The role of cross-linguistic lexical similarity on bilingual word acquisition

Bilinguals face the task of learning words from languages with overlapping phonologies. Floccia et al. (2018) reported larger vocabulary sizes for 24-month-old bilinguals that were learning languages sharing large amount of cognates (e.g., English-Dutch). The mechanisms underlying this effect remain unknown. We explore two compatible scenarios. First, cognates may be learnt earlier than non-cognates, boosting lexical acquisition at early ages. This would account for the difference in vocabulary size associated to the amount of shared cognates across languages. Second, we explore the possibility that the word-forms of one language interact with those form the other language, scaffolding the acquisition of their translation equivalents when their phonologies overlap. This mechanism, in line with the parallel activation account of bilingual speech perception, would provide a plausible explanation to why cognates are acquired ealier by bilinguals. We developed an online tool to collect parental reports of receptive and productive vocabularies from children learning Catalan and/or Spanish, and present data on receptive and productive vocabulary of bilingual toddlers aged 12 to 34 months.

# Introduction

Bilinguals toddlers must learn two distinct set of words (one for each language) that partially overlap in sound and meaning. Previous studies have reported that bilingual children know fewer words than their monolingual counterparts when only language is considered (e.g., English monolinguals know more words in English than English-Spanish bilinguals). When taking both languages into account, bilingual children seem to know, at least, as many words as monolinguals do (Ben-Zeev 1977; Doyle and others 1977; Rosenblum and Pinker 1983; Fernandez et al. 1992; Pearson and Fernández 1994; Bialystok et al. 2010; Hoff et al. 2012, 2014; Core et al. 2013; Blom et al. 2019; but see Pearson, Fernández, and Oller 1993; Houwer, Bornstein, and Putnick 2014). Not all bilinguals show a similar developmental trajectory of lexical acquisition. Recent studies have capitalised on the role of the similarity between the specific pair of languages the infant is learning. One way of measuring language similarity is by estimating the amount of cognates (i.e., form-similar translation equivalents) they share. Floccia et al. (2018) explored the impact of this form of similarity on vocabulary sizes from a large sample of 24-month-old toddlers learning English and an additional language. Toddlers learning languages that shared a large amount of cognates (e.g., English-German) showed larger vocabulary sizes than those learning languages that shared fewer cognates (e.g., English-Spanish). Blom et al. (2019) extended these findings to children aged three to 10 years. These results suggest that the phonological overlap between the word inventories infants are acquiring impacts their trajectory of lexical aquisition.

How does cognateness influence word learning? Floccia et al. (2018) point to the possibility that parallel activation during speech perception may play a role in this cognate facilitation effect. The parallel activation principle suggests that bilinguals activate both languages in parallel during speech perception and production, even when only one of them is being used. This hypothesis can only be understood under the more general account of language non-selectivity of lexical processing, which states that translation equivalents are co-activated at the lexical level (e.g., De Groot and Nas 1991; Costa, Caramazza, and Sebastian-Galles 2000; Dijkstra 2005; Thierry and Wu 2007; Hoshino and Kroll 2008; Yudes, Macizo, and Bajo 2010; Bobb et al. 2020; see Kroll, Gullifer, and Rossi 2013 for review). A growing body of experimental evidence has been given shape to different formal models of bilingual lexical processing, like the Revised Hierarchical model (Kroll and Stewart 1994; Kroll et al. 2010), the Bilingual Interactive Activation model (BIA; Dijkstra and Van Heuven 1998), and its successor BIA+ (Dijkstra and Van Heuven 2002), the Bilingual Language Interaction Network for Comprehension of Speech model (BLINCS; Shook and Marian 2013) and, more recently, Multilink (Dijkstra et al. 2019), which represents an effort for integrating the former to account for language-non selective activation during both comprehension and production. However, these models do not address word learning during language acquistion.

It is feasible that parallel activation plays a role in vocabulary acquisition during the first years of life. On top of Floccia et al. (2018)’s findings, other studies have provided experimental evidence of parallel activation during speech perception in bilingual toddlers (Von Holzen and Mani 2012; Von Holzen, Fennell, and Mani 2019). Yet, it is unclear how parallel activation influences word learning at these ages. What mechanisms are engaged, and how do they relate to vocabulary size? We contemplate two scenarios. First, it is possible that the reported “similarity boost” on vocabulary size is not directly related to the amount of cognates shared between both languages, but rather to the overlap across their correspondent phonological inventories. One of the measures of language similarity Floccia et al. computed is the mean phonological Levenshtein distance across translation equivalents. This involves the calculation of the number of edit operations (e.g., insertions, deletions or replacements) the phonological transcription of a word-form has to go through to become identical to that of its translation equivalent. These scores were aggregated across pais of tanslation equivalents, obtaining a single similarity estimate of for each language pair. This measure embeds confounded information about the “degree of cognateness” and the amount of phonological overlap between both languages. Languages that are phonologically closer tend to share more cognates than those that are phonologically more distant (e.g., Spanish-Italian, Enlgish-Dutch), with few exceptions (e.g., Spanish-Basque). In the light of these data, it is not possible to discern whether vocabulary size is increased by the ealier acquisition of cognates or by the overlap across the phonological systems of the languages of a bilingual. We address this issue by testing two hypotesis.

