corpora in Modern English. This is extended to historical corpora in Section 3, where I compare semantic overlap between cross-linguistic modern CDS and historical lexicons. Finally, Section 4 demonstrates that a type-based threshold learning algorithm to morphological problems yields the same acquisition outcomes in Modern English lexicons taken from CDS and modern non-CDS and to Icelandic lexicons drawn from historical and modern non-CDS.

2 Verbal lexicons derived from child-directed speech and adult corpora

This study establishes the similarity between lexicons derived from adult literary corpora and those derived from corpora of child directed speech. I begin by demonstrating the effect of trimming low frequency vocabulary from the extracted lexicons, and following that, I compare the attested type frequency of various linguistic properties between the adult and CDS-derived lexicons. Types frequencies of these properties are quantitatively similar in these corpora despite superficial differences in specific lexical content.

Adult corpus lexicons are drawn from the Corpus of Contemporary American English (COCA; Davies 2009), which contains millions of lemmatized and POS-tagged words of text drawn from five genres: spoken, popular magazine, fiction, newspaper, and academic. Each genre contains individual subcorpora for each year, and each subcorpus contains between 2.5 and 5.5 million tokens and between 4,200 and 10,200 verb lemmas when those tagged as auxiliaries or modals are excluded.[1] Child input lexicons are drawn from three lemmatized POS-tagged corpora within CHILDES (MacWhinney 2000), each containing roughly 1,000 unique verb lemmas, again with auxiliaries and modals excluded: Brent (*n*=984; Brent & Siskind 2001), Brown (*n*=916; Brown 1973), and MacWhinney (*n*=1042; MacWhinney 1991).[2] These were chosen for their large size relative to other CDS corpora, each containing about a year's worth of child-directed speech. I focus on verbs here for consistency across studies and because they show more interesting inflectional patterns in English than other syntactic categories do. That said, the Zipfian statistical corpus distributions of verb lemmas, inflectional categories, and so on, are the same as those obeyed by other categories (Chan 2008; Finley 2018), which is demonstrated in practice by learning behavior in computational morphology learners (e.g., Lignos et al. 2010). The results can therefore be extended to other syntactic categories.

The most frequent verb lemmas are tabulated for each CHILDES corpus and COCA subcorpus, and four estimates are made from each with the following frequency cutoffs: $n = all$, 1,042 (all types in the largest of the CHILDES corpora),

[1] Since auxiliaries are excluded, the token frequencies for *be*, *have* and *do* only count the instances of these as main verbs. Since English has few auxiliary and modal types, the choice of whether to include them or not does not meaningfully affect the results.

[2] Lemmas are extracted from the morphological annotations provided in these corpora

500, and 100. The three frequency-trimmed conditions represent the lexicons of late, middle, and early learners respectively. Given the total vocabulary size estimates of Hart & Risley (2003) and Szagun et al. (2006), a learner who only knows about 100 verbs is certainly less than three years old, while one who knows 500 is perhaps closer to school age.

2.1 Raw lexical similarity

Measuring lexical overlap between extracted lexicons illustrates the effect of trimming infrequent vocabulary. *Jaccard similarity* (Jaccard 1901) $|A \cap B|/|A \cup B|$ is employed to measure the set overlap between each pair of lexicons (self-similarity excluded). The metric has a range [0,1] where higher is more similar.

Figure 1 shows the range of Jaccard similarities between CDS and COCA corpora on the left and between COCA corpora of different genres on the right. Two observations stand out. First, similarities are much higher for all frequency-trimmed conditions than for $n = all$, which suggests that items which are not shared between corpora are predominately low-frequency. Second, though CDS-COCA similarities are lower than COCA-COCA similarities, their ranges overlap once frequency trimming has been applied, which means that some CDS corpora are more similar to some COCA corpora than some COCA corpora are to one another. Specific lexical items are not necessarily well-shared across corpora regardless of genre, but frequency trimming improves the situation significantly.

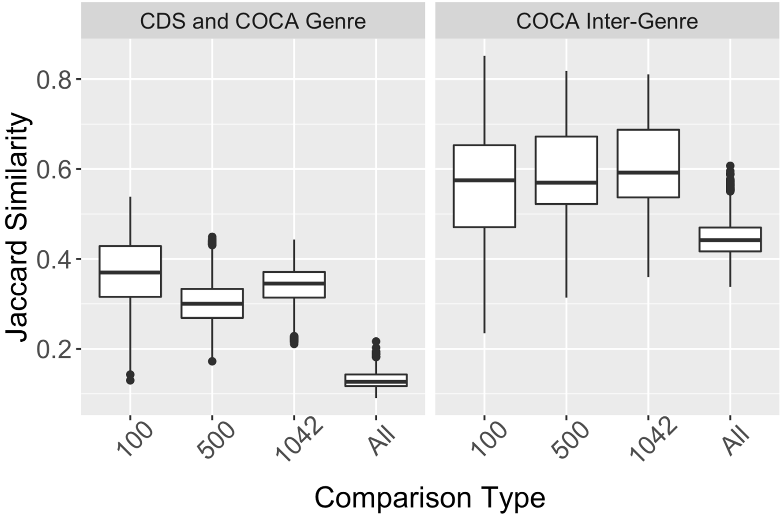

Figure 1: Corpus-derived raw lexicon overlap by comparison type and lexicon size.

