Development of Linguistic-Mediated Abstraction: Insights from Word Ladders task

Caterina Villani (caterina.villani6@unibo.it)
University of Bologna, Department of Modern Languages, Literatures, and Cultures, via Cartoleria 5
Bologna, 40124 Italy

Adele Loia (adele.loia2@unibo.it)
University of Bologna, Department of Modern Languages, Literatures, and Cultures, via Cartoleria 5
Bologna, 40124 Italy

Marianna Marcella Bolognesi (m.bolognesi@unibo.it)
University of Bologna, Department of Modern Languages, Literatures, and Cultures, via Cartoleria 5
Bologna, 40124 Italy

Abstract

What is the developmental trajectory of language-mediated abstraction skills, and to what extent are these skills influenced by semantics? We address these questions asking children to generate semantic relations of categorical inclusion for words varying in concreteness. Results show that abstraction improves over time, independent of age, with both concrete and abstract concepts organized into hierarchical taxonomies. However, abstract concepts allow shorter ladders, making them harder to categorize, especially for younger children. These findings underscore the distinction between concreteness and specificity as separate dimensions of abstract reasoning, and they lend empirical support to theoretical models that treat these facets of abstraction as dissociable.

Keywords: Abstraction; language development; concreteness; specificity; Word Ladders task.

Introduction

Abstraction, a hallmark of human cognition, is the ability to form general concepts or ideas by extracting similarities and general tendencies from direct experience, language, or other concepts (Reilly et al., 2024). Although concepts can be represented independently from words, linguistic labels often act as cues (Lupyan, 2012; Lupyan & Lewis, 2019) that help to create and organize our knowledge into coherent categories. Crucially, words can define different types of categories, spanning from very concrete to very abstract ones ("bike" vs "religion"). Moreover, the same category can be represented at different levels of precision or inclusiveness, from very specific to very general ("electric bike", "Hinduism" vs "vehicle", "belief").

Word concreteness and specificity are two interrelated yet distinct mechanisms underpinning abstraction. However, most studies of abstraction have examined these factors in isolation, paying little attention to how they interact in conceptual processing, language acquisition, and human categorization.

Regarding specificity, classic works on categorical organization in the human mind have primarily focused on concrete categories, demonstrating that basic-level terms, those neither overly general nor overly specific, are more readily comprehended and processed (Rosch et al., 1976; Hajibayova, 2013 for a review). Developmental studies show that categories at different levels of abstraction are acquired at distinct stages. Early word learning favors basic-level categories (e.g., Clark, 1993; Mervis & Crisafi, 1982), while conventional lexical forms of taxonomic categories —both superordinate and subordinate—typically emerge later, around 7–8 years of age (Lucariello & Nelson, 1985; Nelson, 1988). Recent studies reveal that a richer knowledge of very general (superordinate) categories predicts faster vocabulary growth (Lewis et al., 2021; Rissman & Lupyan, 2023).

Regarding concreteness, research on abstract knowledge indicates that abstract words like "democracy" are harder to learn than concrete words like "hammer" (Gleitman et al., 2005). Overall, abstract words are processed more slowly and less accurately than concrete ones (Paivio, 1991; Schwanenflugel, 1991; Kroll & Merves, 1986; Schwanenflugel, Harnishfeger & Stowe, 1988), are acquired later (Gilhooly & Logie, 1980; Della Rosa et al., 2010; Villani et al., 2019), and often collect multiple situations and experiences under one label rather than point to a single, bounded referent (Davis, Altman, & Yee, 2020). Moreover, studies suggest that emotional valence affects the acquisition of abstract words, with positive and negative abstract words being acquired earlier than neutral ones (Kousta et al., 2011; Ponari et al., 2018; Vigliocco et al., 2009). Most studies on abstract and concrete concepts, however, did not control word specificity, raising the question of whether this variable introduces a confound.

Recent studies report a mild positive correlation between concreteness and specificity ratings (Bolognesi et al., 2020; Bolognesi & Caselli, 2023; Ravelli et al., 2024) and a negative effect of specificity on semantic decision latencies (the more a word is specific, the faster it is processed), with no interaction with concreteness (Lamarra, Villani, & Bolognesi, 2024), suggesting that the two variables support different mechanisms in conceptual processing. A semantic content analysis of words balanced for concreteness and specificity shows that semantic categories are distributed along a continuum of abstraction ladder, with both concrete and abstract categories yielding either long taxonomies (e.g., animal kingdom, social entities) or shorter ones (e.g., natural elements, logical relations) (Villani et al., 2024). While

In D. Barner, N.R. Bramley, A. Ruggeri and C.M. Walker (Eds.), *Proceedings of the 47th Annual Conference of the Cognitive Science Society*

promising, these findings are based on subjective judgments rather than speakers' productions and, therefore, do not accurately reflect the organization of their mental lexicon.

The present study uses a generative approach in which participants are asked to construct ladders of words linked by genericity/specificity relations, to investigate how the active vocabulary of words varying in specificity mastered by school kids evolves over the span of one school year, for different types of concrete and abstract concepts.

