Measuring a Basic Developmental Vocabulary Across Many Languages

Anonymous CogSci submission

Abstract

Early language skill is predictive of later life outcomes, and is thus of great interest to developmental psychologists and clinicians. The Communicative Development Inventories (CDIs), including a parent-reported inventory of early-learned vocabulary items, has proven to be a valid and reliable instrument for measuring children's early language skill. CDIs have been adapted to dozens of languages, and cross-linguistic comparisons thus far show both consistency and variability in language acquisition trajectories. Here, we use item-response theory models to examine the psychometric properties of translation-equivalent concepts that have been included on CDIs in several languages, with the goal of identifying a short list of concepts that are of approximately equal difficulty across the majority of the languages. Using a list of XX items, we test how well this

Keywords: early language learning; CDI; psychometrics; cross-linguistic comparison; Swadesh vocabulary;

Introduction

Children's early language skill is predictive of educational outcomes, other developmental milestones, etc.

Assessing children's early language skill is important for researchers, clinicians and parents, as early language skill is predictive of educational outcomes years later. The MacArthur-Bates Communicative Development Inventories [CDIs; Fenson et al. (2007)] are a set of parent report forms that offer a holistic assessment of children's productive and receptive language skills. CDIs are low-cost to administer and produce reliable and valid estimates of early vocabulary and other aspects of early language (Fenson et al., 1994). The CDIs offer more comprehensive data than a short interaction in the lab with a child (Fenson et al., 2000), because they ask parents to report on vocabulary comprehension as well as production, and other milestones, such as communicative gesture use and use of word combinations. Vocabulary size is assessed via a checklist format, which allows caregivers to quickly scan and recognize words their child produces or understands, rather than relying on recall alone. Because of these properties, CDI forms have been adapted to dozens of languages. Data from CDIs are archived in a central repository [Wordbank; Frank, Braginsky, Yurovsky, & Marchman (2017)], and insights from these data have been used to inform theories of early language learning (Frank, Braginsky, Yurovsky, & Marchman, 2021).

Although CDIs measure a variety of other constructs related to early language, our focus here is on vocabulary assessment. Across languages, measures of vocabulary on the CDIs are very tightly correlated with other aspects of early language like gesture and grammatical competence (Bates et al., 1994; Frank et al., 2021). These studies indicate that the language system looks to be "tightly woven" (Frank et al., 2021), in the sense that early vocabulary size serves as a good proxy measure of children's overall language skill.

The CDI:WS form is comprised of 22 semantic categories representing common early-learned words, including nouns (subdivided into e.g., Body Parts, Toys, and Clothing), action words (e.g. verbs), descriptive words (e.g. adjectives), and closed-class words such as pronouns.

Although the CDIs have many advantages, one drawback is that it is difficult for researchers to adapt a CDI for a new language.

Due to these challenges, there have been a variety of efforts to create shortened versions of the CDI. The 100-item short-form CDIs (Fenson et al., 2000, 2007) are derived from the 680-item CDI:WS form, with items selected based on difficulty and item-to-full score correlations, while also attempting to represent the diversity of semantic and linguistic categories. While the scores of the short-form CDI:WS are highly correlated with scores on the full CDI:WS, there is evidence for a ceiling effect for children older than 27 or 28 months of age.

The basis of CAT is item-response theory modeling [IRT; Embretson & Reise (2013)], a technique for the analysis of test data that allows the inference of both the abilities of test takers and the difficulty (and other information) of individual test questions along shared and standardized dimensions. CAT models use this item information – typically extracted from a larger dataset collected via standard testing methods – to select questions of the appropriate difficulty for a particular test-taker. An individual CAT includes a number of components, including the bank of possible items and their difficulties, as well as an algorithm that uses the responses received thus far to choose the next item to give to a test taker, and a rule for when to stop (e.g., after a fixed number of items or after a desired precision has been reached).