2.2 Lexical property similarity

But what matters for learning is often not the individual linguistic items in the input so much as the properties of those items. As discussed in the introduction, the type frequency of some property, the number of items in the lexicon exhibiting it rather than which specific items those are, is what is drives productivity learning. This time, I compare the same adult COCA and CHILDES-derived lexicons in terms of the type frequencies of three linguistic properties. These were chosen for coverage: first, *Latinate* verbs are a morphophonological class which is acquired relatively late, around the start of school (Tyler & Nagy 1989; Jarmulowicz 2002), second, *irregular verbs* are morphological, learned much earlier, and factor into the classic *Past Tense Debates* about productivity in the acquisition literature (Rumelhart & McClelland 1986; Pinker & Prince 1988; Pinker & Ullman 2002: *inter alia*), and third, *double object alternator verbs* are syntactic and semantic in nature (Rappaport Hovav & Levin 2008), and their acquisition is one of the classic case studies in argument structure learning (Baker 1979; Pinker et al. 1987; Gropen et al. 1989; Yang 2016). The results show that there is less variation between corpora in terms of type frequencies in terms of lexical identity, and that the CDS-derived lexicons are in general quantitatively similar to the adult lexicons in the frequency-trimmed conditions.

2.2.1 Irregular verbs

So-called irregular verbs in English are those that undergo stem changes or suppletion when forming the past tense and past participle, e.g., *sing* $\sim$ *sang* $\sim$ *sung*, *go* $\sim$ *went* $\sim$ *gone*, or *tell* $\sim$ *told* $\sim$ *told*. A learner acquiring English verbal morphology must work out which of these verbs are inflected according to some generalizable pattern and which are truly one-off "irregulars" that must be listed or memorized (Berko 1958; Pinker & Prince 1994). Figure 2 shows the mean number of strong verb lemmas by genre for each frequency cutoff n. It is plain from visual inspection alone that CDS and the COCA genres become much more alike when the rare items are trimmed from COCA. It is also striking that academic writing rather than CDS appears to be the greatest outlier for each trimmed condition.

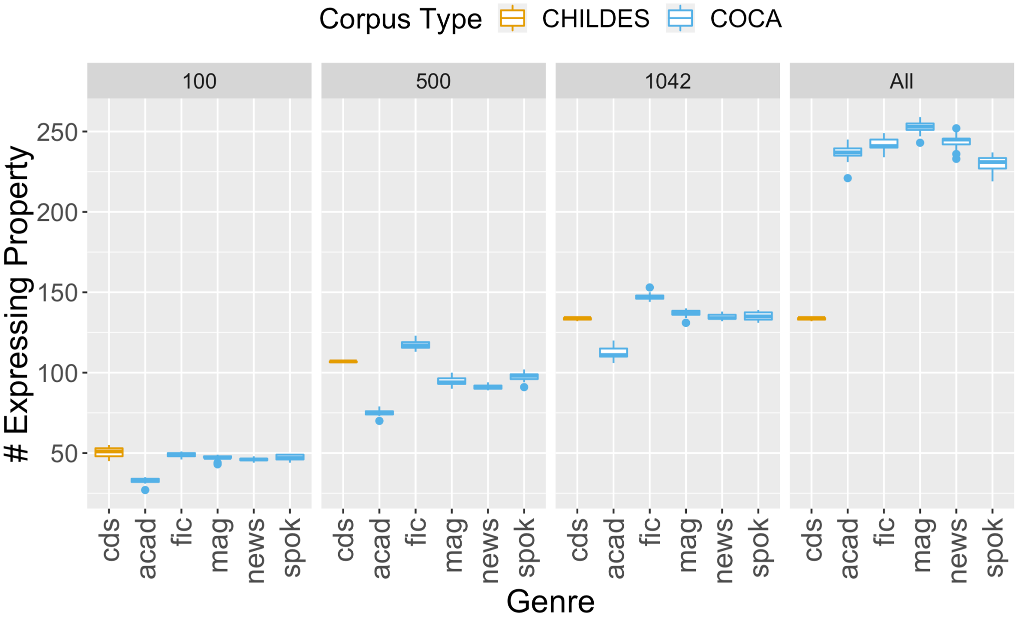

Figure 2: Number of irregular verbs attested per corpus by genre and corpus size.

At $n =$*all*, the adult lexicons contain far more irregular verbs than the CDS-derived lexicons simply because they are taken from larger corpus samples, but when trimmed to $n = 1042$ and 500, CDS falls within the range of the adult lexicons, while at $n = 100$, CDS overlaps with fiction. A regression predicting the number of strong verbs by CDS/adult status finds no significant difference between CDS and adult lexicons in any of the frequency-trimmed conditions – if one were presented with the box plots in Figure 2 with the genre labels and colors removed, it would not be possible to identify which box corresponded to CDS in the trimmed conditions.

2.2.2 Double object / to-dative alternator verbs

The acquisition of DO/to-dative verbs (e.g., *give*, *send* and *tell*)[3] is one of the classic problems in argument structure acquisition. Their attestation in these corpora reveals the same kind of pattern as the irregular verbs: again, trimming the low token frequency items from the COCA-derived lexicons brings them in line with the CDS lexicons.

