Children Face Ambiguous Input Early On: A Cross-Linguistic Corpus Analysis of Homophones

Youtao Lu and Yukie Nagai

1. Introduction

Human languages can be ambiguous at multiple levels, yet children are able to resolve many forms of ambiguity early in development. For example, by six months of age, infants begin to categorically perceive phonemes in their native languages, even when they show substantial overlap acoustically (e.g., Eimas et al., 1971; Kuhl et al., 1992). This sensitivity foreshadows children's ability to represent familiar words with fine-grained phonemic detail by around 14 months of age (e.g., Swingley & Aslin, 2002), a capacity that crucially contributes to the rapid expansion of vocabulary observed at this age (e.g., Bergelson, 2020).

However, the ability to discriminate phonemes does not directly extend to learning novel words that differ from familiar words by only a single phoneme, such as *tog*. Studies have shown that such words, known as *phonological neighbors*, still pose substantial challenges to 18-month-old English-learning children (Swingley & Aslin, 2007). This difficulty is interpreted as reflecting uncertainty about whether a subtle phonological difference signals a lexical contrast or merely a non-contrastive deviation in pronunciation, rather than a limitation in perceptual discrimination. Given the sparsity of the phonological space occupied by their vocabulary, children may expect word forms to be rather distinctive. Therefore, novel phonological neighbors tend to be interpreted as atypical exemplars of familiar words rather than as novel lexical items.

Under this interpretation, phonological neighbors may be regarded by children at this stage as functionally analogous to homophones, words with identical phonological forms but unrelated meanings. As an extreme form of lexical ambiguity, homophones violate the idealized one-to-one mapping between form and meaning, an assumption proposed to guide children's expectations in early language development (Slobin, 1985). Learning homophones has been found to be challenging even for children above four years of age: although they

* Youtao Lu, International Research Center for Neurointelligence (WPI-IRCN), UTIAS, The University of Tokyo, lu.youtao@mail.u-tokyo.ac.jp. Yukie Nagai, International Research Center for Neurointelligence (WPI-IRCN), UTIAS, The University of Tokyo. This work was supported by the World Premier International Research Center Initiative (WPI), MEXT, Japan, and by the Japan Society for the Promotion of Science (JSPS) through a Postdoctoral Fellowship for Research in Japan (No. P24004) and the associated Grant-in-Aid for JSPS Fellows (KAKENHI No. 24KF0103). We thank Devesh Persaud and Julie Nakamura for their assistance with data processing and inspection.

 Proceedings of the 50th annual Boston University Conference on Language Development, ed. Romi Hill et al., 453-466. Somerville, MA: Cascadilla Press.

can successfully learn mappings between novel objects and licit phonological forms in general, they struggle when the form has already been associated with a familiar referent (e.g., Doherty, 2004).

On the surface, homophones would appear to pose challenges for the use of spoken language in general, as they cannot be accurately interpreted based on the auditory signal alone. Nevertheless, homophones are widespread across human languages and exhibit systematic distributional patterns: they tend to be associated with shorter, more frequent, and phonotactically better-formed phonological forms (Piantadosi et al., 2012). These patterns have been explained from the perspective of communicative efficiency. By reusing phonological forms, homophones reduce the memory and articulatory demands for speakers, and assigning easily produced forms to multiple meanings maximizes this benefit. For listeners, the cost of disambiguation is often mitigated by contextual information and is therefore outweighed by these efficiency gains.

It is perhaps unsurprising that adults, as proficient language users, can efficiently exploit linguistic context to disambiguate homophones (for a review, see Simpson, 1994; most evidence comes from studies of visually presented homophones that also have identical orthographic forms). Importantly, previous studies have also shown that providing sufficient linguistic contextual information that clearly distinguishes a novel word from a familiar homophonous counterpart enables homophone learning (Casenhiser, 2005; Dautriche et al., 2018). Children as young as 19 months of age can associate a novel object with a familiar verb when it is presented in noun contexts, thereby learning a novel noun homophone of that verb. They can also learn a novel noun homophone of a familiar noun, provided that the corresponding referents are sufficiently dissimilar and that these distinctions are made explicit during teaching. Dautriche and colleagues also conducted a preliminary investigation of the availability of such contextual cues. Their corpus analyses of Wikipedia passages in Dutch, English, French, and German showed that homophones are more likely to belong to different syntactic categories or to have distinctive semantic representations. It should be noted, however, that the sophisticated written language analyzed in this study differs substantially from the language children hear. Specifically, child-directed speech relies on a relatively limited vocabulary composed largely of short, high-frequency words (Jones et al., 2023), which, as discussed previously, are more likely to be homophones. Moreover, child-directed speech typically consists of shorter utterances with simpler syntactic structures (e.g., Phillips, 1973), which may limit the availability of linguistic context despite the aid it provides to children's learning of homophones.