Previous work has applied CAT and related techniques to CDI forms, leveraging the availability of large datasets from previous CDI studies where parents filled out the full forms (Chai, Lo, & Mayor, 2020; Makransky, Dale, Havmose, & Bleses, 2016; Mayor & Mani, 2019). For example, Makransky et al. (2016) used IRT models fit to normative data from the CDI:WS to develop CAT versions (Fenson et al., 2007). They conducted a simulation study comparing full scores on the CDI to different fixed-length CAT versions (with 5–400 items). Scores from a CAT of only 50 items had a correlation of $r = 0.95$ with the full CDI. However, for the youngest age group (16–18 month-olds), the correlation with the full CDI was somewhat lower ($r = 0.87$).

Our goal in the current work is to… In particular, our contributions are:

We will first introduce Item Response Theory (IRT), fit the models to the datasets, and then examine We end by discussing the strengths and weaknesses of our approach.

Methods

Item Response Theory Models

A variety of IRT models targeting different types of testing scenarios have been proposed (see Baker, 2001 for an overview), but for the dichotomous responses that parents make for each item (word) regarding whether their child can produce that word, we will use the popular 2-parameter logistic (2PL) model that we have previously found is best justified for CDI data (see Kachergis & Frank, 2022).

The 2PL model jointly estimates for each child j a latent ability θ_j (here, language skill), and for each item i two parameters: the item's difficulty b_i and discrimination a_i, described below. In the 2PL model, the probability of child j producing a given item i is

$$P_i(x_i = 1|b_i, a_i, \theta_j) = \frac{1}{1 + e^{-Da_i(\theta_j - b_i)}}$$

Thus, children with high latent ability (θ) will be more likely to produce any given item than children with lower latent ability, and more difficult items will be produced by fewer children (at any given θ) than easier items. The discrimination (a_i) modifies the slope of the logistic (in the classic Rasch or 1PL model, the slope is always 1):

Datasets

Language	WS items	WS N	WG items	WG N
Norwegian	731	9304	395	2926
English (American)	680	8828	396	4130
Danish	725	3714	410	2398
Portuguese (European)	639	3012	293	1314
Turkish	711	2422	418	1115
Mandarin (Taiwanese)	696	1897	354	757
Spanish (Mexican)	680	1853	428	833
English (Australian)	558	1520		
Korean	641	1376	284	618
Cantonese	804	1295		
German	588	1181		
Slovak	609	1066	307	657
Mandarin (Beijing)	799	1056	227	230
Russian	728	1037	424	768
French (Quebecois)	664	929	366	598
Swedish	710	900	341	464
Spanish (Argentinian)	699	784		
Italian	670	752	408	648
French (French)	690	665	212	222
Spanish (European)	588	593	277	412
Hebrew	605	518	439	62
Latvian	723	500	402	183
Czech	553	493		
Croatian	717	377	380	250
Hungarian	802	363		
Dutch	704	303	160	317
Greek (Cypriot)	815	176		
Spanish (Peruvian)	600	105	112	87
Kigiriama	696	100	260	132
English (Irish)	660	99		
Irish	691	99		
Kiswahili	705	90	216	51
Finnish	581	70		
Persian	558	50	367	115

Table 1: Number of CDI:WG and CDI:WS items and subjects (N) per language.

We report IRT analyses for twenty-six languages from Wordbank (Frank et al., 2017), comprising production data from CDI:WS vocabulary checklists.[1] Data from the first twenty-six rows of Table 1 (Norwegian through Dutch) will be used to select a pool of words with approximately equal cross-linguistic difficulty. CDI:WS production data from an additional eight languages (Table 1: Greek (Cypriot) through Persian) had too few participants to be analyzed with IRT, but will be used to test how well the selected pool of generalize to new languages.

Participants The CDI:WS production dataset consists of the combined Wordbank production data for 47527 children

[1] `http://wordbank.stanford.edu/contributors`

aged 16-30 months on 23020 items across 34 forms. Figure 1C and 1D shows children's production scores vs. age for the CDI:WG and CDI:WS datasets respectively. Note that the socioeconomic distributions of these datasets are not matched. (See Frank et al. (2021) for a discussion of possible effects of socioeconomic status on vocabulary development.)

Instruments The 680-item vocabulary checklist of the English CDI:WS form is organized into 22 semantic categories (e.g., furniture, games and routines, people). The Spanish CDI:WS vocabulary checklist consists of 680 words, organized into 23 semantic categories. The English CDI:WG vocabulary checklist is comprised of XXX of the easier vocabulary items from the English CDI:WS, and the Spanish CDI:WG is a subset of XXX of the items from the Spanish CDI:WS.

