Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5251–5260
Marseille, 20-25 June 2022

COSMOS: Experimental and Comparative Studies of Concept Representations in Schoolchildren

Jeanne Villaneau[1], Farida Saïd[2]
[1,2]IRISA-D6 - MEDIA ET INTERACTIONS,[2]LMBA
[1,2]Université de Bretagne Sud, France
{jeanne.villaneau, farida.said}@univ-ubs.fr

Abstract

COSMOS is a multidisciplinary research project investigating schoolchildren's beliefs and representations of specific concepts under control variables (age, gender, language spoken at home). Seven concepts are studied: *friend, father, mother, villain, work, television* and *dog*. We first present the protocol used and the data collected from a survey of 184 children in two age groups (6-7 and 9-11 years) in four schools in Brittany (France). A word-level lexical study shows that children's linguistic proficiency and lexical diversity increase with age, and we observe an interaction effect between gender and age on lexical diversity as measured with MLR (Measure of Lexical Richness). In contrast, none of the control variables affects lexical density. We also present the lemmas that schoolchildren most often associate with each concept. Generalized linear mixed-effects models reveal significant effects of age, gender, and home language on some concept-lemma associations and specific interactions between age and gender. Most of the identified effects are documented in the child development literature. To better understand the process of semantic construction in children, additional lexical analyses at the n-gram, chunk, and clause levels would be helpful. We briefly present ongoing and planned work in this direction. The COSMOS data will soon be made freely available to the scientific community.

Keywords: children survey, conceptual representations, lexical richness, lexical diversity, mixed-effects models

1. Introduction

COSMOS[1] ("Construction Sémantique du MOnde et Stéréotypes" *[Semantic Construction of the World and Stereotypes]*) is a French research project bringing together linguists and data scientists to study conceptual representations in schoolchildren aged 6 to 11. The project is supported by the MSHB (Maison des Sciences de l'Homme en Bretagne), a research unit that promotes interdisciplinary collaborative research at the heart of human and social sciences.

The project consists of conducting surveys in which children put certain concepts into words and, through their verbalizations, study their conceptual representations at two points in their schooling while considering sociodemographic and cultural factors. Beyond the interest of this knowledge per se, the outcomes have a didactic aim: to understand and evaluate the process of acquiring a semantic competence that goes beyond the referential use of words and, in the long run, to provide tools that will help prevent caricatural or fallacious presuppositions.

There seem to be universal stages in children's cognitive development, although these are subject to debate. *Concrete operational stage* occurs between ages 7 and 11: children become less egocentric, can solve complex problems, and acquire classification skills (Babakr et al., 2015). For example, by age 6-7, children can coordinate multiple social categories. By 10-12 years, development involves generalizations about concrete objects and events and the ability to abstract or think hypothetically (Fisher and Bullock, 1984). We are specifically interested in comparing children's conceptualizations between the beginning and end of this stage.

Operationally, the literature reports that surveys with children must be tailored to their cognitive and social maturity. They can be interviewed as young as 6 or 7 years old in qualitative studies, with extreme caution and using their own words. Questions and instructions should be simple, and the risk of suggestibility is high (Borgers et al., 2000; de Leeuw, 2011). Our data collection protocol was designed with these concerns in mind.

In section 2, we describe the protocol used during the surveys, and provide some statistical descriptives of the data collected. We also examine children's language proficiency through the length of their responses and the length of the words they used.

We present in section 3 a lexical analysis of the collected data at the word level. The first part (section 3.1) concerns the lexical richness of the children's verbalizations and its variability by age, gender, and home language. In the second part, we present, for each concept (section 3.2), its most associated lemmas and the effect of age, gender, and home language on the significant concept-lemma associations. We illustrate the results with quotes from children's responses and, where available, relate them to the child development literature.

In the last section (section 4), we present our ongoing work at the word level and future work at other levels of segmentation of the collected texts.

[1] `https://www.mshb.fr/projets_mshb/cosmos/6694/`

The COSMOS data will soon be freely available to the scientific community.