This Study

We investigated the development of language-mediated abstraction skills in a longitudinal study, asking children aged 9-13 to produce hyponyms and hypernyms for a set of target words, at the beginning and the end of the school year. We operationalized abstraction skills in terms of *productivity* (i.e., sum of words produced) and *validity* (i.e., sum of valid words in the taxonomy) in word ladders to investigate (1) how language-mediated abstraction ability unfolds across ages and over time, and (2) how word semantics influence children's ability to construct ladders and therefore to perform language-mediated abstractions over time/age. We tackled these research questions in two analyses, using a printed version of Word Ladders, a mobile app developed by the Abstraction research group to collect linguistic data on the semantic relation of categorical inclusion, also referred to as hypernymy/hyponymy in linguistics, or IS-A relation.

Method

Participants A total of 195 children, aged 9 to 13, were recruited from a primary and middle school in Imola, Italy. 37 children with special educational needs or learning disorders were excluded, leaving a final sample of 158. This included two fifth-grade classes (N=41, 20 female, age 9-10), two sixth-grade classes (N=37, 22 female, age 10-11), two seventh-grade classes (N=41, 22 female, age 11-12) and two eighth-grade classes[1] (N=42, 30 female, age 12-13). The study was approved by the Ethics Committee of the University of Bologna. Informed consent and privacy policy were signed by the school principal following approval obtained from children's parents or caregivers.

Materials The stimuli were 16 Italian nouns including 4 concrete words ("cake", "toy", "car", "horse"), 4 abstract words ("freedom", "time", "mind", "dream"), 4 emotional words ("joy", "affection", "fear", "anger") and 4 neutral social role words ("teacher", "doctor", "driver", "caregiver"). Concrete, abstract, and emotional words were selected from the ANEW Italian database (Montefinese et al., 2014) based on familiarity ratings >7. Social role words, not in the ANEW database, were chosen by the experiments from grammatically neutral Italian occupational terms, as they denote human referents with abstract, socially defined identities. Table 1 in the Appendix shows the descriptive statistics of the stimuli selected from the ANEW database, along with results of ANOVA and post-hoc comparisons on psycholinguistic (familiarity and frequency) and semantic variables (imageability, concreteness, valence, and arousal).

Procedure Each class was tested individually at the beginning (T1: November 2023) and at the end of the school year (T2: May 2024). After a brief welcome, the experimenters verbally introduced the task and explained the semantic relation IS-A using a graphic tutorial of the Word Ladders app (Figure 1A). Pupils were informed that they would participate in a game involving the creation of word ladders. They were provided with paper protocols containing 16 words, each embedded in a ladder (Figure 1B). The aim was to construct the longest word ladders by adding steps above and below the initial prompt. For example, given the word "dog", they should add on the steps above increasingly general words like "mammal, "animal", "living being", and more specific ones on the steps below, such as "labrador". Instructions highlighted that the words added should be more general than the previous ones and should describe "a type of." Examples with correct and incorrect semantic relations were provided. For instance, "bread" is a type of "food" and "baguette" is a specific type of bread, but "flour" is not. Similarly, "number" is a type of "quantity" and "negative numbers" are a specific type of number, but "calculator" is not. The same target words and instructions were used in both experimental sessions, with word order randomized for each child. Each child performed the task individually, with the experimenters and teacher supervising without intervening. Instructions highlighted that longer ladders could earn more points, but only words linked by the "is a type of" relation would score. Pupils were told that ladder lengths varied and that completing every step was optional; each session lasted approximately one hour.

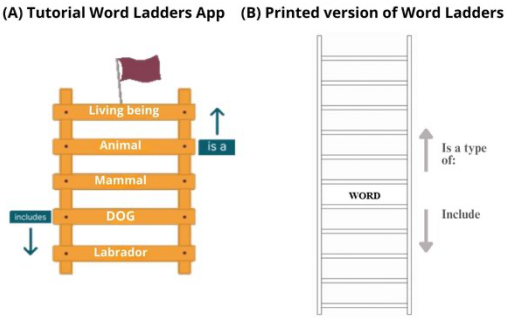

Figure 1: Illustration of Word Ladders task in mobile app (A) and a printed (B) version of experimental protocol.

[1] At T2, only one of the two eighth-grade classes participated in the experimental session due to school organizational constraints. The other class attempted to provide data under the sole supervision of the classroom teacher, who independently conducted the data collection. However, because of substantial missing data (70 %) and related methodological concerns, these data were excluded from all analyses.

Coding The responses were manually entered into a spreadsheet by one of the authors. The words above the target word were coded as hypernyms, while those below were coded as hyponyms. Demographic information (gender, age) was also recorded. For each ladder, we calculated two measures: *raw_length,* the sum of hypernyms and hyponyms produced in response to a word (including the target word), indicating lexical productivity, and *clean_length,* the sum of valid words in the ladder (including the target word), indicating the validity of the taxonomy. For instance, given the target word "cake", the words "food", "nourishment", "physical energy", and "help" were produced as hypernyms, while "tiramisù" and "sugar" were produced as hyponyms. The raw_length is 7 while the clean_length is 4, corresponding to the valid words in the taxonomy: "nourishment", "food", "cake", "tiramisù". The coding of valid words to obtain *clean_length* was done separately for the first and second data collections. In the first data collection, a primary coder initially coded all data using WordNet taxonomies (Miller et al., 1990; Fellbaum, 1998;) and common-sense classifications (see Coding Scheme in the OSF repository), with a second coder, naïve to the study aim, independently coding all the data for inter-rater reliability. Cohen's Kappa indicated a high reliability (unweighted κ=0.84, weighted κ=0.95) (Cohen, 1960). For the second data collection, two coders independently coded 50% of the data. Any disagreements were resolved through consensus.