## Hypothesis 1: Cognates are acquired earlier than non-cognates

First, we test whether the increase in vocabulary size reported by Floccia et al. for languages sharing a greater amount of cognates is indeed caused by cognates being acquired earlier. In line with this account, Bosch and Ramon-Casas (2014) found higher productive vocabulary scores in 18 month-old Catalan-Spanish bilinguals, compared to monolinguals, due to the higher presence of cognates their lexicon. We will adopt a similar approach to Braginsky et al. (2019), who modelled cross-sectional data from individual items to explore what properties were associated with an earlier acquisition. We will analyse the acquisition trajectories of a selection of 717 translation equivalents in a cross-section sample of Catalan/Spanish bilinguals aged 14 to 34 months of age. This approach allows to analyse how the proportion of infants that know each item changes across ages, providing a more suitable framework to put developmental hypotheses to test, instead of taking a “snapshot” of the state of the lexical of children at a single age.

## Hypothesis 2: Cognates are acquired closer in times than non-cognates

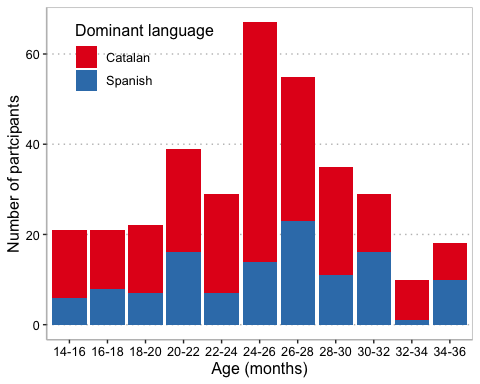
One mechanism that could acccount for an ealier acquisition of cognates is that parallel activation makes word forms acquired in one language available during the acquisition of word-forms in the other language. One consequence of this situation is that translation equivalents scaffold the acquisition of each other. This would point to different mechanisms engaged in the acquisition of cognates vs. non-cognates. While the phonology of cognate translation equivalents overlaps, and thus could be co-activated, the phonology of non-cognates does it in a much lesser degree. Consider the case of a Catalan-Spanish bilingual toddler that has already acquired the Catalan words *porta* (/ˈpɔr.tə/, door) and *taula* (/ˈtaw.ɫə/, table), but not yet their correspondent translation equivalents in Spanish (*puerta*, /ˈpwer.ta/; *mesa* /ˈme.sa/). According to our hypothesis, hearing the word *puerta* in presence of a door should lead to the activation of *porta*, which should facilitate the acquisition of the former. On the contrary, when hearing *mesa* in presence of a table, it is unlikely that the phonology of *taula* will be activated. This mechanism should lead to cognate translation equivalents being learnt closer in time than non-cognate translation equivalents. To test this hypothesis, we will extract the age of acquisition of each word-form in the questionnaire (as estimated by the model used to test Hypothesis 1), and will calculate the difference in time between the acquisition of such word-forms and their translation equivalents. We will test how cognateness affects such outcome.

# Method

## Participants

We collected data from 562 bilinguals from the Metropolitan Area of Barcelona, between 21th November, 2017 and 22th June, 2020. All families participated voluntarily. This study was approved by the Comitè d’Ètica de la Investigació amb Medicaments (CEIm) from Hospital del Mar (Barcelona, Spain), code XXXXXXXXX.

We assessed toddlers’ language background asking parents for an estimated proportion of exposure to each language. We defined participants as Spanish- or Catalan-dominant according to the language they were exposed the most to. Participants were considered bilingual - therefore included - if exposed 50-90% of the time to the non-dominant language. We excluded participants with >10% exposure to a third language. Fig. 1 shows how participants were distributed across age bins and language profiles.