Given that children find it difficult to learn homophones without sufficient support from linguistic context, it appears paradoxical that they are exposed to words that are highly likely to be homophonous, often in context that provides limited cues for disambiguation. However, the limited vocabulary of child-directed speech also means that within a homophone set, not all items are presented to children at the same time. Rather, they may appear at very different stages of language development or with markedly different frequencies. Such separations can substantially reduce the number of homophones that children

actually encounter. They may even alter the distributional properties of homophones as observed in general language use, such that the number of homophonous words associated with a given phonological form no longer correlates with its frequency, length, or phonotactic well-formedness.

One final consideration in the present study concerns cross-linguistic differences in the prevalence of homophones, which persist despite the shared properties discussed above. For example, the prevalence of homophones differs across English, Dutch, and German, with English exhibiting more homophones than Dutch, and Dutch more than German (Piantadosi et al., 2012). Specifically, Japanese was found to exhibit an exceptionally high prevalence of homophones, potentially as a consequence of its restricted phonological system (Lu, 2022). Moreover, when the number of homophones in observed vocabularies was compared with that in randomly simulated, phonology-based vocabularies cross-linguistically, only Japanese showed no robust evidence of homophone avoidance (Trott & Bergen, 2020). These differences raise the possibility that homophone learning poses a heightened challenge for children acquiring homophone-rich languages such as Japanese, and it remains an open question whether it can be mitigated by richer context, greater separation among homophones, or both.

The foregoing discussion suggests that children's difficulty in learning homophones may, in principle, be mitigated in two ways: homophones may be systematically avoided in children's language input, and linguistic context may support successful disambiguation when homophones are encountered. The present study examines these two possibilities in children's early language input and further investigates how they vary across languages with different prevalence of homophones.

2. Method

Using data from the CHILDES database (MacWhinney, 2000), we conducted corpus analyses of mother's speech to children aged 12 to 35 months in English, French, and Japanese. Within each language, we quantified the distributional patterns of homophones with respect to the frequency and length of their phonological forms. These properties may deviate from those observed in general language use as a consequence of homophone avoidance. We then tested for direct evidence of homophone avoidance by comparing differences in frequency and age of occurrence among members of homophone sets. Contextual support was similarly assessed by quantifying the degree of overlap in syntactic categories among items within homophone sets. Across languages, we examined how the prevalence of homophones correlates with these within-language measures to understand if homophones are further separated or more strongly supported by context in homophone-rich languages.

2.1. Data

The analyses utilized four longitudinal corpora of monolingual children: one in English (the Providence Corpus, Demuth et al., 2006), one in French (the Lyon

Corpus, Demuth & Tremblay, 2008), and two in Japanese (the MiiPro Corpus and the Miyata Corpus, Miyata, 2012). We selected 12 months as the starting age and 35 months as the ending age for the analyses. While data from all child-mother dyads extended to the ending point, the starting point varied across dyads, with the latest beginning at 18 months. We nevertheless retained this uneven starting point because the evidence that children between 19 and 20 months of age can learn homophones with sufficient support from linguistic context suggests that relevant language input is likely to have begun earlier, particularly in homophone-rich languages such as Japanese.

These corpora were selected for their comparability along several dimensions. All corpora consist of transcriptions of extensive video or audio recordings of spontaneous mother–child interactions in naturalistic home environments. Recording sessions lasted approximately one hour and were held at least monthly within the selected age range, without skipped sessions. No instructions regarding the events or topics of the interactions were provided, although mothers were sometimes encouraged in advance to elicit children's utterances.

Since the involvement of recorders and other adults was limited, we did not restrict the analyses to maternal utterances immediately followed by children's responses, even though this approach could, in principle, more accurately restrict the analyses to child-directed speech. An additional consideration was that the corpora differed in their criteria for segmenting multi-utterance speech, and applying this filter would therefore have excluded valid data in a disproportionate and corpus-dependent manner.

Transcriptions of the English and French corpora were provided in the respective writing systems. Transcriptions of four dyads in the Japanese corpora were provided in kanji, the logographic writing system of Japanese, while those of three additional dyads were in romaji, an alphabetic transcription system that encodes pronunciation and is not used natively. The latter were excluded because romaji transcriptions are pronunciation-based and do not preserve the distinctions required to identify homophones. The final dataset comprised speech from six English dyads, five French dyads, and four Japanese dyads.