2. Data Collection

2.1. Protocol

We collaborated with educators to adapt the SPA protocol (Galatanu, 2018) for children. We chose seven concepts for practical reasons (interviews duration, human resources available for the surveys) and relied on the literature on cognition and child development to select them. They cover different semantic fields related to the child's social life, namely (1) family: *"un père"* (*a father*), *"une mère"* (*a mother*); (2) social interactions: *"un ami"* (*a friend*), *"un méchant"* (*a villain*); (3) entertainment: *"télévision"* (*television*); (4) values: *"travail"* (*work*); (5) natural world: *"un chien"* (*a dog*).

Two age groups of children were investigated: 6-7-year-olds (first grade) and 9-11-year-olds (fourth and fifth grades). In order to compare the two populations, we chose a single protocol adapted to both age groups. Eight interviewers participated in the survey: 5 were part of the COSMOS research team, and 3 were adult volunteers from the education community. The interviewers participated in preliminary interviews to standardize their practices.

Headteachers provided written consent for their school's involvement and, for each participating class, an information sheet and consent form were sent home to all children. We had written parental consent for all participating children, and children provided oral assent before each session. Meetings were conducted in the schools (library, activities room or other spare room) during school time.

Each child left the classroom for a face-to-face interview of ten to fifteen minutes with an interviewer, and a detailed transcript of the session was produced. The procedure was as follows: the interviewer explains the course of the session, clarifies that there are no right or wrong answers and that the children's responses will remain anonymous. The interviewer encourages the child to speak if needed, but is careful not to influence him/her. The interviewer writes down everything the child says, respecting the exact wording.

The children were asked two questions for each concept as illustrated below for the *father* concept:

1. What words come to your mind if I say *"a father"*?

2. What do you think *"a father"* is?

The concepts were presented in the same random order for all children: *friend, father, work, villain, television, mother, dog*. The order of presentation could affect children's oral production, and, in this sense, the design was not balanced for order effects. However, this issue is not addressed here, not least because testing a given order/gender/age group interaction on all possible orders, even limiting to semantic field orders, leaves very few subjects, if any, per cell in each age group.

2.2. Collected data

The surveys took place between March 2020 and March 2021 in 4 schools located in Brittany (France): school A in a peri-urban area, school D in a city-center area, and schools B and C in other urban areas.

The socio-demographic information collected from the children was age, gender, grade, and the languages spoken at home. 184 children were interviewed, 126 in the 9-11 age group and 58 in the 6-7 age group. The distributions of children by school (A, B, C, D), gender (F, M) and age are given in Table 1. There are significantly more boys in school A and significantly more girls in school B ($\chi^2(3) = 15.32$, $p < .001$, *Cramer's V*$= 0.29$), and we observe an over-representation of the 9-11-year olds in all schools because of the difference in the age ranges of the two groups.

School	F6-7	M6-7	F9-11	M9-11	Total
A	7	15	14	28	64
B	12	3	15	7	37
C	5	10	17	11	43
D	4	2	16	18	40
Total	28	30	62	64	184

Table 1: Number of children by age, gender and school

The first language spoken at home is, in decreasing frequency: French, Arabic, Turkish, Maore, Lingala, Bambara, Armenian, Albanian and Russian. We divided the children into two groups according to their first home language: 'French' (F1) and 'Other' (E1). Table 2 gives the distribution of home language by age, gender and school. There is an under-representation of French as the first language in school A and its over-representation in school B ($\chi^2(3) = 33.64$, $p < .001$, *Cramer's V*$= 0.43$).

Home language		French	Other
All		58.15	41.85
Age group	6-7	62.1	37.9
	9-11	56.4	43.6
Gender	F	55.3	44.7
	M	61.1	38.9
School	A	29.7	70.3
	B	78.4	21.6
	C	74.4	25.6
	D	67.5	32.5

Table 2: Distribution (in %) of the first language by age, gender and school

The children tended to answer the first question with lists of words or nominal groups and the second question with clauses. Following the specifications laid down by (Berman and Slobin, 1994) for spoken language, we define a clause as a unified predicate describing a single situation (an activity, event, or state).