## Questionnaire

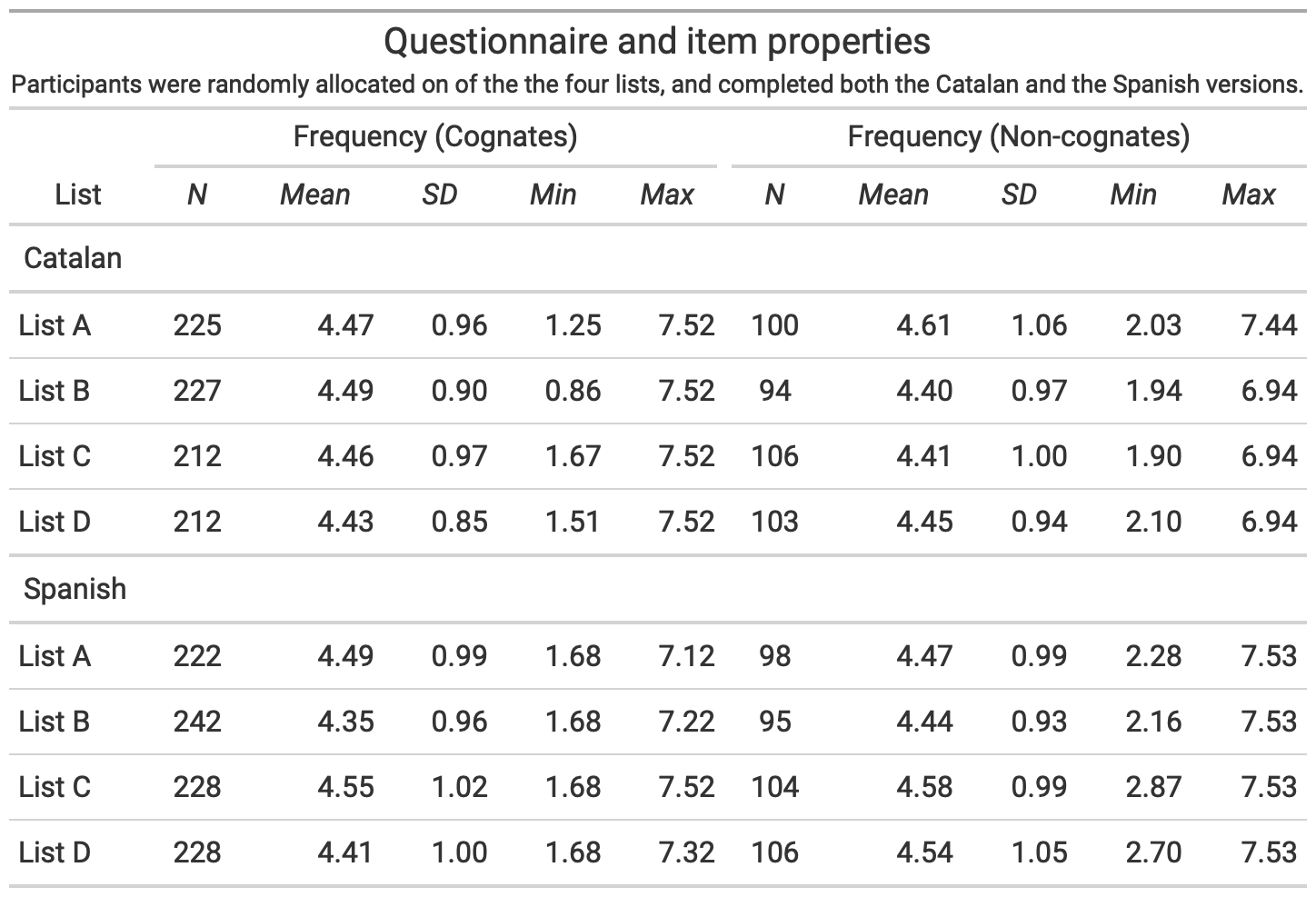
Parents completed an online questionnaire that gathered information about their demographic status, language environment of the child, and that included two vocabulary inventories: one for Catalan and one for Spanish. We implemented the questionnaire in formr (Arslan, Walther, and Tata 2020), a free, open-source platform. Participants were randomly allocated into one of four list of items. Each list contained a distinct inventory of translation equivalents. We included items from a diverse sample of 26 semantic/funcional categories: Action words, adventures, animals, body parts, clothes, color, descriptive words, food and drink, furniture and rooms, household items, on-line, outside, parts of animals, parts of things, people, pronouns, quantifiers, question words, time, toys, and vehicles. We discarded the following categories: adverbs, auxiliary words, connectives, interjections and games and routines. Those categories that did not entail more than 16 items - thus resulting in less than four items after dividing it in four lists - were not divided: all of their items were included in the four lists. Another subset of items that were part of the trial lsits of some experiments in the lab were also included in all versions.

Each item of the inventory presented a word-form, to which parents were instructed to report whether the toddlers was able to understand it, understand ans say it, or neither. The Catalan inventory contained 796 items (582 cognates, 195 non-cognates) and the Spanish inventory contained 801 items (584 cognates, 197 non-cognates). Items of one language were translation equivalents from the other (e.g., whenever *gos* [dog] was included in the Catalan inventory, the *perro* [dog] was included in the Spanish inventory), roughly following a one-to-one mapping. When there were two acceptable translation equivalents for a given word, we included both (e.g., Catalan *acabar* [*to finish*] and Spanish *acabar* and *terminar*), or merged them into the same one (e.g., Spanish *mono* [*monkey*] and Catalan *mono/mico*). We excluded from the analysis multi-word items (e.g., *barrita de cereales* [cereal bar]) and items that included more than one word-form (e.g., *mono / mico*).