2.2. Processing

Data were processed separately for each language. Individual word tokens were extracted from pre-segmented transcriptions of maternal speech using CLAN (MacWhinney, 2000). Tokens were associated with the annotated syntactic categories in which they appeared, and no cross-linguistic normalization of syntactic categories was performed. Word tokens were the grouped into lexical entries based on orthographic form and annotated stem. Occurrences of the same word across ages and dyads were collapsed at this stage. Tokens corresponding to proper nouns, interjections, and unidentifiable usages (e.g., fragments and imitation of children's vocalizations) were excluded from further analyses.

For the purposes of the present study, stem annotations were revised so that usages of the same word across different syntactic categories shared a single stem. These revisions were validated by assistant coders who are native speakers of the

respective languages. Any modifications arising from manual inspection were discussed between the coders and the authors, and discrepancies were resolved through discussion. During this process, we identified substantial inconsistencies in the annotation of a set of Japanese words transcribed in hiragana, the phonologically based writing system of Japanese. Due to both the relatively high token frequency of these words and the opacity of the existing annotations, comprehensive manual correction was not feasible. These items, including all single hiragana word forms except *da* and *de*, as well as five word forms consisting of two hiragana units (*nda*, *an*, *nto*, *beta*, *kai*), were therefore excluded from subsequent analyses. Words with the same pronunciation that were transcribed in kanji were unaffected. Since all excluded items were annotated with multiple stems and should be regarded as homophones, their exclusion is expected to result in a slight underestimation of homophone prevalence in Japanese.

For each lexical entry, summary measures, including token frequency, the syntactic categories in which it occurred, and the earliest age at which the word appeared in children's input, were recorded. Children's age in months was directly extracted from transcript metadata. Lexical entries were then linked to their phonological forms using the CALLHOME lexicons for English and Japanese (Kingsbury et al., 1997; Kobayashi et al., 1996) and the *Lexique* database for French (New et al., 2004). Entries for which no phonological form could be identified were excluded. Manual inspection was conducted when a single lexical entry could be linked to multiple phonological forms. In cases of free variation, the most common form, as determined by coders, was retained. When different forms corresponded to distinct meanings (e.g., *DEsert* vs. *deSERT* in English), or when pronunciation was context dependent (as is common in Japanese), coders examined the original utterance to determine the appropriate phonological form.

Lexical entries were subsequently grouped by phonological form, and another round of manual inspection was conducted to confirm that homophonous entries corresponded to distinct meanings. Entries associated with highly related meanings, typically reflecting inflectional variants, were combined. Stress patterns annotated in the English lexicon were treated as phonological distinctions. Although pitch-accent patterns in Japanese can similarly differentiate words with identical segmental forms, such information was not available in the Japanese lexicon and was therefore not considered, potentially leading to an overestimation of homophone prevalence in Japanese. Finally, because Japanese employs multiple writing systems, the same word may be transcribed using different orthographic forms. Therefore, after grouping lexical entries by phonological form, we identified entries that shared both phonological form and stem and merged those that represented different orthographic realizations of the same word.

Our analyses are based on measures computed at the level of individual phonological forms. Details of these measures and analyses are provided in the next two sections.

2.3. Measures

The first three measures are taken for all phonological forms.

The *number of homophones* was defined as the number of homophonous lexical entries minus one, following Piantadosi et al. (2012). This measure thus corresponds to the number of additional meanings associated with a given phonological form. Its distribution can be appropriately analyzed using quasi-Poisson regression models, consistent with the analytical approach adopted in the previous study.

Form Length was measured as the number of syllables in each phonological form. In French, syllable counts were obtained directly from the *Lexique* database. In English, syllable counts were derived from stress markers annotated for each syllable in the phonological forms provided by the CALLHOME lexicon. In Japanese, form length was estimated by counting the number of vowels in the phonological transcription, with each lengthened vowel treated as a single vowel, despite being transcribed as two identical consecutive vowels. However, the phonological transcriptions provided in the CALLHOME lexicon do not distinguish lengthened vowels from sequences of identical vowels spanning two syllables. As a result, syllable counts were underestimated in a small number of cases. Form length was standardized within each language by subtracting the language-specific mean from the value of each phonological form and dividing by the corresponding standard deviation across all phonological forms.

Form Frequency was operationalized as the negative log probability. For each phonological form, token frequencies of all associated lexical entries were summed and divided by the total token frequency of all lexical entries in the language, and the negative logarithm of the resulting proportion was taken. Forms with higher frequency thus correspond to lower negative log probabilities. Negative log probabilities were standardized within each language by subtracting the language-specific mean and dividing by the corresponding standard deviation across all phonological forms.

The following three measures were computed for phonological forms associated with more than one lexical entry. Each measure was calculated pairwise, and when more than two lexical entries were associated with a given phonological form, the mean of all pairwise comparisons was used.