Data Analysis Data analysis and visualizations of the two analyses were carried out with R (RCoreTeam, 2019) and RStudio (v4.2.3). We modeled the *raw_length* (productivity) and *clean_length* (validity) for each word ladder separately with a mixed effect binomial logistic regression using gmler() function from "lmer4" R package (Bates et al., 2015). Significance of the main effects and interactions were assessed with Wald Chi-Square tests using Anova () function from "car" R's package (Fox & Weisberg, 2011). Post-hoc contrasts were carried out with the "emmeans" R package (Lenth et al., 2024), using Tukey's adjustment for multiple comparisons. All data and scripts are available on OSF: https://osf.io/qbr6j/

Analysis 1

Analysis 1 tests whether abstraction skills improve over time and if this improvement correlates with children's ages. Given that conceptual representation is flexible and develops over time, we expected variations in word ladders created at the beginning and the end of the school year, reflecting performance improvement. Additionally, we hypothesized a positive correlation between age and the productivity and validity of ladders. Specifically, as children age increases, their ladders should become longer and more valid. We analyzed *raw_length* and *clean_length* in two separate models, with Time of data collection (T1 vs. T2) as a fixed effect, and participants and words as random intercepts. Spearman correlations were then conducted to explore the relationship between Age (9–10, 10–11, 11–12, 12–13) and each dependent variable, first on the overall data and then separately for T1 and T2.

Results and Discussion

Time We found a significant main effect of Time on both the productivity and validity of word ladders. Raw length was significantly higher at T2 (M = 5.06, SD = 1.89) than T1 (M = 4.43, SD = 1.89), $\chi^2(1) = 69.148$, $p < .001$. Similarly, clean length was higher at T2(M = 3.89, SD = 1.75) than T1 (M = 2.75, SD = 1.35), $\chi^2(1) = 414.35$, $p < .001$.

Age Overall, we found a very small positive correlation between age and raw length (r = 0.04), and a very small negative correlation between age and clean length (r = -0.01). At T1, we found a slightly strong positive correlation between age and raw length (r = 0.08), and a weak positive correlation with clean length (r = 0.04). At T2, the positive correlation between age and raw length drops to r = 0.02, and the correlation between age and clean length approaches zero (r = -0.01). In line with our first hypothesis, we found a clear improvement in both productivity and validity of ladders created by children from the beginning to the end of the school year. Our second hypothesis is only partially supported. Results show weak correlations between children's age and the productivity and validity of word ladders across both time points.

Analysis 2

Analysis 2 tests whether semantics affect children's performance on the Word Ladders task across ages and over time. In line with previous studies showing an early acquisition of concrete over abstract words, we hypothesized that ladders based on concrete words will be longer and more valid than those based on abstract words. Emotion words are learned earlier and perceived as less abstract than neutral abstract words. Similarly, social role-related words combine both concrete and abstract elements, representing human beings who embody various social phenomena. Considering these factors, it may be easier to construct ladders on emotional words and social role-related words compared to more abstract words. These effects may also vary by age and time: younger children might initially find it more challenging to create long and valid ladders for abstract words than for concrete words, and their improvement over time may be more pronounced for abstract words. To test these hypotheses, we expanded our analysis by modeling the dependent variable *raw_length* and *clean_length* separately, using Age (9-10, 10-11, 11-12, 12-13), Semantic Type (abstract, concrete, social role, emotions), Time (T1 vs. T2), and their interaction as fixed effects, with participants and words as random effects.

Results and Discussion

Productivity: Raw Length We found a significant main effect of Semantic Type, $\chi^2(3) = 80.5195$, $p < .001$, and Time, $\chi^2(1) = 68.0861$, $p < .001$, and significant interactions between Age and Semantic Type, $\chi^2(9) = 24.1942$, $p = .004$, Age and Time, $\chi^2(3) = 29.3265$, $p < .001$, and Semantic Type

and Time, $\chi^2(3) = 10.6322$, $p = .013$. No other main effect or interaction reached significance, all $p_s = .06$.

Post-hoc contrasts: Semantic Types Children produced shorter ladders for abstract concepts than for concrete, $z = -7.918$, $p < .0001$, and social role concepts, $z = -6.738$, $p < .0001$. They also generated shorter ladders for emotional concepts than for social role ones, $z = -4.615$, $p < .0001$. In contrast, they produced longer ladders for concrete than for emotional concepts, $z = 5.797$, $p < .0001$. No difference was found between abstract and emotional concepts, $z = -2.130$, $p = .14$, and between concrete and social role concepts, $z = 1.183$, $p = .637$ (Figure 2).

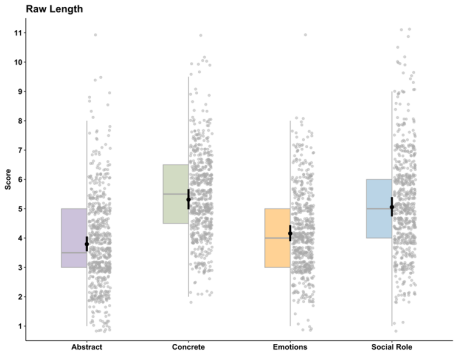

Figure 2: Raw Length scores (ladder productivity) as a function of semantic type. Raw data in the background are aggregated over participants.