When a CDI:WG form was administered, caregivers were asked to indicate for each vocabulary item whether their child 1) understands that word ("comprehends") or 2) both understands and says ("produces") that word. Leaving the item blank indicates that the child neither comprehends nor produces that word. When a CDI:WS forms was administered, caregivers were asked to indicate for each vocabulary item on the instrument whether or not their child can recognizably produce (say) the given word.

The current study solely investigates production, thus "produces" responses were coded as 1 and all other responses were coded as 0. Our datasets consist of a dichotomous-valued response matrix for each language, of size N subjects $\times$ W words.

Results

All models, simulations, and other materials are available on OSF[2].

Cross-linguistic similarities

We look at the Spearman correlation between the item difficulty of each language compared to each other language. We might expect this to recapitulate the historical relationship between languages, with more similar languages having more similar item difficulties (e.g., Quebecois and European French).

Difficulty by CDI Category

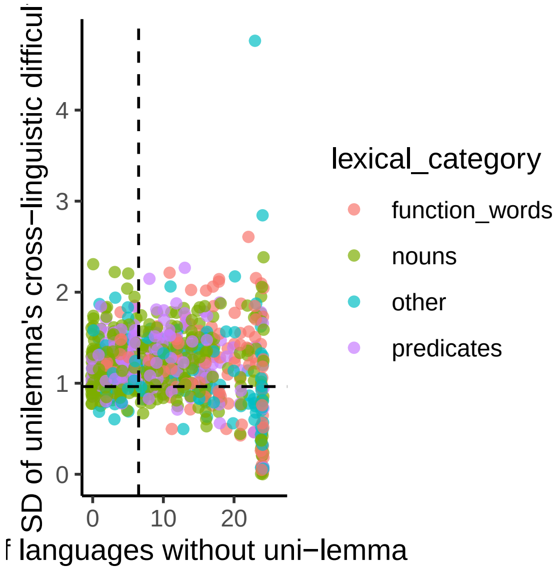
Which categories are hardest? Easiest? Which categories show the greatest variability? The least?

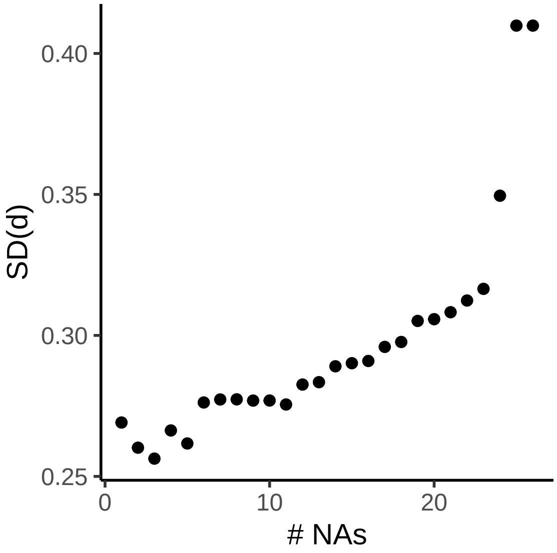
Candidate Items Shown below are the total number of items per language that also have uni-lemmas.

Language	Items
Czech	553
English (Australian)	558
German	588
Spanish (European)	588
Hebrew	605
Slovak	609
Portuguese (European)	639
Korean	641
French (Quebecois)	664
Italian	670
English (American)	680
Spanish (Mexican)	680
French (French)	690
Mandarin (Taiwanese)	696
Spanish (Argentinian)	699
Dutch	704
Swedish	710
Turkish	711
Croatian	717
Latvian	723
Danish	725
Russian	728
Norwegian	731
Mandarin (Beijing)	799
Hungarian	802
Cantonese	804

CDI:WS forms have a median of 693 uni-lemmas defined (range: 553 (Czech) to 804 (Cantonese)), and there are a total of 1840 uni-lemmas defined across the 26 languages. Only 61 uni-lemmas appear on all 26 CDI:WS forms: too few to comprise a short form. To expand the pool of items, we consider thresholds on both 1) the variability of an item's difficulty across languages (lower is better), and 2) the number of CDI:WS forms on which it appears (more is better). Below we examine the standard deviation of uni-lemmas' cross-linguistic difficulty as a function of how many languages that uni-lemma is missing from. For now we consider items with less than median variability in difficulty that are included on 10 or more of the 26 CDI:WS forms.