The *difference in frequency* between two lexical entries was calculated as the absolute difference between their raw token frequencies divided by the sum of their raw token frequencies.

The *difference in age of occurrence* between two lexical entries was calculated as the absolute difference between their earliest age of occurrence, measured in months.

The o*verlap in syntactic categories* between two lexical entries was calculated as the number of shared syntactic categories divided by the total number of syntactic categories attested for the two entries in the corpora. All attested categories were weighted equally.

2.4. Analyses

Analyses of measures computed for all phonological forms followed the approach of Piantadosi et al. (2012). A quasi-Poisson regression model was fitted

using the *lme4* package (ver. 1.1.35.1; Bates et al., 2015) in R (R Core Team, 2021). The number of homophones was specified as the dependent variable, with form length, form frequency, and language included as fixed effects. All two-way interactions and the three-way interaction were also included. The language factor was re-leveled to obtain coefficient estimates within each language.

Analyses of measures computed within homophone sets followed the approach of Dautriche et al. (2018). We compared measures computed from the observed homophone sets to those obtained from simulated sets constructed under randomness. In the simulations, both the number of homophone sets and the size of each set were held constant to match the observed data, and lexical entries from the observed homophone sets were randomly shuffled to construct simulated sets. This procedure was repeated 1,000 times. For the observed data and each simulated dataset, measures were averaged across homophone sets. The mean obtained from the observed data was treated as a single observation and, under the null hypothesis of random grouping of lexical entries given the number and sizes of homophone sets, was expected to fall within the distribution of means from the 1,000 simulated datasets. Because this distribution approximated normality by the central limit theorem, we calculated a z-score for each measure in each language by subtracting the mean of the simulated distribution from the observed mean and dividing the difference by the standard deviation of the simulated distribution. The corresponding p-values were used to assess whether the observed homophone sets deviated from the simulated baseline.

In addition to comparisons with random baselines, we conducted pairwise cross-linguistic comparisons of measures computed within homophone sets to examine whether they are modulated by homophone prevalence. Wilcoxon rank-sum tests were used, as the measures did not necessarily approximate normality.

3. Results

Our data processing yielded 9,514 phonological forms in English, 5,069 in French, and 2,736 in Japanese, all of which were included in further analyses. Among these phonological forms, 172 in English (1.81%), 174 in French (3.43%), and 181 in Japanese (6.62%) were associated with more than one lexical entry, thus corresponding to homophone sets. This cross-linguistic difference in the proportion of homophone sets, together with differences in their sizes (i.e., the number of homophonous lexical entries), resulted in different mean numbers of homophones, with each phonological form on average associated with 0.019 additional meanings in English, 0.043 in French, and 0.095 in Japanese.

The quasi-Poisson regression analysis revealed significant coefficients for form length in all three languages (English: $\beta = -1.66$, $p < .001$; French: $\beta = -2.05$, $p < .001$; Japanese: $\beta = -1.44$, $p < .001$), indicating that shorter phonological forms tend to be associated with more lexical entries. Coefficients for form frequency were also significant (English: $\beta = -.60$, $p < .001$; French: $\beta = -.65$, $p = .002$; Japanese: $\beta = -0.88$, $p < .001$), indicating that more frequent phonological forms tend to be associated with more lexical entries. As illustrated in Figure 1 using the

ggplot2 package (ver. 4.0.0; Wickham, 2016), both patterns replicate those reported for written language in Piantadosi et al., (2012).

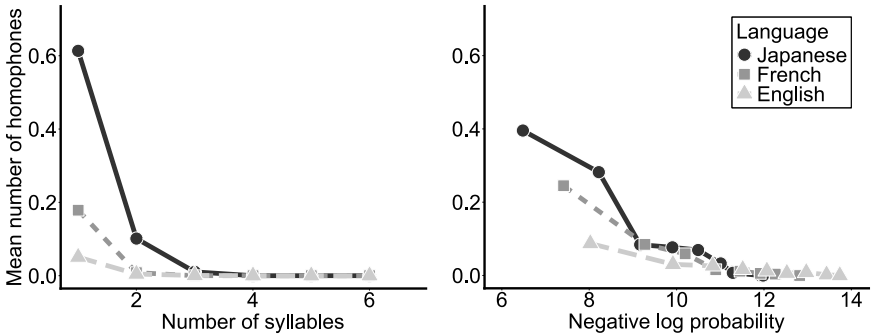

Figure 1. Number of homophones by form length and frequency. The y-axis shows the mean number of homophones associated with each phonological form. In the left panel, the x-axis represents form length (measured in the number of syllables); in the right panel, it represents form frequency (measured as negative log probability). Points in the left panel indicate means computed within natural bins defined by form length, whereas points in the right panel indicate means computed within ten equally sized frequency bins within each language. Languages are distinguished by point and line shape as well as by grayscale.