[3] full list from Levin 1993: §2.1 (119)

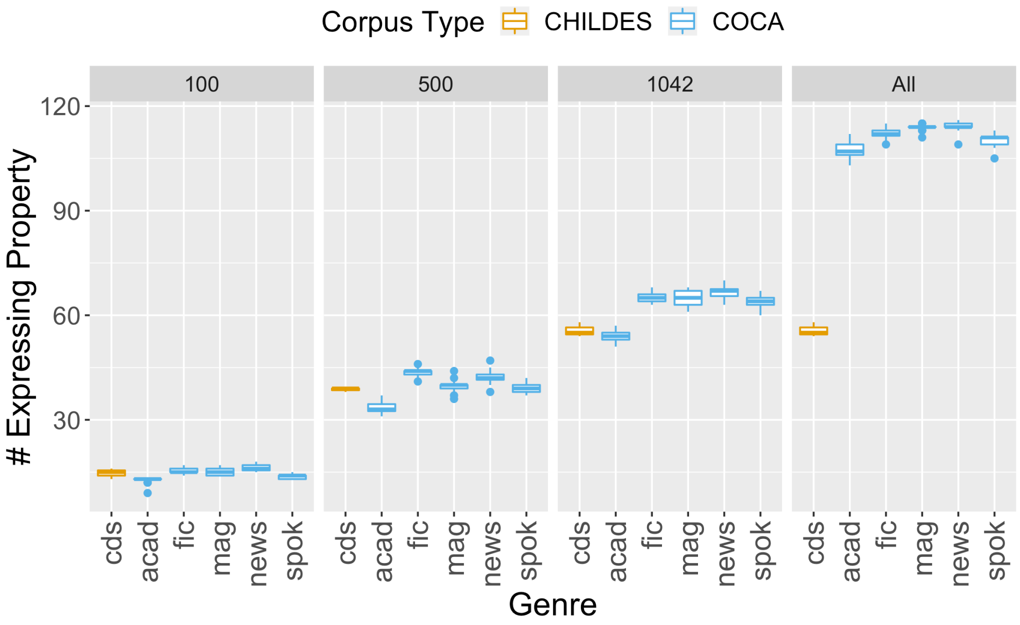

Figure 3: Number of double object alternator verbs attested per corpus by genre and corpus size.

There is no significant difference between CDS and adult lexicons at $n = 500$ or 100, and while CDS is statistically different from adult at $n = 1042$, it is not different from academic, and the difference between CDS and adult means decreased from a factor of about 200% to near 10%.

2.2.3 Latinate verbs

Unlike the previous properties, Latinate verbs are saliently associated with genre (Levin et al. 1981; Levin & Novak 1991), and many, but not all are high-register (COCA contains *encapsulate*, *irradiate*, *reconstitute*, but also *confuse*, *offer*, and *remember*). Additionally, the morphophonological generalizations associated with English Latinate vocabulary are acquired late, typically not until children enter school. As such, we expect there to exist significant quantitative differences between the rate of Latinate verbs in CDS-derived and adult-derived lexicons as shown in Figure 4.

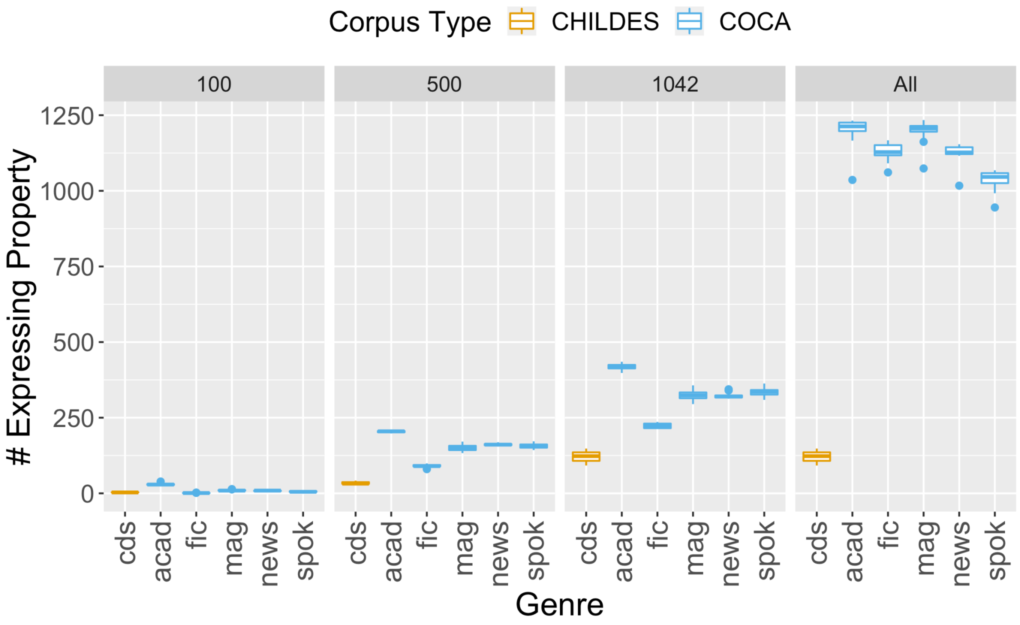

Figure 4: Number of polysyllabic Latinate verbs attested per corpus by genre and corpus size.

This prediction bears out since every test shows a significant difference except for $n = 100$. Nevertheless, frequency trimming brings the type frequencies of CDS and non-CDS much closer together since Latinate vocabulary is disproportionately present among low-frequency items in every COCA genre. Notably, academic lexicons once again differ from all other genres.