## Data analysis

For testing Hypothesis 1, we modelled the proportion of infants that understood or produced each item within age bins of 2 months. Our model assumes that this proportion follows a logistic curve accross ages, an that the age of acquisition of each item is generated from a linear model that adjusts for the dominance of the language of the item (*item dominance*, L1 vs. L2), the cognate status of the word (*cognateness*, Non-cognate vs. Cognate), and the frequency of the word form in its correspondent language. Our prediction is that the age of acquisition will be lower for cognate words, compared to that of non-cognate words.

Items were classified into cognates and non-cognates by a trained Catalan-Spanish bilingual with broad experience in phonology and phonetics. Items were classified as L1 if the item was included in the inventory of the language the participant was dominant in, and as L2 if otherwise For instance, the item *porta* (door, in Catalan) was classified as L1 when answered by Catalan-dominant participants, and as L2 when answered by Spanish-dominant participants. We retrieved frequencies from SUBTLEX-ESP (Cuetos et al. 2011) and SUBTLEX-CAT (Boada et al. 2019) for Spanish and Catalan words, respectively. Word frequency was calculated as the Zipf score of the word (Zipf 1949; Heuven et al. 2014). Missing frequency scores were imputed using multiple imputation (Buuren and Groothuis-Oudshoorn 2011). Table 1 shows the classification of items in cognates and non-cognates and their frequency scores across the four lists of the inventories.



We fit two separate logistic models on comprehensive and productive data, respectively. Logistic curves are characterised by three parameters(see Mahr 2019, and Mahr 2020 for an excellent introduction to this approach, in which we based this analysis): (a) an *asymptote* (upper boundary of the proportion of infants that have acquired the item), (b) the *steepness* (how fast the a given item is acquired by infants), and (c) a *mid-point* (the age at which steepness is maximum). We work on the assumption that the trajectory of aquisition of items follows a sigmoidal shape (Mayor and Plunkett 2011), and that the age of acquisition of a word is fairly captured by the position of the mid-point of its curve. This approach allows us to model and estimate the age of acquisition of items explicitely. Our model estimated the asymptote and stepness as grand means across items (i.e., intercepts). Mid-points were estimated by a linear zero-one-beta-regression model[[1]](#footnote-1)(Ospina and Ferrari 2012) that included our predictors of interest (*age bin*, *item dominance* (L1/L2, sum-coded), *cognateness* (*non-cognate/cognate, sum-coded), the interaction between* item dominance\* and *cognateness*, and frequency (standardised Zipf score)]. We also included a random intercept for each translation equivalent.

We estimated the parameters of our model using the Bayesian framework. Bayesian models use the Bayes rule to combine prior knowlegde about the distribution of a given parameter, and the likelihood of each of the values it can take given tha observed data. We relied on the available data on Wordbank (Frank et al. 2017) to generate our priors (see Appendix 1 for more details about our priors). Finally, to test the contribution of the *cognateness* predictor to the estimation of mid-points, we compared the fit of the model with a reduced (null) model that did not include the *cognateness* factor using Pareto-smoothed importance sampling (PSIS; Vehtari, Gelman, and Gabry 2017; Vehtari et al. 2019). We fit both null and alternative models using the R environment (R Core Team 2013; RStudio Team 2015) and the brms package (Bürkner 2017), which relies on the probabilistic language Stan (Carpenter et al. 2017) to estimate posterior distributions. Data and model results will be processed and visualised using the R packages the tidyverse family of packages (Wickham et al. 2019) and the tidybayes package (Kay 2020).

To test Hypothesis 2, we run a Bayesian anova using the difference in mid-points between the two word-forms of each translation equivalent (in Catalan and Spanish, respectively), that included *cognateness* as predictor. Using the BayesFactor R package (Morey and Rouder 2018), we computed a Bayes factor (*BF*) that compared the likelihood of a linear model that included *cognateness* as a predictor against the likelihood of a linear model that did not, under the light of our data.