The regression analysis also revealed substantial differences across languages. The coefficient for language was significant when Japanese was contrasted with English ($\beta = 1.73$, $p < .001$) and with French ($\beta = 1.79$, $p < .001$), indicating that phonological forms in Japanese are associated with a greater number of lexical entries. In contrast, the coefficient was not significant when English and French were contrasted ($p = .85$). The interaction between form length and language was significant for the contrast between Japanese and French ($\beta = .61$, $p = .01$), indicating a stronger association between form length and the number of lexical entries in French. Finally, the interaction between form length and form frequency was significant in Japanese ($\beta = -.21$, $p < .001$) but not in English ($p = .74$) or French ($p = .75$), suggesting that the effects of these two predictors are not additive in Japanese but instead reinforce each other. The three-way interaction did not reach significance, providing limited evidence that this pattern is a robust cross-linguistic difference (Japanese vs. English: $p = .12$; Japanese vs. French: $p = .12$).

Comparisons of measures within homophone sets between observed values and random baselines showed that the observed differences and overlaps in general did not deviate from those expected under random grouping, as illustrated in Figure 2. The only significant deviation was observed in French, and it was opposite in direction to the hypothesized pattern of homophone avoidance: the difference in frequency within observed homophone sets was smaller than the simulated sets (*z-score* $= -2.56$, $p = .005$).

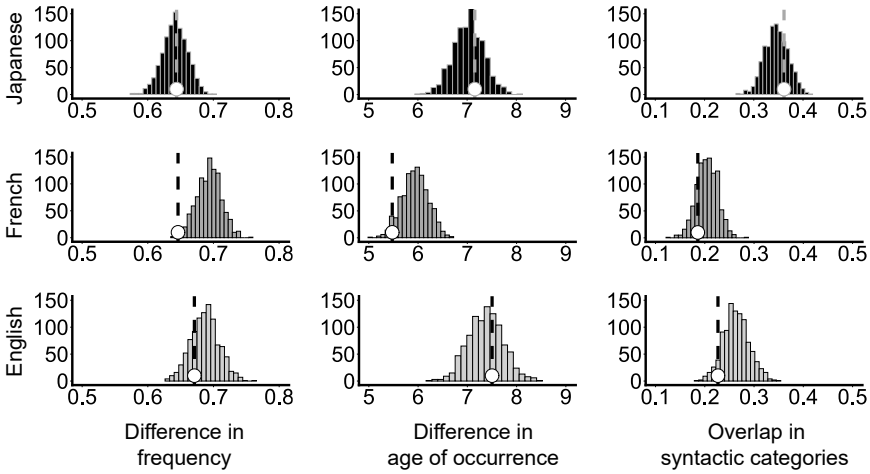

Figure 2. Observed means versus simulated mean distributions of measures within homophone sets.
Panels are arranged such that rows correspond to languages and columns correspond to measures. In each panel, the histogram shows the distribution of mean values obtained from 1,000 simulated homophone-set lists, serving as a random baseline for the corresponding language and measure. The point indicates the mean calculated from the observed homophone sets for the same language and measure.

Significant cross-linguistic differences were also identified in Wilcoxon rank-sum tests contrasting measures computed within homophone sets, as illustrated by the distributions shown in Figure 3. Ranked-biserial r values, which can be interpreted as effect-size estimates of these comparisons, were calculated using the *effectsize* package (ver. 1.0.1; Ben-Shachar et al., 2020) and are reported to aid interpretation. Specifically, homophone sets in Japanese exhibited greater overlap in syntactic categories than those in English ($W = 18{,}489$, $r = .19$, $p < .001$) or French ($W = 19{,}160$, $r = .22$, $p < .001$). Homophone sets in French showed smaller differences in the age of occurrences than those in English ($W = 12{,}116$, $r = -.23$, $p < .001$) or Japanese ($W = 11{,}807$, $r = -.21$, $p < .001$). Neither pattern fully aligns with the prediction that greater homophone prevalence should correspond to stronger separation by age of occurrence or syntactic context.