2.3 Interim conclusions

These studies show that type frequencies in corpora derived from child-directed speech are statistically similar to frequency-trimmed corpora derived from adult literary genres even though they differ in their specific lexical contents. In every instance, frequency trimming brings CDS-derived and non-CDS-derived type counts much closer together, and in most cases there is no statistically significant difference between the two trimmed lexicon categories. Adult corpora may be reasonably substituted for CDS corpora for the purpose of modeling grammar learning in child language acquisition, since it is these type frequencies that are directly relevant and frequency trimming is just a normal step for approximating child vocabulary size and composition when analyzing CDS for productivity.

3 Verbal lexicons derived from child-directed speech and historical corpora

Child language acquisition is often implicated as a driving force in language change (Sweet 1899; Halle 1962; Kiparsky 1965; Anderson 1973; Baron 1977; Lightfoot 1979; Niyogi & Berwick 1996; Kroch 2001; Yang 2002; Cournane 2017: *inter alia*), and some programs which do not privilege child language acquisition still acknowledge a special role for children (Labov 1989), though there are also prominent dissenters (Croft 2000; Meisel 2011; Diessel 2012: *inter alia*). Children of the past must have acquired language in a way similar to modern children (a straightforward consequence of linguistic uniformitarianism (Labov 1972; Walkden 2019)) so the obvious obstacle to investigating the relationship between acquisition and change, whether or not the position is empirically supported, is more practical than theoretical: it is hampered by the lack of access to children of the past.

This study extends the previous analysis back through time to compare the contents of modern CDS-derived and (frequency-trimmed when applicable) historical lexicons. Since linguistic properties like the presence of "irregular" inflection are not conserved across languages, this study compares the meanings contained in each lexicon instead. Items are matched between two lexicons if there is a shared translation between them. For example, English *slide* is matched with Spanish *resbalar* 'slip,' Latin *lābī* 'slip, glide,' and Proto-Germanic **slīdaną* 'slide.'[4] Since correspondences between the lexicons are no longer one-to-one, Jaccard similarity does not make sense here. Instead, raw percent overlap is calculated as $|A \cap B|/\min(|A|,|B|)$. Overlaps are systematically higher than Jaccard similarities because the denominator is smaller.

English CDS (Brown) and Spanish CDS from CHILDES (FernAguado, Hess, OreaPine, Remedi, Romero, and SerraSole (Romero et al. 1992; Hess Zimmermann 2003; MacWhinney 2000; Aguado-Orea & Pine 2015)) are compared to two pre-modern lexicons: Latin from all Old and Classical texts in the Perseus online edition (Smith et al. 2000), and Proto-Germanic (PGmc) taken from all securely reconstructable strong verbs in Seebold (1970).[5] The Proto-Germanic strong verbs are chosen because they are not semantically coherent and provide a sufficiently large set for comparison, and frequency cutoffs are established for each corpus-derived lexicon to bring them in line with the size of PGmc. To establish a within-language CDS baseline, the overlap procedure was performed between the Brown

[4] A full list of correspondences is provided in the supplementary material.

[5] I thank Donald Ringe for his help in sorting through Seebold.

and Brent corpora with the same frequency cutoff applied to both, and Brown and Spanish were compared as a cross-language CDS baseline.

Table 1: Modern CDS, historical and prehistoric lexicon raw percent overlaps..

Comparison	Sizes	Overlap
EN Brown-EN Brent	260/257	81.71%
English-Spanish	260/263	73.07%
English-PGmc	260/258	66.67%
Spanish-PGmc	263/258	71.32%
English-Latin	260/260	75.77%
Spanish-Latin	263/260	79.62%

Table 1 reveals a spread of about 15 points between lowest and highest raw percent overlap scores. The within-language English-English baseline is the highest at about 82%, while the cross-language CDS baseline is somewhat lower at 73%. The Latin comparisons are higher than the CDS baseline, while the Proto-Germanic numbers are a few points lower. The high overlap between the reconstructed and modern lexicons is likely due to the fact that words are securely reconstructable only if they are retained in multiple daughter branches, and that the words that are likely to be retained tend to be frequent everyday terms – the same kind that we expect to find in CDS. For example, the Proto-Germanic words for 'bite,' 'wait,' 'fall,' 'pull,' 'sing,' and 'help' are reconstructable because they were retained in its daughters, and their equivalents are all present in both the modern English CDS corpora since they are common everyday terms.[6]

It seems that cultural differences account for the extra discrepancy between Proto-Germanic and CDS. The PGmc lexicon contains many terms for farming ('sow,' 'plant,' 'thresh'), household chores ('weave,' 'knead,' 'bank a fire') and other aspects of culture ('cast lots,' 'be a retainer') that modern urban children are unlikely to know, but which children growing up in Iron Age agricultural societies must have. These cultural terms account for roughly 3.1 points of overlap, which when added in would put the PGmc comparisons in line with the English-Spanish overlap.

All in all, lexical overlap is conserved between CDS, adult historical corpora, and reconstructed lexicons about as well as between CDS lexicons. They contain largely the same kinds of meanings despite their varied origins, and differences between lexicons can be partially account for by cultural differences rather than corpus differences. The lists collated in the supplementary material show that higher frequency items are more likely to match than low frequency items, even among

[6] The supplementary material contains a full list of examples.

different CDS corpora. This reiterates the point from Section 2 that low token frequency items are more likely to be corpus-specific than high-frequency items.