Post-hoc contrasts: Semantic Type and Age Children aged 9-10 generated shorter ladders for abstract concepts compared to concrete, $z = -7.714$, $p < .0001$, emotions, $z = -2.622$, $p = .0434$, and social role concepts, $z = -5.425$, $p < .0001$. They also produced shorter ladders for emotional concepts compared to social role concepts, $z = -2.815$, $p = .02$. In contrast, they produced longer ladders for concrete concepts compared to emotional ones, $z = 5.122$, $p < .0001$. There was no difference in ladders productivity between concrete and social role concepts, $p = .09$. For children aged 10-11, 11-12, and 12-13, a consistent pattern emerged. In all three age groups, ladders for abstract concepts were shorter than those for concrete concepts (10-11$_{\text{year-olds}}$ $z = -4.385$, $p < .0001$; 11-12$_{\text{year-olds}}$ $z = -4.974$, $p < .0001$; 12-13$_{\text{year-olds}}$ $z = -7.269$, $p < .0001$) and social role concepts (10-11$_{\text{year-olds}}$ $z = -4.236$, $p < .0001$; 11-12$_{\text{year-olds}}$ $z = -4.699$, $p < .0001$; 12-13$_{\text{year-olds}}$ $z = -6.287$, $p < .0001$). Ladders for concrete concepts were longer than those for emotional concepts (10-11$_{\text{year-olds}}$ $z = 3.097$, $p = .0105$; 11-12$_{\text{year-olds}}$ $z = 4.552$, $p < .0001$; 12-13$_{\text{year-olds}}$ $z = 5.264$, $p < .0001$), while ladders for emotional concepts were shorter than for social role ones (10-11$_{\text{year-olds}}$ $z = -2.949$, $p = .0168$; 11-12$_{\text{year-olds}}$ $z = -4.278$, $p = .0001$; 12-13$_{\text{year-olds}}$ $z = -4.273$, $p = .0001$). Across these age groups, there were no differences in ladders productivity between abstract and emotional concepts (10-11$_{\text{year-olds}}$ $p = .56$; 11-12$_{\text{year-olds}}$ $p = .97$; 12-13$_{\text{year-olds}}$ $p = .17$) or between concrete and social role concepts (10-11$_{\text{year-olds}}$ $p = .1$; 11-12$_{\text{year-olds}}$ $p = .99$; 12-13$_{\text{year-olds}}$ $p = .75$). We also found that ladders for abstract concepts were significantly shorter in 9-10$_{\text{year-olds}}$ compared to 10-11$_{\text{year-olds}}$, $z = -3.370$, $p = .0042$, and 11-12$_{\text{year-olds}}$, $z = -3.235$, $p = .0067$. Conversely, ladders for abstract concepts were longer in 10-11$_{\text{year-olds}}$, $z = 2.938$, $p = .0174$, and 11-12$_{\text{year-olds}}$, $z = 2.794$, $p = .0267$, compared to 12-13$_{\text{year-olds}}$. Instead, there were no differences between age groups in terms of ladders productivity for concrete, $p_s > .95$, social role, $p_s > .06$, and emotional concepts, $p_s > 13$.

Post-hoc contrasts: Age and Time At T1, ladders were significantly shorter in 9-10$_{\text{year-olds}}$ compared to 10-11$_{\text{year-olds}}$ $z = -3.313$, $p = .005$, and 12-13$_{\text{year-olds}}$ $z = -2.568$, $p = .05$. No other comparisons reached significance, all $p_s > .26$. At T2, ladders were significantly shorter in 11-12$_{\text{year-olds}}$ compared to 12-13$_{\text{year-olds}}$ $z = 3.360$, $p = .0043$. No other comparisons reached significance, $p_s > .07$. We also found that ladders were significantly longer at T2 compared to T1 in 9-10$_{\text{year-olds}}$, $z = -5.676$, $p < .0001$, 10-11$_{\text{year-olds}}$ $z = -2.466$, $p = .01$, and 11-12$_{\text{year-olds}}$, $z = -7.247$, $p < .0001$. No difference was found between T1 and T2 in 12-13$_{\text{year-olds}}$, $p = .51$.

Post-hoc contrasts: Semantic Type and Time Children produced longer ladders at T2 than T1 for concrete, $z = -3.994$, $p = .0001$, emotions, $z = -3.119$, $p = .0018$, and social role concepts, $z = -6.138$, $p < .0001$. No difference was found for abstract concepts, $p = .23$. At both times, children produced shorter ladders for abstract concepts than for concrete (T1: $z = -6.554$, $p < .0001$; T2: $z = -7.731$, $p < .0001$) and social role concepts (T1: $z = -4.795$, $p < .0001$; T2: $z = -7.316$, $p < .0001$). They also produced shorter ladders for emotional concepts than for social role concepts, T1: $z = -3.444$, $p = .003$; T2: $z = -4.895$, $p < .0001$, while generating longer ladders for concrete concepts compared to emotional ones (T1: $z = 5.205$, $p < .0001$; T2: $z = 5.311$, $p < .0001$). No differences were found between concrete and social role concepts (T1: $p = .29$; T2: $p = .97$) and between abstract and emotional concepts (T1: $p = .52$; T2: $p = .06$) (Figure 3).

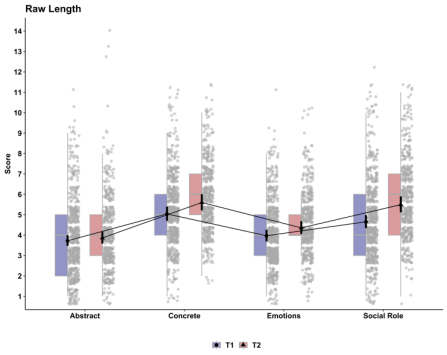

Figure 3: Raw Length scores (ladder productivity) for each semantic type as a function of time. Raw data in the background are aggregated over participants.