1. What words come to your mind if I say *"a father"*?
   un papa, un papy [a dad, a grandpa]
2. What do you think *"a father"* is? *quelqu'un qui s'occupe de nous; qui était très content de nous avoir [someone who cares about us; who was very happy to have us]*

(boy, 6 years old)

The children's responses were transcribed using the exact wording provided by the interviewers, augmented with commas and semicolons. Semicolons always indicate a transition to another clause, while commas are used to ease reading and allow for enumeration, especially in the first question. Dividing the text into clauses remains a challenging issue.

During the interviews, many children did not distinguish between the two questions. We then chose to aggregate their responses to the two items; this produced 184 texts, one per child, and 46 600 words in total.

We first examined children's language proficiency through text and word length, which are known to be good indicators of this specific skill, particularly in lexical resources (Ruth and Bracha, 2010; Kang and Yan, 2018). We used two-way Anovas to investigate the effects of gender, age, and home language on children's language proficiency. We found out a significant main effect for age on both text length ($F(1, 176) = 17.1$, $p < .001$, $\omega^2 = .08$) and word length ($F(1, 176) = 4.79$, $p = .03$, $\omega^2 = .02$). Tukey's post hoc corrections showed that language proficiency is significantly higher in the 9-11 group than in the 6-7 group (text length: t =3.71, $p < .001$, and word length: $t = 2.19$, $p = .03$). Statistical descriptives of the two indicators are given in tables 3 and 4 respectively, for all children and by age category.

Age group	mean	std	min	max	med
all	253.2	148.3	18	989	231
6-7	193.9	108	18	435	186.5
9-11	280.5	156.5	65	989	251.5

Table 3: Number of words per child.

Age group	mean	std	min	max	med
all	4.16	0.32	3.51	5.30	4.10
6-7	4.10	0.34	3.51	5.24	4.06
9-11	4.18	0.30	3.61	5.30	4.11

Table 4: Length of words per child.

The following section focuses on the effects of gender, age, and home language on children's lexical richness in terms of lexical diversity and density and on the most common concept-lemma associations.

3. Data analysis

We used various measures to assess the lexical richness of the corpus and constructed frequency distributions for the most significant "concept-lemma" pairs. We used these data as dependent variables in mixed-effects models (Baayen et al., 2008) with gender, age, and home language as predictors and school as a context variable whose effect was treated with random intercepts. Fixed and random effects were incrementally added to a minimal model, and we compared the models using the likelihood ratio test. We describe in the following the best-fitting model for each analysis. Visual inspection of the residual plots revealed no obvious deviations from homoscedasticity or normality in all presented cases. Computations were performed with the function *lmer* from the package *lme4* (Bates et al., 2012) in the statistical environment R (R Core Team, 2021).

For ease of reading, we refer to 6-7-year-olds as 6-7 and 9-11-year-olds as 9-11 in the following.

3.1. Linguistic features

Lexical richness (LR) is a multidimensional feature of written and spoken language which refers to lexical sophistication and language proficiency. Many metrics have been proposed in the literature, the effectiveness of which is controversial (Van Hout and Vermeer, 2010; Zhang and Wu, 2021). Two characteristics are currently employed in writing and speaking to describe lexical development and language acquisition: lexical diversity and lexical density (Johansson, 2008).

Lexical density estimates linguistic complexity from the proportion of lexical words (i.e. nouns, verbs, adjectives and some adverbs) used. The most common measure of lexical density is the ratio of nouns, verbs, and adjectives (and often adverbs) to the total number of words. It is indicative of lexical density since content words contain more information than function words. Other options exist: noun ratio, ratio of lexical words to utterances, ratio of pronouns to words since pronouns refer to objects (Johansson, 2008), Etc.

Lexical diversity measures the number of different words used in a text: the more varied the vocabulary, the higher the lexical diversity. The traditional measure of lexical diversity is the TTR (Type-Token Ratio), the ratio of different words to the total number of words. The TTR is sensitive to the length of the text: the rate of different words increases less as the text gets longer due to lexical repetition. More sophisticated measures have been proposed to cope with this problem, as the index of Guiraud, VoCD, and MTLD (Measure of Textual Lexical Diversity). The index of Guiraud is the square root of TTR. VoCD (Malvern et al., 2004) is obtained from a series of random text samplings; it is suitable for texts of 50 words and more. MTLD is based on a moving window approach: it is the mean length of the sequential word strings in a text that maintains a given TTR value (default TTR=0.732)