[2] OSF repository: `https://osf.io/XXX/`.





Of these, a total of 1312 uni-lemmas are included in more than one language, and only 656 uni-lemmas are included in 10 or more of the languages. We will start by considering this more restricted list.

To evaluate how variable items are in their cross-linguistic difficulty, we calculate the standard deviation (SD) of each uni-lemma's difficulty. The median SD is 1.17 (SD=0.41), so we consider the items with SD less than half the median SD (i.e., SD $<$ 0.96). These 106 candidate "Swadesh CDI" items are summarized by semantic category below (see OSF for full list). 52% of Swadesh CDI words are nouns, 21% are predicates, 22% are function words, and 5% are other. This breakdown is comparable to the lexical category percentages on the 680-item English CDI:WS (46% nouns, 24% predicates, %15 function words, and 15%), minimally suggesting that this method of selecting candidate items is not biased by lexical category. The candidate items, ultimately selected by lower variability, were also present on more forms than typical in the selection set: on average, each item appeared on 20.0754717 forms.

Comparing the IRT parameters of the Swadesh CDI words to the rest of the items showed that the candidate Swadesh items are significantly easier (mean Swadesh $d = 0.03$, other items' mean $d = -0.51$, $t(2737) = 12.74$, $p < .001$). If the Swadesh list is too focused on early-learned items, this may result in ceiling effects for older children. It may be prudent to consider expanding the Swadesh list to include some more difficult items. The discrimination parameter (i.e., slope) of the Swadesh items did not significantly differ from the other items.

category	n
household	13
action_words	12
animals	11
descriptive_words	9
body_parts	8
food_drink	8
people	7
clothing	5
furniture_rooms	5
locations	5
sounds	5
games_routines	4
outside	4
places	3
time_words	3
vehicles	3
connecting_words	1

Table 2: Number of good cross-linguistic words by semantic category.

compare to short form use average IRT parameter per item to calculate ability rather than sumscore

Discussion

Acknowledgements

We would like to thank all of the contributors to Wordbank, from the researchers who created and adapted the CDIs to those who collected the data (as well as the participants), to those who have created and maintained Wordbank over the years.

References

10 Baker, F. B. (2001). *The basics of item response theory*. ERIC.

Bates, E., Marchman, V., Thal, D., Fenson, L., Dale, P. S., Reznick, J. S., ... Hartung, J. (1994). Developmental and stylistic variation in the composition of early vocabulary. *Journal of Child Language*, *21*(1), 85–123.

Chai, J. H., Lo, C. H., & Mayor, J. (2020). A Bayesian-inspired item response theory-based framework to produce very short versions of MacArthur-Bates Communicative

Development Inventories. *Journal of Speech, Language, and Hearing Research*, *63*(10), 3488–3500.

Embretson, S. E., & Reise, S. P. (2013). *Item response theory*. Psychology Press.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., … Stiles, J. (1994). Variability in early communicative development. *Monographs of the Society for Research in Child Development*, i–185.

Fenson, L., Marchman, V. A., Thal, D. J., Dale, P. S., Reznick, J. S., & Bates, E. (2007). *MacArthur-Bates Communicative Development Inventories: User's guide and technical manual (2nd ed.)*. Baltimore, MD: Brookes.

Fenson, L., Pethick, S., Renda, C., Cox, J. L., Dale, P. S., & Reznick, J. S. (2000). Short-form versions of the MacArthur Communicative Development Inventories, *21*, 95–116.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2017). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677.

Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2021). *Variability and consistency in early language learning: The wordbank project*. MIT Press.

Kachergis, M., G., & Frank, M. C. (2022). Online computerized adaptive tests of children's vocabulary development in english and mexican spanish. *Journal of Speech, Language, and Hearing Research*, *65*(6), 2288–2308.

Makransky, G., Dale, P. S., Havmose, P., & Bleses, D. (2016). An item response theory-based, computerized adaptive testing version of the MacArthur-Bates Communicative Development Inventory: Words & Sentences (CDI:WS). *Journal of Speech, Language, and Hearing Research*, *59*(2), 281–289.