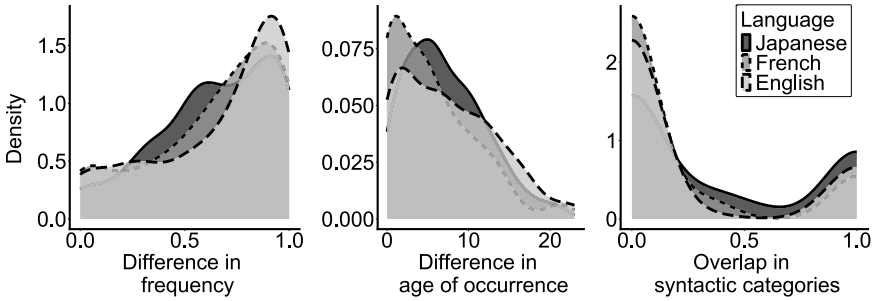

Figure 3. Cross-linguistic comparisons of measures within homophone sets. Each panel shows the distribution of the observed values for the corresponding measure within homophone sets, as labeled at the bottom. The y-axis represents kernel density estimates. Languages are distinguished by line shape and grayscale.

4. Discussion

In this cross-linguistic study of homophones in children's language input, we examined mothers' speech to children longitudinally from 12 to 35 months of age in English, French, and Japanese to address two questions: whether homophones are systematically avoided in children's language input, and whether their learning is supported by contextual cues. Our results showed that, relative to randomly constructed baselines, homophones in children's language input did not exhibit larger differences in frequency or age of occurrence, nor did they show reduced overlap in syntactic categories. Overall, we found little evidence for systematic homophone avoidance or contextual support. Instead, homophones in children's language input showed distributional properties similar to those reported for written language, such that shorter and more frequent phonological forms tend to be associated with more lexical entries (Piantadosi et al., 2012).

The cross-linguistic pattern of homophone prevalence in children's language input also resembles that reported for language use more generally (e.g., Lu, 2022; Trott & Bergen, 2020): homophones are more prevalent in Japanese than in French, and more prevalent in French than in English. We refrained from making direct comparisons of exact values across studies, which might otherwise suggest that homophones are less prevalent in children's language input, given the substantial differences in data sources and processing procedures, Instead, this resemblance in cross-linguistic pattern suggests that, even if some degree of homophone avoidance exists, it is likely a byproduct of the limited vocabulary used in child-directed speech, rather than an adaptation specifically tailored to mitigate the challenges of learning homophones.

This interpretation, that homophones arise without adaptation to mitigate learning difficulty, is further supported by our cross-linguistic comparisons of measures of homophone separation and disambiguation. Homophones in Japanese mothers' speech did not show larger differences in frequency or age of

occurrence than those in English and, contrary to expectations, exhibited greater overlap in the syntactic environments in which they occurred. This pattern again provides little evidence for systematic bias in the distribution of homophones in children's language input. The smaller differences in age of occurrence observed for French homophones appear to reflect a language-specific pattern. Our preliminary inspection suggests that this pattern may be driven by early-occurring homophonous function words, such as *a* ("has") and *à* ("to" or "at"). We plan to examine this possibility further by restricting future analyses to content words.

Thus, using maternal speech data, the present study did not replicate the previously reported syntactic gap between homophones (Dautriche et al., 2018). In addition to differences in data register (child-directed speech vs. written texts) and baseline construction (sampling within homophonous lexical entries vs. across all lexical entries), we note that the previous study did not explicitly restrict analyses to lexical entries with distinct meanings. As a result, homophonous conjugated forms of the same word may have contributed to an overestimation of syntactic differentiation among homophones.

The lack of evidence for homophone avoidance raises the original question: why do homophones persist in children's language input if the learning challenge they pose is not actively mitigated? One possibility is that homophones may also confer benefits that offset their costs. As suggested by Piantadosi et al. (2012), the reuse of phonological forms may reduce memorization demands. This advantage may be further amplified by the tendency for phonological forms of homophones in maternal speech to also be shorter and more frequent.

It should be acknowledged that this account still assumes that the cost of disambiguation can be reduced through contextual support. Although the present study did not find robust evidence for distinctive syntactic contexts in maternal speech, homophone learning may nevertheless be supported by semantic contexts. In addition, caregivers may provide non-linguistic communicative cues, such as pointing or demonstrating, to facilitate homophone learning. Importantly, this process does not require caregivers to be consciously aware of homophones in their speech. When children hear a familiar phonological form, they may actively search for its referent, which will be absent when a new meaning is intended in the case of homophones. Such shifts in children's attention may prompt caregivers to enrich their speech with additional cues and to repeat novel associations during more stable episodes of joint engagement, thereby supporting learning.

A second possibility is that, counterintuitively, a set of homophones may be easier to learn before any one of them has been fully established as a lexical association. Conceptually, this prevents an earlier or more frequently occurring association from acting as a strong inhibitor of the formation of a new association. This route of parallel learning for homophones has not yet been examined directly, but suggestive evidence can be found in studies of phonological neighborhoods (e.g., Swingley & Aslin, 2007) and minimal pairs (e.g., Yoshida et al., 2009). As discussed previously, English-learning children at 18 months show difficulty learning novel phonological neighbors of familiar words, whereas at 14 months they successfully learn pairs of neighboring novel words (e.g., *bin* and *din*). This contrast, while seemingly straightforward, should be interpreted with caution,

given substantial differences in experimental design across studies. We plan to pursue this possibility further through both meta-analyses of relevant work and empirical investigations of homophone learning under sequential versus parallel learning conditions.