# Appendix

## Appendix 1: Session info

R version 3.6.3 (2020-02-29) Platform: x86\_64-apple-darwin15.6.0 (64-bit) Running under: macOS Catalina 10.15.4

Matrix products: default BLAS: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRblas.0.dylib LAPACK: /Library/Frameworks/R.framework/Versions/3.6/Resources/lib/libRlapack.dylib

locale: [1] en\_US.UTF-8/en\_US.UTF-8/en\_US.UTF-8/C/en\_US.UTF-8/en\_US.UTF-8

attached base packages: [1] stats graphics grDevices utils datasets methods base

other attached packages: [1] rlang\_0.4.6 here\_0.1 data.table\_1.12.8  
[4] gt\_0.2.1 patchwork\_1.0.0.9000 ggridges\_0.5.2  
[7] janitor\_2.0.1 lubridate\_1.7.9 readxl\_1.3.1  
[10] forcats\_0.5.0 stringr\_1.4.0 dplyr\_1.0.0  
[13] purrr\_0.3.4 readr\_1.3.1 tidyr\_1.1.0  
[16] tibble\_3.0.1 ggplot2\_3.3.2 tidyverse\_1.3.0  
[19] magrittr\_1.5 knitr\_1.28 papaja\_0.1.0.9942

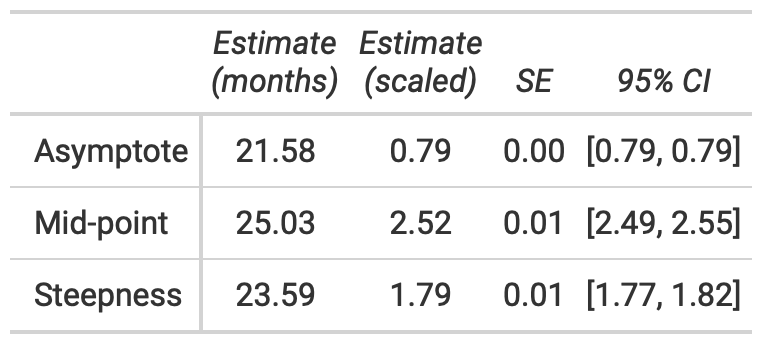
loaded via a namespace (and not attached): [1] Rcpp\_1.0.4.6 lattice\_0.20-38 ps\_1.3.3 assertthat\_0.2.1  
[5] rprojroot\_1.3-2 digest\_0.6.25 R6\_2.4.1 cellranger\_1.1.0  
[9] plyr\_1.8.6 backports\_1.1.8 reprex\_0.3.0 evaluate\_0.14  
[13] httr\_1.4.1 pillar\_1.4.4 rstudioapi\_0.11 callr\_3.4.3.9000  
[17] blob\_1.2.1 checkmate\_2.0.0 rmarkdown\_2.3 labeling\_0.3  
[21] webshot\_0.5.2 munsell\_0.5.0 broom\_0.5.6 compiler\_3.6.3  
[25] modelr\_0.1.8 xfun\_0.15 pkgconfig\_2.0.3 htmltools\_0.5.0  
[29] tidyselect\_1.1.0 fansi\_0.4.1 crayon\_1.3.4 dbplyr\_1.4.4  
[33] withr\_2.2.0 commonmark\_1.7 grid\_3.6.3 nlme\_3.1-147  
[37] jsonlite\_1.6.1 gtable\_0.3.0 lifecycle\_0.2.0 DBI\_1.1.0  
[41] scales\_1.1.1 cli\_2.0.2 stringi\_1.4.6 farver\_2.0.3  
[45] fs\_1.4.1 snakecase\_0.11.0 xml2\_1.3.2 ellipsis\_0.3.1  
[49] generics\_0.0.2 vctrs\_0.3.1 RColorBrewer\_1.1-2 tools\_3.6.3  
[53] glue\_1.4.1 hms\_0.5.3 processx\_3.4.2 yaml\_2.2.1  
[57] colorspace\_1.4-1 rvest\_0.3.5 haven\_2.3.1 sass\_0.2.0

## Appendix 2: Model

Our model is incomplete, in that we have not included all parameters that may play a role in the real mechanisms that generate our observed data. Previous studies have found robust effects of sex (e.g., Frank et al. (2017)), word length/number of phonemes (e.g., Frank et al. (2017)), etc. Computational limitations have prevented us from specifying and fitting a more exhaustive model. Following the model-building strategy suggested by Schad, Betancourt, and Vasishth (2020), we have specified an intitial model that captures the fundamental structure of our phenomenon of interest and that can be computed with our available resources. Then, we have expanded such model incorporating a new predictor that adresses our hypotheses.

## Appendix 3: Priors

We fitted a logistic model using the proportion of infants that acquired each item as output, and age-bin (defined as we did with our data) as input, estimating the three parameters afore mentioned (asymptote, mid-point, and steepness). We included data from 1.311910^{4} items from 20 languages[[2]](#footnote-2), from monolingual infants aged 16 to 30 months of age. Table 1 shows the estimates of this model, which were rescaled and included as priors in our main model to test Hypothesis 1.



# References

Arslan, Ruben C., Matthias P. Walther, and Cyril S. Tata. 2020. “Formr: A Study Framework Allowing for Automated Feedback Generation and Complex Longitudinal Experience-Sampling Studies Using R.” *Behavior Research Methods* 52 (1): 376–87. <https://doi.org/10.3758/s13428-019-01236-y>.