Mayor, J., & Mani, N. (2019). A short version of the MacArthur-Bates Communicative Development Inventories with high validity. *Behavior Research Methods*, *51*(5), 2248–2255.

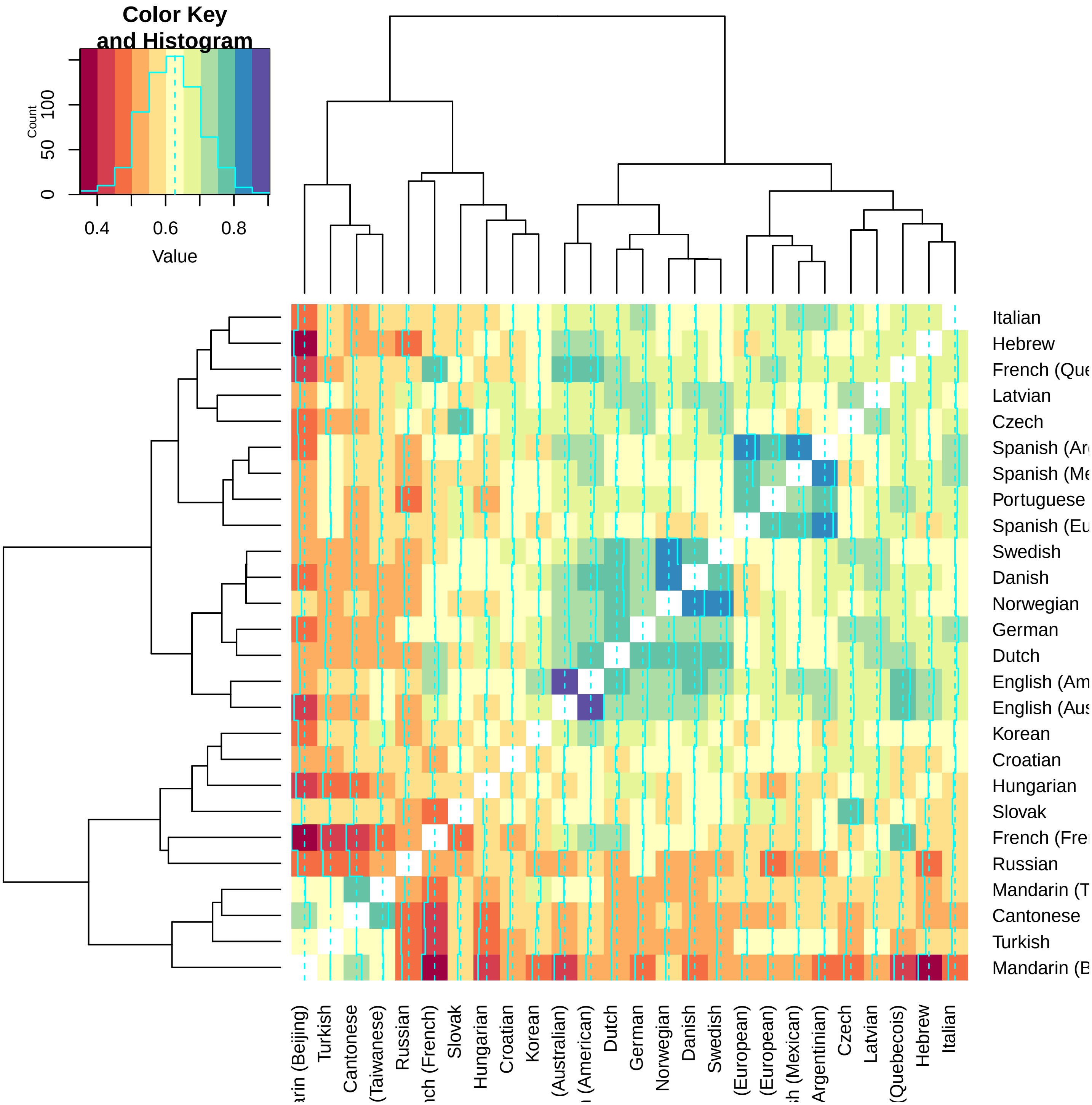


Figure 1: Cross-linguistic similarity (Spearman correlation) of IRT item difficulty from the CDI:WS.