Validity: Clean Length We found significant main effects of Semantic Type, $\chi^2(3) = 69.1481$, $p < .001$, and Time, $\chi^2(1) = 410.2634$, $p < .001$, and significant interactions between Age and Time, $\chi^2(3) = 11.1227$, $p = .011$, and Semantic Type

and Time, $\chi^2(3) = 9.0108$, $p = .029$. No other main effects or interactions reached significance, all $p_s > .52$.

Post-hoc contrasts: Semantic Type Children produced fewer valid words in ladders for abstract concepts than for concrete, $z = -7.457$, $p < .0001$, emotional, $z = -2.602$, $p = .04$, and social role concepts, $z = -6.272$, $p < .0001$. In contrast, they produced more valid words in the ladders for concrete concepts compared to emotional concepts, $z = 4.866$, $p < .0001$, and fewer valid words in the ladders for emotional concepts than social role concepts, $z = -3.676$, $p = .0014$). There was no difference in the validity of ladders between concrete and social role concepts, $p = .63$ (Figure 4).

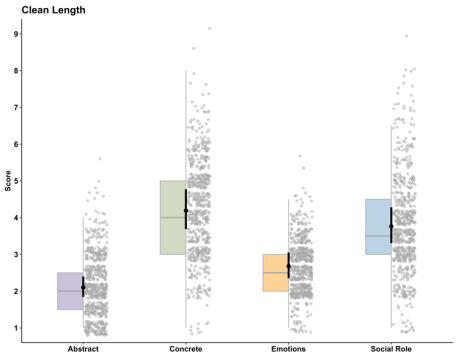

Figure 4: Clean Length scores (ladder validity) as a function of semantic types. Raw data in the background are aggregated over participants.

Post-hoc contrasts: Age and Time We found no difference in the validity of ladders between the age groups at either time point, T1 $p_s > .26$ and T2 $p_s > .21$. However, the number of valid words in the ladders improved significantly at T2 compared to T1 in all age groups: 9-10$_{\text{year-olds}}$ $z = -10.140$, $p < .0001$; 10-11$_{\text{year-olds}}$ $z = -11.199$, $p < .0001$; 11-12$_{\text{year-olds}}$ $z = -12.136$, $p < .0001$; 12-13$_{\text{year-olds}}$ $z = -5.720$, $p < .0001$.

Post-hoc contrasts: Semantic Type and Time Ladders validity increased significantly from T1 to T2 for each semantic type, all $p_s < .0001$. At both times, children produced fewer valid words in ladders for abstract concepts than for concrete concepts, T1: $z = -7.710$, $p < .0001$; T2: $z = -6.675$, $p < .0001$, and social role concepts, T1: $z = -5.962$, $p < .0001$; T2: $z = -6.121$, $p < .0001$. They also produced more valid words in ladders for concrete concepts than for emotional concepts, T1: $z = 4.954$, $p < .0001$; T2: $z = 4.476$, $p < .0001$, and fewer valid words for emotional concepts than for social role concepts, T1: $z = -3.197$, $p = .007$; T2: $z = -3.919$, $p = .0005$. No difference was found in the validity of ladders between concrete and social role concepts at either time point, T1: $p = .29$; T2: $p = .94$. However, a key difference emerged: at T1, children produced fewer valid words in ladders for abstract concepts compared to emotional concepts, $z = -2.776$, $p = .0282$, whereas at T2, such difference was not observed, $p = .1198$ (Figure 5).

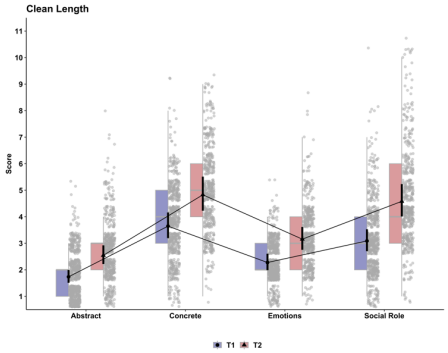

Figure 5: Clean Length scores (ladder validity) for each semantic type as a function of time. Raw data in the background are aggregated over participants.

Overall, as predicted, children produced longer and more valid ladders for concrete concepts than for abstract concepts. Ladders for social role concepts were similar in length and validity to those for concrete ones, but longer and more valid than those for abstract and emotional concepts. Ladders for emotional concepts were shorter than those for concrete ones, with similar productivity but higher validity than ladders for abstract concepts. A similar pattern emerged across the two data collection times. However, while the validity of ladders increased from T1 to T2 for all semantic types, the ladders productivity increased only for concrete, emotional, and social role concepts, while abstract concepts consistently yielded shorter ladders. Interestingly, at T2, no difference was found between the validity of ladders for abstract and emotional concepts. Notably, we found age differences primarily in ladders involving abstract concepts. While there was no difference in ladders productivity for concrete, emotional, and social role concepts across age groups, the youngest children (9-10) created shorter ladders for abstract concepts than emotional ones, a pattern not seen in older groups (10-13). The youngest children (9-10) also produced shorter ladders overall than older children (10-11 and 11-12), showing no difference compared to the oldest group (12-13). In contrast, 12-13-year-olds produced longer ladders overall than the 10-11 and 11-12 groups, with no significant differences between the latter two. Consistent with this, ladders productivity improves from T1 to T2, with significant increases in younger children, but no improvement in the oldest group (12-13). Contrary to our predictions, we found no significant interaction between semantics and the validity of ladders, suggesting a similar pattern in all age groups. Nevertheless, ladders validity increased from T1 to T2 for all age groups, with no significant differences between them.