The most substantial limitation of the present study concerns the opacity of word segmentation in Japanese. Whereas words in written French and English can be straightforwardly identified as units separated by spaces, no comparable convention for word segmentation exists in Japanese. This difference has two nontrivial consequences for the present analyses.

First, opacity in segmentation may have led to inconsistencies in word coding between the speech corpora and the phonological lexicon, resulting in reduced coverage of phonological forms in Japanese. Post hoc inspection indicated that phonological forms could be identified for over 80% of lexical entries in English (9,816/11,366) and French (7,326/8,995) prior to additional processing, but this proportion was considerably lower in Japanese (34.53%; 3,117/9,027). Such mismatches are not inherently related to homophony. However, words that are prone to segmentation mismatches are likely to be words with lower frequency or with more complex morphological structure, which are themselves less likely to be homophones. Their exclusion could therefore lead to an overestimation of homophone prevalence in Japanese. Conversely, excluded words may have been homophonous with included words that were treated as non-homophones, thus biasing estimates of homophone prevalence in the opposite direction. Regardless of the direction of this influence, such mismatches may also affect assessments of frequency differences and syntactic overlap within homophone sets.

Second, the opacity of segmentation in Japanese, together with properties of Japanese grammar, introduced systematic cross-linguistic inconsistencies. For example, the present progressive in Japanese is expressed by combining a verb in its *-te* form with the auxiliary verb *iru*, which can also function as an independent verb meaning “exist”. Annotators therefore reasonably treated the main verb and the auxiliary verb as separate items, whereas the corresponding expression in English is typically coded as a single item, without separating the main verb and the *-ing* marker. As a result, morphosyntactic constructions that are functionally comparable across languages were represented differently at the lexical level, potentially limiting the cross-linguistic comparability of the present analyses.

As noted in the data processing procedure, the present analyses were further constrained by limitations in the available annotations, including inconsistencies in stem annotation and the absence of pitch accent information for Japanese. These limitations introduce potential biases in different directions, making their combined impact on the observed patterns difficult to predict. More precise estimates of the distributional properties of homophones will therefore benefit from future corpora of child-directed speech with richer and more standardized phonological and morphological annotation.

5. Conclusion

Building on previous findings that homophones pose challenges for children's word learning unless supported by contextual cues, we examined the distributional properties of homophones in maternal speech in English, French, and Japanese. Our results indicate that homophones in maternal speech are not systematically avoided through greater separation in frequency or age of occurrence, nor are they preferentially distributed across distinct syntactic context. Moreover, homophones are not more strongly separated or more richly supported by syntactic context in a homophone-rich language such as Japanese. Taken together, these findings call for a reconsideration of the role homophones play in children's early language environments and point to alternative routes through which homophones may be learned.

References

Bates, Douglas, Mächler, Martin, Bolker, Ben, & Walker, Steve (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Ben-Shachar, Mattan S., Lüdecke, Daniel, & Makowski, Dominique (2020). effectsize: Estimation of Effect Size Indices and Standardized Parameters. *Journal of Open Source Software*, *5*(56), 2815. https://doi.org/10.21105/joss.02815

Bergelson, Elika (2020). The comprehension boost in early word learning: Older infants are better learners. *Child Development Perspectives*, *14*(3), 142–149. https://doi.org/10.1111/cdep.12373

Casenhiser Devin M. (2005). Children's resistance to homonymy: an experimental study of pseudohomonyms. *Journal of Child Language*, 32(2), 319–343. https://doi.org/10.1017/s0305000904006749

Dautriche, Isabelle, Fibla, Laia, Fievet, Anne-Caroline, & Christophe, Anne (2018). Learning homophones in context: Easy cases are favored in the lexicon of natural languages. *Cognitive Psychology*, *104*, 83–105. https://doi.org/10.1016/j.cogpsych.2018.04.001

Demuth, Katherine, Culbertson, Jennifer, & Alter, Jennifer (2006). Word-minimality, epenthesis and coda licensing in the early acquisition of English. *Language and Speech*, *49*(2), 137–174. https://doi.org/10.1177/00238309060490020201

Demuth, Katherine, & Tremblay, Annie (2008). Prosodically-conditioned variability in children's production of French determiners. *Journal of Child Language*, *35*(1), 99–127. https://doi.org/10.1017/s0305000907008276