4 Deploying an acquisition model

This study compares outcomes when a learning algorithm is applied to CDS, modern non-CDS and historical corpora. First, I compare the acquisition of Modern English productive past -*ed* on lexicons sourced from CDS and adult corpora. Following that, I apply the same algorithm to a past tense generalization in Old and Modern Icelandic to draw conclusions about child development in the past.

In both cases, I apply the Tolerance Principle (TP) following Yang (2016: ch. 4.1), though any type-based acquisition model could be used here. The TP stated in (1) is a model of productivity learning that defines a threshold θ for how many exceptional types a hypothesized grammatical pattern can tolerate before it becomes untenable and the learner resorts to memorization and listing instead. The threshold is derived such that it lies at the point where it becomes more economical for a language user to learn a pattern plus exceptions rather than no pattern (Yang 2016: pp. 48-51).

(1) **Tolerance Principle**:
If R is a productive rule applicable to N candidates, then the following relation holds between N and e, the number of exceptions that could but do not follow R:

$$e \leq \theta_N \text{ where } \theta_N := \frac{N}{\ln N}$$

4.1 Modern English Past +ed

To investigate whether CDS-derived and adult-derived lexicons yield similar learning outcomes, I model the acquisition of the English productive past-forming -*ed* pattern. The acquisition of English past tense is a complex and classic problem in morphological learning which has triggered decades of debate (Berko 1958; Rumelhart & McClelland 1986; Pinker & Prince 1988; 1994; Ramscar 2002; Kirov & Cotterell 2018: *inter alia*), and the acquisition of a default past -*ed* pattern is one piece of the challenge. In terms of the Tolerance Principle, the pattern being acquired is one that applies -*ed* (with the appropriate morphophonology) to a verb to produce its past tense form. All verb types learned up to a given point in development count towards the N in the formula, while those verb types learned with irregular pasts by that point make up e. Specific lexical items do not matter in the calculation, nor do the values of e and N past establishing whether or not e lies below the tolerance threshold.

Yang (2016) finds that the English lexicon is such that early learners who know 500 or fewer verbs know too many irregulars relative to regularly derived past verbs to learn +*ed* productively. The situation is marginal at 800, and learners can finally reliably acquire the productive past once they know 1,000 verbs.

I reproduce these results. 1,000 CDS-derived lexicons with 1,000 items each are sampled from the 1,515 unique lemmas attested together in English Brent, Brown, or MacWhinney weighted by their token frequencies across those corpora, then the same sampling is performed on the 1,500 most common COCA lemmas to create 1,000 sample adult-derived lexicons. The TP is calculated on each lexicon for the top N=100, 150, and 200 through 1,000 items. For all CDS-derived and adult-derived lexicons, the results at $N = 100$, 500, and 1,000 are identical to what Yang (2016) reports for both sample types: every lexicon fail to generalize past -*ed* at low N but succeed by $N = 1,000$ as shown in Table 2. On its own, this TP calculation would imply that a learner would not acquire a productive past -*ed* until knowing near 1,000 unique verbs' past forms, but see Yang (2016: ch. 4.1.2)

Table 2: Percent of sampled corpora generalizing +*ed* by the TP ($e < N/\ln N$).

Corpus Type	**100**	**300**	**500**	**800**	**1,000**
CDS Samples	0%	0%	0%	0.02%	**100%**
Adult Samples	0%	0%	0%	68.4%	**100%**
Yang 2016	no	no	no	no	**yes**

What differences do exist cluster around the $N = 800$ point that Yang reports as marginally non-productive. When plotted in Figure 5, we see that the adult corpus learning curve is shifted somewhat to the left, which reflects the slightly lower average number of irregular verbs in the adult-derived lexicons at that point (117 vs. 127). This is effectively a sample-dependent relative developmental delay of the kind reported in Maratsos (2000) and Yang (2002). Regardless, the final learning state is identical for every single adult and CDS sample.