General Discussion

This study explored the development of language-mediated abstraction through the Word Ladders task, where school-age children generated semantic relations of categorical inclusion for a set of words, at two points during the school year. We predicted that abstraction skills, operationalized as ladders

productivity and validity, would improve with increasing age and time, and vary by the semantics of the prompts.

Analysis 1 confirms that abstraction skills improve over time, with ladders created at the end of the school year being more productive and more valid than those at the beginning. This suggests that children become better at identifying valid taxonomic relationships as they progress in education. Surprisingly, these skills do not correlate with age. The lack of an age effect may be due to task training throughout the year, which helps younger children develop skills similar to those of older peers. Additionally, the Word Ladders task seems more sensitive to word meaning than to age-related cognitive differences, with age only influencing performance in relation to semantics, as shown in the second analysis.

Analysis 2 shows that semantics affects children's language-mediated abstractions. As expected, ladders for concrete concepts were longer and more valid than other semantic types, while abstract concepts showed the opposite pattern. Emotional and social-role concepts produced intermediate results: ladders for emotional concepts were more valid but equally productive as those for abstract ones, while ladders for social role concepts were comparable to concrete concepts. This suggests that the abstraction process is easier with words tied to real-world experience, while identifying hierarchical relations for abstract concepts remains challenging (see Banks & Connell, 2022; Persichetti et al., 2024). This aligns with theories classifying abstract concepts as "hard words" to acquire and process, and with findings that semantic categories characterized by different levels of concreteness also vary along the specific-general continuum, resulting in either long or short taxonomies (Villani et al., 2024). Crucially, we replicate and extend this research using an innovative generative task from a developmental perspective.

Our findings also contribute to language development research, particularly on abstract language. Indeed, we found that while all children showed similar ladders productivity for concrete, emotional, and social concepts, the youngest children (9-10) produced shorter ladders for abstract concepts than for emotional concepts, a pattern absent in older children (10-13). Regarding ladders validity, we found no age-related differences across semantic types, indicating that children of all ages can identify meaningful taxonomic relations, regardless of concept type. Yet, abstract and emotional concepts yielded less valid ladders than concrete and social role concepts. Interestingly, by the second data collection, the ladders validity for abstract concepts was comparable to emotional ones. This suggests that abstract thinking develops later than emotional processing, and as children's emotional vocabulary refines, their ability to form complex associations with abstract concepts improves. These results align with the "Abstract-via-Emotion" hypothesis (Kousta et al., 2011; Vigliocco et al., 2009; Ponari et al., 2018), which posits that emotional states provide a bootstrap mechanism for the development of abstract words and concepts.

Methodologically, our study has significant implications. Instead of relying on Likert-scale ratings to collect linguistic data, we used a gamified approach through the Word Ladders task. This task enhances students' ability to retrieve both specific and general words from memory, bolstering their linguistic proficiency. Indeed, productivity and validity of the ladders improved across two data collection periods, with variations by semantic type and age group. Improvements were especially evident for concrete, emotional, and social role concepts, reflecting advances in children's taxonomic knowledge. However, abstract concepts showed less improvement due to their inherently shorter taxonomic ladders. Notably, younger children demonstrated the most significant progress, likely due to ongoing cognitive and linguistic development. In contrast, the oldest group (12–13) showed no notable improvement, potentially due to the consolidation of their cognitive abilities and stabilized abstract language development. Nonetheless, ladders validity improved across all age groups and semantic types, indicating overall progress in generating valid taxonomic ladders, even for abstract concepts. These improvements can be attributed to the "training effect" of the Word Ladders task, which fosters abstraction and linguistically mediated skills, an outcome often difficult to achieve with traditional methods.

Theoretically, our findings add layers to studies on conceptual knowledge, showing that, like concrete concepts, abstract concepts also vary in specificity. While taxonomies for abstract concepts tend to be shorter, researchers should account for this factor, particularly in studies comparing the processing of abstract and concrete words and concepts, where specificity is often overlooked (see, Bolognesi et al., 2020, for a discussion).

Certainly, this work is not without limitations. While we focused on categorical inclusion, the data excluded from the coding of valid words (*clean_length*) contained other semantic relations, such as free associations ("time"-"clock") and part-whole relations (e.g., "car"-"wheel"). In the future, we aim to analyze these relations to better understand the role of language in abstraction processes, particularly in generalization and linguistic compositionality. Another limitation concerns the small number of words selected for balanced familiarity and time constraints, as well as the lack of other socio-demographic factors (e.g., kids' reading habits), which we are addressing by scaling up data collection via the Word Ladders mobile application[2] to include a larger sample of Italian and English speakers from various cultural and educational backgrounds.

To conclude, our findings contribute to the literature on language development, showing that the ability to perform conceptual abstraction improves over time, regardless of age. Crucially, both concrete and abstract concepts support the development of hierarchical taxonomies. However, abstract concepts allow the development of short taxonomic ladders, making them more challenging to categorize, particularly for younger children.