(McCarthy and Jarvis, 2010). In MTLDbi, the MTLD calculation is performed twice: once in left-to-right text order and once in right-to-left text order. There is no agreement on the best measure of lexical diversity, but the Guiraud index, VoCD, and MTLD are cited as the most reliable (Jarvis, 2013; McCarthy and Jarvis, 2010; Johansson, 2009). The MTLD measure is highly appropriate for short texts (Koizumi, 2012), which makes it relevant for our data. A. Vermeer (2004) argues that lexical density and diversity measures are not related to word difficulty. She proposes a measure of lexical richness based on classes of lemmas (voclists), according to their frequencies in a global corpus (Vermeer, 2004). Specifically, lemmas are divided into nine classes, and the relative coverage of the corpus by the classes is taken as *model*.

The MLR metric estimates a text's vocabulary size as a linear combination of parameters $q_i, i = 1 \cdots 9$, where q_i is the ratio between the text and the *model* coverages of class i. The minimum MLR value is 1 if all the words in the text belong to the first class of the thousand most frequent lemmas. A comparison between classical measures of lexical diversity showed that MLR succeeds in discriminating between two groups of students while measures such as TTR and VoCD fail to (Van Hout and Vermeer, 2010).

We built the global corpus used for MLR computation with Manulex, a French lexical resource related to schoolchildren (Lété et al., 2004). Manulex provides word occurrence frequencies in a corpus of 54 textbooks (1.9 million words). However, it does not provide the frequencies of grammatical words, so we have completed it with the frequencies of the 1500 most common words of the French language given by Eduscol, an official website of the French national education system[2].

We then divided the lemmas into 9 classes to construct a 'model' coverage, as in Vermeer (2004) and computed MLR scores as in Van Hout and Vermeer (2010). We excluded proper nouns from the calculation because most of them refer to names of friends mentioned by the children and are not relevant for our study.

We ran GLMMs with MTLD, MTLDbi, and MLR scores as dependent variables. The best-fitting models are presented in Table 5; they show a significant effect of age on lexical diversity, whatever the metric, and an interaction between age and gender on MLR scores. Specifically, Table 6 highlights that lexical diversity is significantly higher for older children, and when measured with the MLR metric, lexical diversity is significantly lower for non-French native boys compared to girls of the same background and French native boys. On the other hand, age, gender, and home language have no significant effect on lexical density.

MTLD				
Fixed effects	mean (β)	SE	t	p
Intercept	35.77	.99	36.01	***
Age	3.31	.82	4.04	***
Random effect	Variance		SD	
School intercept	1.24		1.11	
MTLDbi				
Fixed effects	mean (β)	SE	t	p
Intercept	32.01	1.09	31.27	***
Age	4.25	.74	5.71	***
Random effect	Variance		SD	
School intercept	1.96		1.40	
MLR				
Fixed effects	mean (β)	SE	t	p
Intercept	2.56	.10	26.45	***
Age	.45	$< .01$	4.81	***
Gender	$< .01$	$< .01$	5.46	ns
Language (HL)	$< .01$	$< .01$	-0.52	ns
Gender*HL	.26	$< .01$	2.96	**
Random effect	Variance		SD	
School intercept	$< .01$		$< .01$	

Note. SD = standard deviation; SE = standard error. Number of observations = 184. p-values: ***$p < .001$, **$p < .01$, ns: not significant.

Table 5: Summary of Linear Mixed-Effects Models with MTLD, MTLDbi and MLR as outcome variables, with random intercepts for each school, and with age, gender and home language as predictors

3.2. Concept-Lemma associations

This section focuses on the lemmas most often associated with the concepts under study and investigates whether sociocultural variables affect these associations. We sorted the lemmas by concept based on the number of children who used them at least once. We then used generalized binomial mixed-effects models to explore the effects of age, gender, and home language on selected concept-lemma associations. We considered age, gender, and home language as predictors, while school served as a contextual variable. Table 7 shows the results of the binomial GLMMs for the concept "mother." Age affects significantly the "mother-mom" and "mother-to help" associations. Table 8 provides the list of concept-lemma associations for which significant effects of our predictors were found. These results are detailed, concept by concept, in the following.