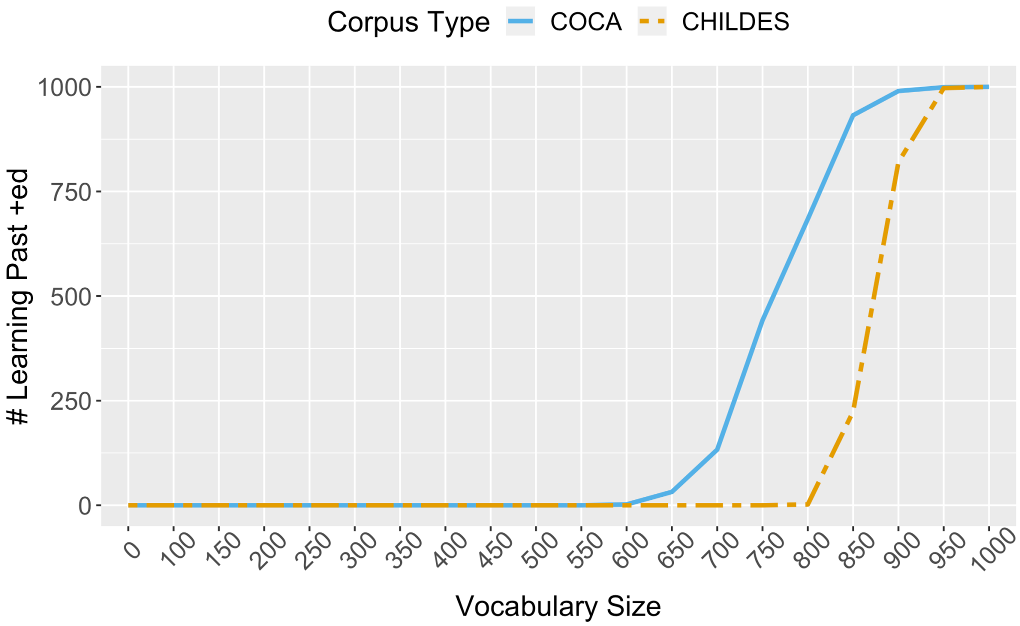

Figure 5: Proportion of learners acquiring productive +*ed* past by vocabulary size.

4.2 Icelandic Strong Verbs

Finally, I apply the Tolerance Principle to a problem in both Old and Modern Icelandic to compare modern and historical learning trajectories. The remarkable diachronic stability of Icelandic morphology renders it uniquely suitable for this study since it allows us to set up a null hypothesis: patterns that emerge among the highest frequency items in a Modern Icelandic text should be apparent in a Modern Icelandic text as well. We could run the same test on Old English or Latin, but we would have no hypothesis to test since their modern descendents are so different.

Icelandic, like English, has a significant number of verbs that express past tense by stem mutation (so-called *strong verbs*, e.g., *syngja* ∼ *söng* 'sing'), and a much larger number which express the past through suffixation (multiple classes of *weak verbs*, e.g., *dvelja* ∼ *dvaldi* 'dwell, reside,' *svara* ∼ *svaraði* 'answer, respond'). It is up to child learners to sort out whether any patterns exist over these verbs that indicate which type of inflection to productively employ. This turns out to be quite challenging – even eight-year-old Icelandic children still make a non-trivial number of errors in which they substitute one class for another (Ragnarsdóttir et al. 1999).

I consider one such generalization that illustrates this pattern of learning: the relationship between monosyllabic verbs (e.g, *dá* 'adore, worship,' *ná* 'get, obtain,' *sjá* ‘see, perceive’) and strong inflection. Most verbs in this set are weak, but a few are strong as well, so a learner has to determine which ones belong to the productive pattern, if any, and which should be learned as exceptions. To investigate this quantitatively, I extract all verbs which are attested at least once in the past tense from the Old Icelandic and Modern Icelandic texts in the POS-tagged and lemmatized Icelandic Parsed Historical Corpus (IcePaHC; Wallenberg et al. 2011), which results in two sets of 735 and 921 unique verb types respectively. Next, I apply the same sampling procedure as in the above section to generate 1,000 sample lexicons from each era to model the learning trajectories of “typical” learners exposed to these verbs in their input. The resulting developmental trajectories are presented in Figure 6.

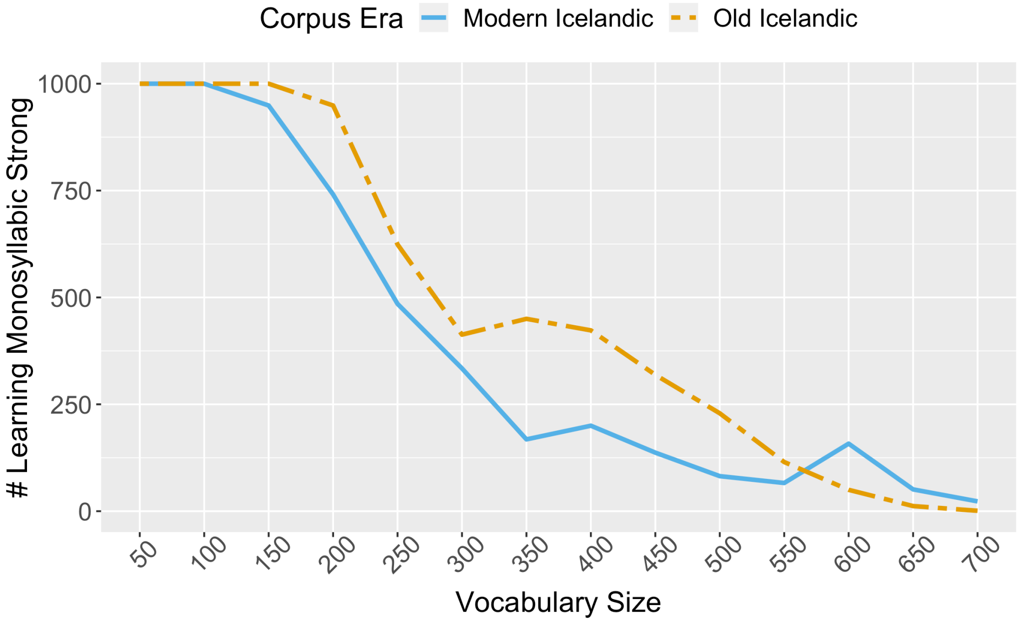

Figure 6: Proportion of learners acquiring productive strong inflection for verbs with monosyllabic stems.