[2] Android download: http://shorturl.at/kqAOS
Apple download: http://shorturl.at/buIS2

Acknowledgments

The authors would like to thank all members of ABSTRACTION research group (GRANT AGREEMENT: ERC-2021-STG-101039777) for comments and discussions. They are, in alphabetical order, Andrea Amelio Ravelli, Giulia Rambelli, and Tommaso Lamarra. The authors would also like to thank all the students and teachers of the IC7 public school in Imola (BO) for participating in this study.

Fundings: Caterina Villani and Marianna M. Bolognesi were funded by ABSTRACTION (GRANT AGREEMENT: ERC-2021-STG-101039777). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them. Adele Loia was funded by PNRR - PRIN 2022 (2022EPTPJ9) "WEMB: Word Embeddings from Cognitive Linguistics to Language Engineering and back", financed by the European Union – Next Generation EU, Mission 4 Component 2 Investment 1.1 CUP J53D23007100001.

References

Banks, B., & Connell, L. (2022). Category production norms for 117 concrete and abstract categories. *Behavior Research Methods*, *55*(3), 1292–1313. https://doi.org/10.3758/s13428-021-01787-z

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

Bolognesi, M., Burgers, C., & Caselli, T. (2020). On abstraction: Decoupling conceptual concreteness and categorical specificity. *Cognitive Processing*, *21*(3), 365–381. https://doi.org/10.1007/s10339-020-00965-9

Bolognesi, M. M., & Caselli, T. (2023). Specificity ratings for Italian data. *Behavior Research Methods*, *55*(7), 3531–3548. https://doi.org/10.3758/s13428-022-01974-6

Clark, E. V. (1993). *First language acquisition* (1. publ., 6. print). Cambridge Univ. Press.

Cohen, J. (1960). A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, *20*(1), 37–46. https://doi.org/10.1177/001316446002000104

Davis, C. P., Altmann, G. T. M., & Yee, E. (2020). Situational systematicity: A role for schema in understanding the differences between abstract and concrete concepts. *Cognitive Neuropsychology*, *37*(1–2), 142–153. https://doi.org/10.1080/02643294.2019.1710124

Della Rosa, P. A., Catricalà, E., Vigliocco, G., & Cappa, S. F. (2010). Beyond the abstract—concrete dichotomy: Mode of acquisition, concreteness, imageability, familiarity, age of acquisition, context availability, and abstractness norms for a set of 417 Italian words. *Behavior Research Methods*, *42*(4), 1042–1048. https://doi.org/10.3758/BRM.42.4.1042

Fellbaum, C. (1998). *WordNet: An Electronic Lexical Database*. MIT Press.

Fox, J., & Weisberg, S. (2011). *An R Companion to Applied Regression*. SAGE.

Gilhooly, K. J., & Logie, R. H. (1980). Age-of-acquisition, imagery, concreteness, familiarity, and ambiguity measures for 1,944 words. *Behavior Research Methods & Instrumentation*, *12*(4), 395–427. https://doi.org/10.3758/BF03201693

Gleitman, L. R., Cassidy, K., Nappa, R., Papafragou, A., & Trueswell, J. C. (2005). Hard words. *Language Learning and Development*, *1*(1), 23–64. https://doi.org/10.1207/s15473341lld0101_4

Hajibayova, L. (2013). Basic-level categories: A review. *J. Inf. Sci.*, *39*(5), 676–687. https://doi.org/10.1177/0165551513481443

Kousta, S.-T., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: Why emotion matters. *Journal of Experimental Psychology: General*, *140*(1), 14–34. https://doi.org/10.1037/a0021446

Kroll, J. F., & Merves, J. S. (1986). Lexical access for concrete and abstract words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *12*(1), 92–107. https://doi.org/10.1037/0278-7393.12.1.92

Lamarra, T., Villani, C., & Bolognesi, M. (2024). *Specificity effect in concrete/abstract semantic categorization task*. https://osf.io/j2smv/

Lenth, R. V., Bolker, B., Buerkner, P., Giné-Vázquez, I., Herve, M., Jung, M., Love, J., Miguez, F., Piaskowski, J., Riebl, H., & Singmann, H. (2024). *emmeans: Estimated Marginal Means, aka Least-Squares Means* (Version 1.10.4) [Computer software]. https://cran.r-project.org/web/packages/emmeans/index.html

Lewis, M., Colunga, E., & Lupyan, G. (2021). *Superordinate Word Knowledge Predicts Longitudinal Vocabulary Growth*. https://doi.org/10.31234/osf.io/c8wdt

Lucariello, J., & Nelson, K. (1985). Slot-filler categories as memory organizers for young children. *Developmental Psychology*, *21*(2), 272–282. https://doi.org/10.1037/0012-1649.21.2.272

Lupyan, G. (2012). Linguistically Modulated Perception and Cognition: The Label-Feedback Hypothesis. *Frontiers in Psychology*, *3*. https://doi.org/10.3389/fpsyg.2012.00054

Lupyan, G., & Lewis, M. (2019). From words-as-mappings to words-as-cues: The role of language in semantic knowledge. *Language, Cognition and Neuroscience*, *34*(10), 1319–1337. https://doi.org/10.1080/23273798.2017.1404114

Mervis, C. B., & Crisafi, M. A. (1982). Order of Acquisition of Subordinate-, Basic-, and Superordinate-Level Categories. *Child Development*, *53*(1), 258–266. https://doi.org/10.2307/1129660

Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D., & Miller, K. J. (1990). Introduction to WordNet: An On-line Lexical Database*. *International Journal of*

Lexicography, *3*(4), 235–244. https://doi.org/10.1093/ijl/3.4.235
Montefinese, M., Ambrosini, E., Fairfield, B., & Mammarella, N. (2014). The adaptation of the Affective Norms for English Words (ANEW) for Italian. *Behavior Research Methods*, *46*(3), 887–903. https://doi.org/10.3758/s13428-013-0405-3
Nelson, K. (1988). Where Do Taxonomic Categories Come from? *Human Development*, *31*(1), 3–10. https://doi.org/10.1159/000273198
Paivio, A. (1991). Dual coding theory: Retrospect and current status. *Canadian Journal of Psychology / Revue Canadienne de Psychologie*, *45*(3), 255–287. https://doi.org/10.1037/h0084295
Persichetti, A. S., Shao, J., Denning, J. M., Gotts, S. J., & Martin, A. (2024). Taxonomic structure in a set of abstract concepts. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1278744
Ponari, M., Norbury, C. F., & Vigliocco, G. (2018). Acquisition of abstract concepts is influenced by emotional valence. *Developmental Science*, *21*(2), e12549. https://doi.org/10.1111/desc.12549
Ravelli, A. A., Bolognesi, M. M., & Caselli, T. (2024). Specificity ratings for English data. *Cognitive Processing*. https://doi.org/10.1007/s10339-024-01239-4
RCoreTeam. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.r-project.org/
Reilly, J., Shain, C., Borghesani, V., Kuhnke, P., Vigliocco, G., Peelle, J. E., Mahon, B. Z., Buxbaum, L. J., Majid, A., Brysbaert, M., Borghi, A. M., De Deyne, S., Dove, G., Papeo, L., Pexman, P. M., Poeppel, D., Lupyan, G., Boggio, P., Hickok, G., … Vinson, D. (2024). What we mean when we say semantic: Toward a multidisciplinary semantic glossary. *Psychonomic Bulletin & Review*. https://doi.org/10.3758/s13423-024-02556-7
Rissman, L., & Lupyan, G. (2023). *The Power of the Lexicon: Eliciting Superordinate Categories With and Without Labels*. OSF. https://doi.org/10.31234/osf.io/5xucp
Rosch, E., Mervis, C. B., Gray, W. D., Johnson, D. M., & Boyes-Braem, P. (1976). Basic objects in natural categories. *Cognitive Psychology*, *8*(3), 382–439. https://doi.org/10.1016/0010-0285(76)90013-X
Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? *The Psychology of Word Meanings*, *1991*.
Schwanenflugel, P. J., Harnishfeger, K. K., & Stowe, R. W. (1988). Context availability and lexical decisions for abstract and concrete words. *Journal of Memory and Language*, *27*(5), 499–520. https://doi.org/10.1016/0749-596X(88)90022-8
Vigliocco, G., Meteyard, L., Andrews, M., & Kousta, S. (2009). Toward a theory of semantic representation. *Language and Cognition*, *1*(2), 219–247. https://doi.org/10.1515/LANGCOG.2009.011
Villani, C., Loia, A., & Bolognesi, M. M. (2024). The semantic content of concrete, abstract, specific, and generic concepts. *Language and Cognition*, 1–28. https://doi.org/10.1017/langcog.2023.64
Villani, C., Lugli, L., Liuzza, M. T., & Borghi, A. M. (2019). Varieties of abstract concepts and their multiple dimensions. *Language and Cognition*, *11*(3), 403–430. https://doi.org/10.1017/langcog.2019.23

Appendix

Table 1. Descriptive statistics for target words used in the study grouped by semantic types (i.e., abstract, concrete, emotional concepts) with ANOVA results and post-hoc comparisons on psycholinguistic and semantic variables, i.e., frequency: CoLFIS (Bertinetto et al., 2005), Repubblica (Baroni et al., 2004), familiarity, imageability, concreteness, valence, arousal (Montefinese et al. 2014).

	Abstract concepts		**Concrete concepts**		**Emotional concepts**			
Measure	***Mean***	***sd***	***Mean***	***sd***	***Mean***	***sd***	***ANOVA (p-value)***	***Post-hoc Comparisons***
Ln_Colfis	6.62	1.06	5.17	1.18	5.57	0.67	$p = 0.16$	
Ln_FreqRep	11.01	1.24	8.95	0.94	9.73	0.8	$p = 0.0519$	
Familiarity	7.25	0.21	7.59	0.19	7.32	0.3	$p = 0.165$	
Imageability	6.04	1.09	8.21	0.2	7.11	0.53	$p < 0.01$	Concrete > Abstract ($p = 0.04$)
								Concrete > Emotional ($p < 0.001$)
								Emotional vs. Abstract ($p = 0.13$)
Concreteness	4.09	0.8	8.05	0.37	5.26	0.25	$p < 0.01$	Concrete > Abstract ($p < 0.001$)
								Concrete > Emotional ($p < 0.001$)
								Emotional > Abstract ($p = 0.02$)
Valence	7.16	1.36	6.93	0.55	5.28	3.2	$p = 0.4$	
Arousal	5.81	0.38	5.63	0.5	6.73	0.5	$p = 0.02$	Emotional > Concrete ($p = 0.04$)
								Emotional > Abstract ($p = 0.02$)
								Abstract vs. Concrete ($p = 0.85$)