Ben-Zeev, Sandra. 1977. “The Influence of Bilingualism on Cognitive Strategy and Cognitive Development.” *Child Development* 48 (3): 1009–18. <https://doi.org/10.2307/1128353>.

Bialystok, Ellen, Gigi Luk, Kathleen F. Peets, and Sujin Yang. 2010. “Receptive Vocabulary Differences in Monolingual and Bilingual Children.” *Bilingualism: Language and Cognition* 13 (4): 525–31. <https://doi.org/10.1017/S1366728909990423>.

Blom, Elma, Tessel Boerma, Evelyn Bosma, Leonie Cornips, Kirsten van den Heuij, and Mona Timmermeister. 2019. “Cross-Language Distance Influences Receptive Vocabulary Outcomes of Bilingual Children.” *First Language*, December, 0142723719892794. <https://doi.org/10.1177/0142723719892794>.

Boada, Roger, Marc Guasch, Juan Haro, Josep Demestre, and Pilar Ferré. 2019. “SUBTLEX-CAT: Subtitle Word Frequencies and Contextual Diversity for Catalan.” *Behavior Research Methods*, March. <https://doi.org/10.3758/s13428-019-01233-1>.

Bobb, Susan C., Katie Von Holzen, Julien Mayor, Nivedita Mani, and Manuel Carreiras. 2020. “Co-Activation of the L2 During L1 Auditory Processing: An ERP Cross-Modal Priming Study.” *Brain and Language* 203 (April): 104739. <https://doi.org/10.1016/j.bandl.2019.104739>.

Bosch, Laura, and Marta Ramon-Casas. 2014. “First Translation Equivalents in Bilingual Toddlers’ Expressive Vocabulary: Does Form Similarity Matter?” *International Journal of Behavioral Development* 38 (4): 317–22. <https://doi.org/10.1177/0165025414532559>.

Braginsky, Mika, Daniel Yurovsky, Virginia A. Marchman, and Michael C. Frank. 2019. “Consistency and Variability in Children’s Word Learning Across Languages.” *Open Mind* 3 (June): 52–67. <https://doi.org/10.1162/opmi_a_00026>.

Buuren, Stef van, and Karin Groothuis-Oudshoorn. 2011. “<span Class="nocase">mice</Span>: Multivariate Imputation by Chained Equations in R.” *Journal of Statistical Software* 45 (3): 1–67. <https://www.jstatsoft.org/v45/i03/>.

Bürkner, Paul-Christian. 2017. “Brms: An R Package for Bayesian Multilevel Models Using Stan.” *Journal of Statistical Software* 80 (1): 1–28. <https://doi.org/10.18637/jss.v080.i01>.

Carpenter, Bob, Andrew Gelman, Matthew D. Hoffman, Daniel Lee, Ben Goodrich, Michael Betancourt, Marcus Brubaker, Jiqiang Guo, Peter Li, and Allen Riddell. 2017. “Stan: A Probabilistic Programming Language.” *Journal of Statistical Software* 76 (1): 1–32. <https://doi.org/10.18637/jss.v076.i01>.

Core, Cynthia, Erika Hoff, Rosario Rumiche, and Melissa Señor. 2013. “Total and Conceptual Vocabulary in Spanish–English Bilinguals from 22 to 30 Months: Implications for Assessment.” *Journal of Speech, Language, and Hearing Research* 56 (5): 1637–49. <https://doi.org/10.1044/1092-4388(2013/11-0044)>.

Costa, Albert, Alfonso Caramazza, and Núria Sebastian-Galles. 2000. “The Cognate Facilitation Effect: Implications for Models of Lexical Access.” *Journal of Experimental Psychology* 26 (5): 1283–96. <https://doi.org/10.1037/0278-7393.26.5.1283>.

Cuetos, Fernando, Maria Glez-Nosti, Analía Barbón, and Marc Brysbaert. 2011. “SUBTLEX-ESP: Spanish Word Frequencies Based on Film Subtitles.” *Psicológica* 32 (2): 133–43.

De Groot, Annette MB, and Gerard LJ Nas. 1991. “Lexical Representation of Cognates and Noncognates in Compound Bilinguals.” *Journal of Memory and Language* 30 (1): 90–123.

Dijkstra, AFJ, and Walter JB Van Heuven. 2002. “The Architecture of the Bilingual Word Recognition System: From Identification to Decision.”

Dijkstra, Ton. 2005. “Bilingual Visual Word Recognition and Lexical Access.” In *Handbook of Bilingualism: Psycholinguistic Approaches*, 179–201. New York, NY, US: Oxford University Press.