There are two takeaways here. First, the average learning trajectories are closely matched between the Old Icelandic and Modern Icelandic learners, which confirms our expectations of morphological conservatism in Icelandic and once again demonstrates the insignificance of genre differences when it comes to the type expression of linguistic properties. Second, it shows that all early learners with small vocabularies can productively apply strong verb inflection to monosyllabic verbs, but that they gradually lose this option as their vocabularies grow and monosyllabic strong verbs are revealed to be the true exceptions. This pattern of early spurious productivity is consistent with the widely observed tendency for “irregulars” (here, strong verbs) to cluster among high token frequency items (Bybee 1985; Baayen 1993; Yang 2016: *inter alia*). It drives modern learners to tenable but ultimately incorrect hypotheses about their languages (e.g., Xu & Pinker 1995; Ragnarsdóttir et al. 1999), and now we can say that it did so for Icelandic learners of the past too.

5 Conclusions

The studies presented here identify substantial similarities between corpora of child-directed speech and both modern and historical adult corpora as well as a reconstructed lexicon. When lexicons derived from child-directed speech and non-CDS corpora are trimmed by token frequency in order to approximate child lexicon sizes, they express type frequencies to a degree that is statistically similar to those in CDS corpora. Since it is these type frequencies that are critical for the acquisition of linguistic generalizations, non-CDS corpora can be used to model aspects of child language learning. These results open up a path for researchers who wish empirically evaluating the relationship between acquisition and change and gives reason to investigate what other relationships may hold between child directed and non-child directed corpora.

Funding information

This work was funded by an NDSEG fellowship awarded to the author by the U.S. Army Research Office.

Acknowledgements

I would like to thank Charles Yang, Mitch Marcus, and Don Ringe along with anonymous *Glossa* reviewers for their insights and suggestions.

Competing interests

The author has no competing interests to declare.

References

Aguado-Orea, Javier & Julian M Pine. 2015. Comparing different models of the development of verb inflection in early child spanish. *PLOS one* 10(3). e0119613.

Anderson, James Maxwell. 1973. *Structural aspects of language change*, vol. 13. Longman.

Aronoff, Mark. 1976. Word formation in generative grammar. *Linguistic Inquiry Monographs Cambridge, Mass.* (1). 1–134.

Baayen, Harald. 1993. On frequency, transparency and productivity. In *Yearbook of morphology 1992*, 181–208. Springer.

Baker, Carl L. 1979. Syntactic theory and the projection problem. *Linguistic Inquiry* 10(4). 533–581.

Baron, Naomi S. 1977. *Language acquisition and historical change*, vol. 36. North-Holland Amsterdam.

Berko, Jean. 1958. The child's learning of english morphology. *Word* 14(2-3). 150–177.

Brent, Michael R & Jeffrey Mark Siskind. 2001. The role of exposure to isolated words in early vocabulary development. *Cognition* 81(2). B33–B44.

Brown, Roger. 1973. *A first language: The early stages.* Harvard U. Press.

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form*, vol. 9. John Benjamins Publishing.

Chan, Erwin. 2008. *Structures and distributions in morphological learning*: Ph. D. dissertation, Dept. of Computer and Information Science, UPenn dissertation.

Cournane, Ailís. 2017. In defense of the child innovator. *Micro Change and Macro Change in Diachronic Syntax* 10–24.

Croft, William. 2000. *Explaining language change: An evolutionary approach.* Pearson Education.

Davies, Mark. 2009. The 385+ million word corpus of contemporary american english (1990–2008+): Design, architecture, and linguistic insights. *International journal of corpus linguistics* 14(2). 159–190.

Diessel, Holger. 2012. Language change and language acquisition. *Historical linguistics of English: An international handbook* 2. 1599–1613.

Elman, Jeffrey. 1998. Generalization, simple recurrent networks, and the emergence of structure. In *Proceedings of the twentieth annual conference of the cognitive science society*. 6. Mahwah, NJ: Lawrence Erlbaum Associates.

Finley, Sara. 2018. Cognitive and linguistic biases in morphology learning. *Wiley Interdisciplinary Reviews: Cognitive Science* 9(5). e1467.

Goodman, Judith C, Philip S Dale & Ping Li. 2008. Does frequency count? parental input and the acquisition of vocabulary. *Journal of child language* 35(3). 515–531.

Gropen, Jess, Steven Pinker, Michelle Hollander, Richard Goldberg & Ronald Wilson. 1989. The learnability and acquisition of the dative alternation in english. *Language* 203–257.

Halle, Morris. 1962. Phonology in generative grammar. *Word* 18(1-3). 54–72.

Hart, Betty & Todd R Risley. 1995. *Meaningful differences in the everyday experience of young american children.* Paul H Brookes Publishing.

Hart, Betty & Todd R Risley. 2003. The early catastrophe: The 30 million word gap by age 3. *American educator* 27(1). 4–9.

Hess Zimmermann, Karina. 2003. El desarrollo lingüístico en los años escolares: análisis de narraciones infantiles. *Unpublished PhD dissertation). El Colegio de México* .

Jaccard, Paul. 1901. Distribution de la flore alpine dans le bassin des dranses et dans quelques régions voisines. *Bull Soc Vaudoise Sci Nat* 37. 241–272.

Jarmulowicz, Linda D. 2002. English derivational suffix frequency and children's stress judgments. *Brain and Language* 81(1-3). 192–204.

Kiparsky, Paul. 1965. *Phonological change.*: Massachusetts Institute of Technology dissertation.

Kirov, Christo & Ryan Cotterell. 2018. Recurrent neural networks in linguistic theory: Revisiting pinker and prince (1988) and the past tense debate. *Transactions of the Association for Computational Linguistics* 6. 651–665.