Doherty, Martin J. (2004). Children's difficulty in learning homonyms. *Journal of Child Language*, *31*(1), 203–214. https://doi.org/10.1017/S030500090300583X

Eimas, Peter D., Siqueland, Einar R., Jusczyk, Peter, & Vigorito, James. (1971). Speech perception in infants. *Science, 171*(3968), 303–306. https://doi.org/10.1126/science.171.3968.303

Jones, Gary, Cabiddu, Francesco, Barrett, Doug J. K., Castro, Antonio, & Lee, Bethany (2023). How the characteristics of words in child-directed speech differ from adult-directed speech to influence children's productive vocabularies. *First Language*, *43*(3), 253–282. https://doi.org/10.1177/01427237221150070

Kingsbury, Paul, Strassel, Stephanie, McLemore, Cynthia, & Mclntyre, Robert (1997). *CALLHOME American English Lexicon (PRONLEX)*. Philadelphia: Linguistic Data Consortium. https://doi.org/10.35111/z1z4-ep76

Kobayashi, Megumi, Crist, Sean, Kaneko, Masayo, & McLemore, Cynthia (1996). *CALLHOME Japanese Lexicon*. Philadelphia: Linguistic Data Consortium. https://doi.org/10.35111/r0q4-qq14

Kuhl, Patricia K., Williams, Karen A., Lacerda, Francisco, Stevens, Kenneth N., & Lindblom, Björn (1992). Linguistic experience alters phonetic perception in infants by 6 months of age. *Science*, *255*(5044), 606–608. https://doi.org/10.1126/science.1736364

Lu, Youtao (2022). *Homophones: A Cross-Linguistic Perspective* [Doctoral dissertation, Brown University]. Brown Digital Repository. https://repository.library.brown.edu/studio/item/bdr:27ygsjqg

MacWhinney, Brian (2000). *The CHILDES Project: Tools for analyzing talk* (3rd ed.). Mahwah, NJ: Lawrence Erlbaum Associates.

Miyata, Susanne (2012). *Japanese CHILDES: The 2012 CHILDES manual for Japanese*. URL http://www2.aasa.ac.jp/people/smiyata/CHILDESmanual/chapter01.html

New, Boris, Pallier, Christophe, Brysbaert, Marc, & Ferrand, Ludovic (2004). *Lexique 2*: A new French lexical database. *Behavior Research Methods, Instruments, & Computers*, *36*(3), 516–524. https://doi.org/10.3758/BF03195598

Phillips, Juliet R. (1973). Syntax and vocabulary of mothers' speech to young children: Age and sex comparisons. *Child Development, 44*(1), 182–185. https://doi.org/10.2307/1127699

Piantadosi, Steven T., Tily, Harry, & Gibson, Edward (2012). The communicative function of ambiguity in language. *Cognition*, *122*(3), 280–291. https://doi.org/10.1016/j.cognition.2011.10.004

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria. URL https://www.Rproject.org/

Simpson, Greg B. (1994). Context and the processing of ambiguous words. In Morton A. Gernsbacher (Ed.), *Handbook of psycholinguistics* (pp. 359–374). Academic Press.

Slobin, Dan I. (1985). Crosslinguistic evidence for the language-making capacity. In Dan I. Slobin (Ed.), *The crosslinguistic study of language acquisition, Vol. 1. The data; Vol. 2. Theoretical issues* (pp. 1157–1256). Lawrence Erlbaum Associates, Inc.

Swingley, Daniel, & Aslin, Richard N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480–484. https://doi.org/10.1111/1467-9280.00485

Swingley, Daniel, & Aslin, Richard N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*(2), 99–132. https://doi.org/10.1016/j.cogpsych.2006.05.001

Trott, Sean, & Bergen, Benjamin (2020). Why do human languages have homophones?. *Cognition*, *205*, 104449. https://doi.org/10.1016/j.cognition.2020.104449

Wickham Hadley (2016). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Yoshida, Katherine A., Fennell, Christopher T., Swingley, Daniel, & Werker, Janet F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science*, *12*(3), 412–418. https://doi.org/10.1111/j.1467-7687.2008.00789.x

Proceedings of the 50th annual Boston University Conference on Language Development

edited by Romi Hill, Xinhan Jiang, Danutham Worapipat, and Aditya Yedetore

Cascadilla Press Somerville, MA 2026

ISSN 1080-692X
ISBN 978-1-57473-047-0 (2 volume set, paperback)

Ordering information

To order a copy of the proceedings or to place a standing order, contact:

Cascadilla Press, P.O. Box 440355, Somerville, MA 02144, USA
phone: 1-617-776-2370, sales@cascadilla.com, www.cascadilla.com