Dijkstra, Ton, and Walter JB Van Heuven. 1998. “The BIA Model and Bilingual Word Recognition.” *Localist Connectionist Approaches to Human Cognition*, 189–225.

Dijkstra, Ton, Alexander Wahl, Franka Buytenhuijs, Nino Van Halem, Zina Al-Jibouri, Marcel De Korte, and Steven Rekké. 2019. “Multilink: A Computational Model for Bilingual Word Recognition and Word Translation.” *Bilingualism: Language and Cognition* 22 (4): 657–79. <https://doi.org/10.1017/S1366728918000287>.

Doyle, Anna Beth, and others. 1977. “Some Issues in the Assessment of Linguistic Consequences of Early Bilingualism.” *Working Papers on Bilingualism*, no. 14.

Fernandez, Maria C., Barbara Z. Pearson, Vivian M. Umbel, D. K. Oiler, and Margie Molinet-Molina. 1992. “Bilingual Receptive Vocabulary in Hispanic Preschool Children.” *Hispanic Journal of Behavioral Sciences* 14 (2): 268–76. <https://doi.org/10.1177/07399863920142006>.

Floccia, Caroline, Thomas D. Sambrook, Claire Delle Luche, Rosa Kwok, Jeremy Goslin, Laurence White, Allegra Cattani, et al. 2018. “I: Introduction.” *Monographs of the Society for Research in Child Development* 83 (1): 7–29. <https://doi.org/10.1111/mono.12348>.

Frank, Michael C., Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2017. “Wordbank: An Open Repository for Developmental Vocabulary Data.” *Journal of Child Language* 44 (3): 677–94. <https://doi.org/10.1017/S0305000916000209>.

Heuven, Walter J. B. van, Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2014. “Subtlex-UK: A New and Improved Word Frequency Database for British English:” *Quarterly Journal of Experimental Psychology*, June. <https://journals.sagepub.com/doi/10.1080/17470218.2013.850521>.

Hoff, Erika, Cynthia Core, Silvia Place, Rosario Rumiche, Melissa Señor, and Marisol Parra. 2012. “Dual Language Exposure and Early Bilingual Development\*.” *Journal of Child Language* 39 (1): 1–27. <https://doi.org/10.1017/S0305000910000759>.

Hoff, Erika, Rosario Rumiche, Andrea Burridge, Krystal M. Ribot, and Stephanie N. Welsh. 2014. “Expressive Vocabulary Development in Children from Bilingual and Monolingual Homes: A Longitudinal Study from Two to Four Years.” *Early Childhood Research Quarterly* 29 (4): 433–44. <https://doi.org/10.1016/j.ecresq.2014.04.012>.

Hoshino, Noriko, and Judith F. Kroll. 2008. “Cognate Effects in Picture Naming: Does Cross-Language Activation Survive a Change of Script?” *Cognition* 106 (1): 501–11. <https://doi.org/10.1016/j.cognition.2007.02.001>.

Houwer, Annick De, Marc H. Bornstein, and Diane L. Putnick. 2014. “A Bilingual–Monolingual Comparison of Young Children’s Vocabulary Size: Evidence from Comprehension and Production.” *Applied Psycholinguistics* 35 (6): 1189–1211. <https://doi.org/10.1017/S0142716412000744>.

Kay, Matthew. 2020. “<span Class="nocase">tidybayes</Span>: Tidy Data and Geoms for Bayesian Models.” Manual. <https://doi.org/10.5281/zenodo.1308151>.

Kroll, Judith F., Jason W. Gullifer, and Eleonora Rossi. 2013. “The Multilingual Lexicon: The Cognitive and Neural Basis of Lexical Comprehension and Production in Two or More Languages.” *Annual Review of Applied Linguistics* 33 (March): 102–27. <https://doi.org/10.1017/S0267190513000111>.

Kroll, Judith F., Janet G. Van Hell, Natasha Tokowicz, and David W. Green. 2010. “The Revised Hierarchical Model: A Critical Review and Assessment\*.” *Bilingualism: Language and Cognition* 13 (3): 373–81. <https://doi.org/10.1017/S136672891000009X>.

Kroll, Judith F., and Erika Stewart. 1994. “Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connection Between Bilingual Memory Representations.” *Journal of Memory and Language* 33 (2): 149–74. <https://doi.org/10.1006/jmla.1994.1008>.

Mahr, Tristan. 2019. “Anatomy of a Logistic Growth Curve.” *Higher Order Functions*. <https://tjmahr.github.io/anatomy-of-a-logistic-growth-curve/>.

———. 2020. “Bayes’ Theorem in Three Panels.” *Higher Order Functions*. <https://tjmahr.github.io/bayes-theorem-in-three-panels/>.

Mayor, Julien, and Kim Plunkett. 2011. “A Statistical Estimate of Infant and Toddler Vocabulary Size from CDI Analysis.” *Developmental Science* 14 (4): 769–85. <https://doi.org/10.1111/j.1467-7687.2010.01024.x>.