Kroch, Anthony S. 2001. *Syntactic change*. na.

Labov, William. 1972. Some principles of linguistic methodology. *Language in society* 1(1). 97–120.

Labov, William. 1989. The child as linguistic historian. *Language variation and change* 1(1). 85–97.

Levin, Beth. 1993. *English verb classes and alternations: A preliminary investigation.* University of Chicago press.

Levin, Harry, Susan Long & Carole A Schaffer. 1981. The formality of the latinate lexicon in english. *Language and Speech* 24(2). 161–171.

Levin, Harry & Margaretta Novak. 1991. Frequencies of latinate and germanic words in english as determinants of formality. *Discourse Processes* 14(3). 389–398.

Lightfoot, David W. 1979. Principles of diachronic syntax. *Cambridge Studies in Linguistics London* 23.

Lignos, Constantine, Erwin Chan, Charles Yang & Mitchell P Marcus. 2010. Evidence for a morphological acquisition model from development data. In *Proceedings of the 34th annual boston university conference on language development*, vol. 2. 269–280.

MacWhinney, Brian. 1978. The acquisition of morphophonology. *Monographs of the society for research in child development* 1–123.

MacWhinney, Brian. 1991. The childes language project: Tools for analyzing talk. *Hillsdale, NJ: Lawrence Erlbaum* 40. 62–74.

MacWhinney, Brian. 2000. *The childes project: The database*, vol. 2. Psychology Press.

Maratsos, Michael. 2000. More overregularizations after all: new data and discussion on marcus, pinker, ullman, hollander, rosen & xu. *Journal of Child Language* 27(1). 183–212.

Meisel, Jürgen M. 2011. Bilingual language acquisition and theories of diachronic change: Bilingualism as cause and effect of grammatical change. *Bilingualism: Language and Cognition* 14(2). 121–145.

Nagy, William E & Richard C Anderson. 1984. How many words are there in printed school english? *Reading research quarterly* 304–330.

Niyogi, Partha & Robert C Berwick. 1996. A language learning model for finite parameter spaces. *Cognition* 61(1). 161–193.

Pierrehumbert, Janet B. 2003. On frequency, transparency and productivity. In *Probabilistic linguistics*, 177–228. MIT Press.

Pinker, Steven, David S Lebeaux & Loren Ann Frost. 1987. Productivity and constraints in the acquisition of the passive. *Cognition* 26(3). 195–267.

Pinker, Steven & Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition* 28(1-2). 73–193.

Pinker, Steven & Alan Prince. 1994. Regular and irregular morphology and the psychological status of rules of grammar. *The reality of linguistic rules* 321. 51.

Pinker, Steven & Michael T Ullman. 2002. The past and future of the past tense. *Trends in cognitive sciences* 6(11). 456–463.

Ragnarsdóttir, Hrafnhildur, Hanne Gram Simonsen & Kim Plunkett. 1999. The acquisition of past tense morphology in icelandic and norwegian children: An experimental study. *Journal of child language* 26(3). 577–618.

Ramscar, Michael. 2002. The role of meaning in inflection: Why the past tense does not require a rule. *Cognitive Psychology* 45(1). 45–94.

Rappaport Hovav, Malka & Beth Levin. 2008. The english dative alternation: The case for verb sensitivity. *Journal of Linguistics* 44(01). 129–167.

Romero, Silvia, Aida Santos & Dora Pellicer. 1992. The construction of communicative competence in mexican spanish speaking children (6 months to 7 years). *Mexico City: University of the Americas* .

Rumelhart, David E & James L McClelland. 1986. Learning the past tenses of english verbs: Implicit rules or parallel distributed processing. In Rumelhart David E McClelland, James L & the PDP Research Group (eds.), *Parallel distributed processing: Explorations in the microstructure of cognition, vol. 2: Psychological and biological models.*, .

Seebold, Elmar. 1970. *Vergleichendes und etymologisches wörterbuch der germanischen starken verben*, vol. 85. Walter de Gruyter.

Smith, David A, Jeffrey A Rydberg-Cox & Gregory R Crane. 2000. The perseus project: A digital library for the humanities. *Literary and Linguistic Computing*

15(1). 15–25.

Sweet, Henry. 1899. *The practical study of languages*. London: Dent.

Szagun, Gisela, Claudia Steinbrink, Melanie Franik & Barbara Stumper. 2006. Development of vocabulary and grammar in young german-speaking children assessed with a german language development inventory. *First language* 26(3). 259–280.

Tyler, Andrea & William Nagy. 1989. The acquisition of english derivational morphology. *Journal of memory and language* 28(6). 649–667.

Walkden, George. 2019. The many faces of uniformitarianism in linguistics. *Glossa: a journal of general linguistics* 4(1).

Wallenberg, Joel C, Anton Karl Ingason, Einar Freyr Sigurdhsson & Eirikur Roegnvaldsson. 2011. Icelandic parsed historical corpus (icepahc). *Version 0.9. Size* 1.

Xu, Fei & Steven Pinker. 1995. Weird past tense forms. *Journal of child language* 22(3). 531–556.

Yang, Charles. 2002. *Knowledge and learning in natural language*. Oxford University Press on Demand.

Yang, Charles. 2016. *The price of linguistic productivity*. Cambridge, MA: MIT Press.