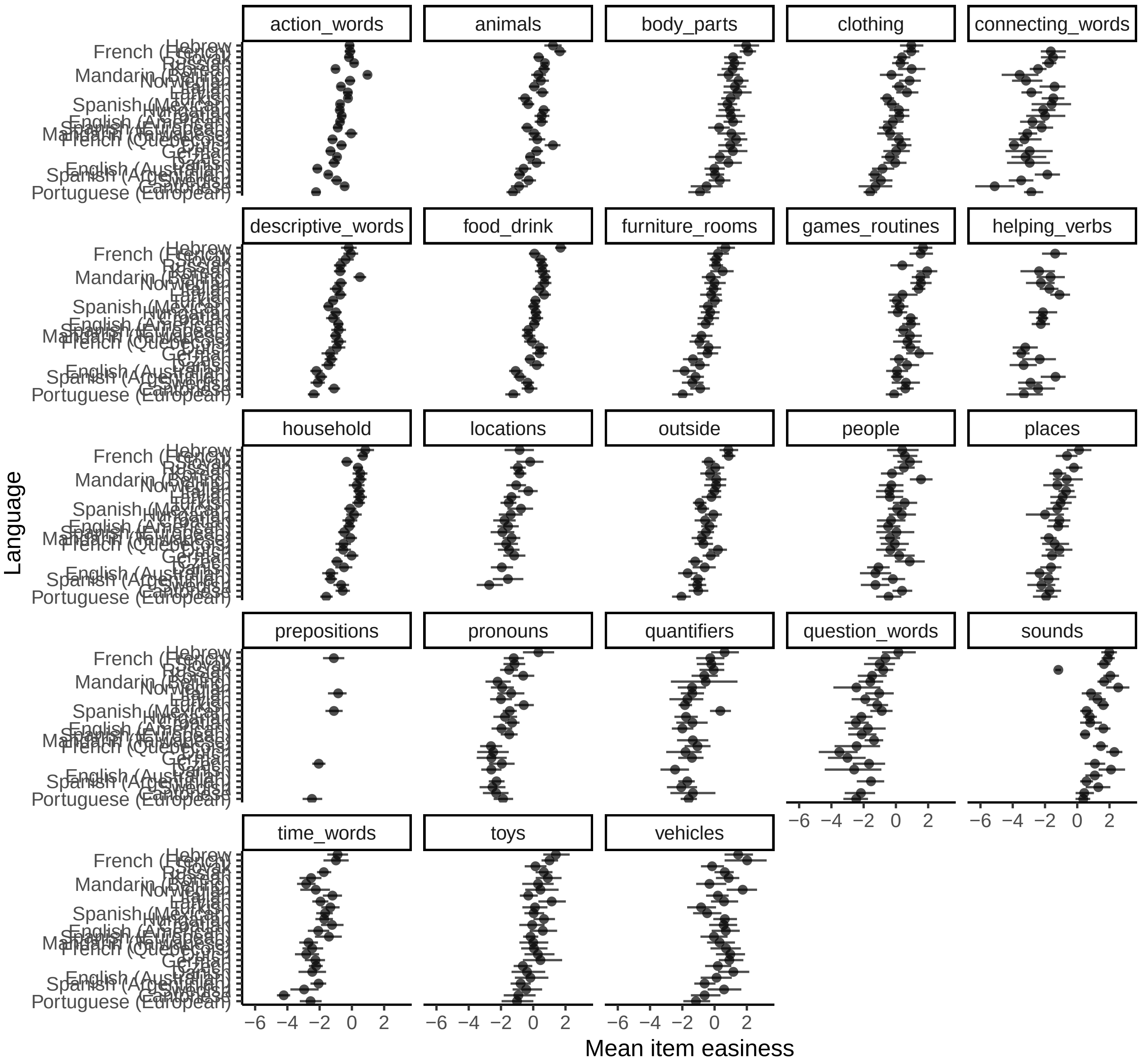


Figure 2: Mean difficulty of CDI words by semantic category. Bars represent bootstrapped 95% confidence intervals.