Morey, Richard D., and Jeffrey N. Rouder. 2018. “BayesFactor: Computation of Bayes Factors for Common Designs.” Manual. <https://CRAN.R-project.org/package=BayesFactor>.

Ospina, Raydonal, and Silvia L. P. Ferrari. 2012. “A General Class of Zero-or-One Inflated Beta Regression Models.” *Computational Statistics & Data Analysis* 56 (6): 1609–23. <https://doi.org/10.1016/j.csda.2011.10.005>.

Pearson, Barbara Zurer, and Sylvia C. Fernández. 1994. “Patterns of Interaction in the Lexical Growth in Two Languages of Bilingual Infants and Toddlers.” *Language Learning* 44 (4): 617–53. <https://doi.org/10.1111/j.1467-1770.1994.tb00633.x>.

Pearson, Barbara Zurer, Sylvia C. Fernández, and D. Kimbrough Oller. 1993. “Lexical Development in Bilingual Infants and Toddlers: Comparison to Monolingual Norms.” *Language Learning* 43 (1): 93–120. <https://doi.org/10.1111/j.1467-1770.1993.tb00174.x>.

R Core Team. 2013. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. <http://www.R-project.org/>.

Rosenblum, Tamar, and Steven A Pinker. 1983. “Word Magic Revisited: Monolingual and Bilingual Children’s Understanding of the Word-Object Relationship.” *Child Development*, 773–80. <https://doi.org/10.2307/1130064>.

RStudio Team. 2015. *RStudio: Integrated Development Environment for R*. Boston, MA: RStudio, Inc. <http://www.rstudio.com/>.

Schad, Daniel J., Michael Betancourt, and Shravan Vasishth. 2020. “Toward a Principled Bayesian Workflow in Cognitive Science.” *arXiv:1904.12765 [Stat]*, February. <http://arxiv.org/abs/1904.12765>.

Shook, Anthony, and Viorica Marian. 2013. “The Bilingual Language Interaction Network for Comprehension of Speech\*.” *Bilingualism: Language and Cognition* 16 (2): 304–24. <https://doi.org/10.1017/S1366728912000466>.

Thierry, Guillaume, and Yan Jing Wu. 2007. “Brain Potentials Reveal Unconscious Translation During Foreign-Language Comprehension.” *Proceedings of the National Academy of Sciences* 104 (30): 12530–5. <https://doi.org/10.1073/pnas.0609927104>.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2017. “Practical Bayesian Model Evaluation Using Leave-One-Out Cross-Validation and WAIC.” *Statistics and Computing* 27 (5): 1413–32. <https://doi.org/10.1007/s11222-016-9696-4>.

Vehtari, Aki, Daniel Simpson, Andrew Gelman, Yuling Yao, and Jonah Gabry. 2019. “Pareto Smoothed Importance Sampling.” *arXiv:1507.02646 [Stat]*, July. <http://arxiv.org/abs/1507.02646>.

Von Holzen, Katie, Christopher T. Fennell, and Nivedita Mani. 2019. “The Impact of Cross-Language Phonological Overlap on Bilingual and Monolingual Toddlers’ Word Recognition.” *Bilingualism: Language and Cognition* 22 (3): 476–99. <https://doi.org/10.1017/S1366728918000597>.

Von Holzen, Katie, and Nivedita Mani. 2012. “Language Nonselective Lexical Access in Bilingual Toddlers.” *Journal of Experimental Child Psychology* 113 (4): 569–86. <https://doi.org/10.1016/j.jecp.2012.08.001>.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D’Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. “Welcome to the <span Class="nocase">tidyverse</Span>.” *Journal of Open Source Software* 4 (43): 1686. <https://doi.org/10.21105/joss.01686>.

Yudes, Carolina, Pedro Macizo, and Teresa Bajo. 2010. “Cognate Effects in Bilingual Language Comprehension Tasks.” *NeuroReport* 21 (7): 507–12. <https://doi.org/10.1097/WNR.0b013e328338b9e1>.

Zipf, George Kingsley. 1949. *Human Behavior and the Principle of Least Effort*. Human Behavior and the Principle of Least Effort. Oxford, England: Addison-Wesley Press.

1. Zero-or-one-inflated beta regression models allow to account for overdispersion then using proportions as the output of a model accounting, while also allowing values to be at boundary (0 and 1). [↑](#footnote-ref-1)
2. Chinese (Cantonese and Mandarin), Croatian, Czech, Danish, English (Australian, British, and American), French (European), German, Hebrew, Italian, Korean, Latvian, Portuguese, Russian, Slovak, Spanish (European and Mexican), and Turkish. All the other available languages were excluded, as no (or insufficient) data were available. [↑](#footnote-ref-2)