3.2.1. Friend *["ami"]*

Table 9 gives the lemmas most associated with the concept *friend* and the number of children who mentioned them.

Playing with peers is the primary context in which young children form friendships (Coelho et al., 2017) and indeed, we observe that *"jouer" [to play]* is the lexical word most associated with the concept *friend.* This association is more significant in 6-7-year-olds than in 9-11-year-olds (72.4% vs. 54%).

[2] https://eduscol.education.fr/186/liste-de-frequence-lexicale.

MTLD			
Age	Estimate	SE	95% CI
6-7	32.46	1.47	[29.58;35.34]
9-11	39.08	1.08	[36.97;41.19]
MTLDbi			
Age	Estimate	SE	95% CI
6-7	27.76	1.42	[24.98;38.55]
9-11	36.26	1.09	[34.12;38.39]
MLR			
Age	Estimate	SE	95% CI
6-7	2.11	0.16	[1.80;2.42]
9-11	3.01	0.11	[2.80;3.22]
Gender*Language	Estimate	SE	95% CI
F, Other	2.73	0.23	[2.27;3.18]
M, Other	2.20	0.19	[1.83;2.57]
F, French	2.39	0.16	[2.07;2.71]
M, French	2.87	0.18	[2.52;3.22]

Note. SE = standard error. Number of observations = 184. CI = confidence interval

Table 6: Estimated marginal means in Linear Mixed-Effects Models with MTLD, MTLDbi and MLR as outcome variables, with random intercepts for each school, and with age, gender and home language as predictors

Mom				
Fixed effects	mean (β)	SE	t	p
Intercept	-.08	.17	-.49	ns
Age	-.65	.17	-3.89	***
Random effect	Variance		SD	
School intercept	$<.01$		$<.01$	
To help				
Fixed effects	mean (β)	SE	t	p
Intercept	-1.54	.24	-6.49	***
Age	.62	.24	2.62	$<.01$
Random effect	Variance		SD	
School intercept	1		1	

Note. SD = standard deviation; SE = standard error. Number of observations = 184. p-values: ***$p<.001$, **$p<.01$, ns: not significant.

Table 7: Summary of binomial GLMMs for the concept "Mother" with random intercepts for each school, and with age, gender and home language as predictors

> *On est ami, c'est un copain, on joue ensemble [we're friends; he's a friend; we play together]* (girl, 6 years old).

The word *"copain" [buddy]*, a colloquial synonym for *friend*, is more significantly associated with *"ami"* among boys than girls (72.4% vs. 54%). However, girls often use the word *"copine"*, a feminine form of *"copain"*, suggesting, consistent with the literature, a large majority of same-gender friendships in these age groups (Zaidman, 2007). Older children also associate more significantly the concept of *friend* with the lexical words *"aider" [to help]* (27% in 9-11 vs 5.2% in 6-7) and *"confiance" [trust]* (19.1% in 9-11 vs. 1.7% in 6-7).

Fixed Effects		
Concepts	Significant Effects	Lemmas
Friend	Age Gender Gender*Language	*to play, to help, trust* *buddy* *trust*
Father	Age Language	*dad, family, to like* *to take, care*
Mother	Age	*mom, to help*
Television	Gender*Language	*game*
Work	Age	*to work, school, money*
Dog	Age Language	*to play, friend* *to like*

Table 8: Significant fixed effects for concept-lemma associations

Lemma	*jouer* *[to play]*	*copain* *[buddy]*	*aimer* *[to like]*
Children	110	51	44

Lemma	*aider* *[to help]*	*confiance* *[trust]*
Children	37	25

Table 9: Selected lemmas and number of children who associated them with the concept *friend*.

> *"C'est quelqu'un que tu joues avec, que tu peux lui faire confiance" [it's someone you play with, that you can trust]* (boy, 9 years old).

9-11-year-olds often mentioned sharing secrets as a sign of trust and friendship, consistent with the literature (Liberman and Shaw, 2018):

> *"se dire des secrets; avoir confiance" [telling each other secrets; trust]* (boy, 10 years old).

A developmental psychology approach draws on the observed gender separation between boys' and girls' groups (Zaidman, 2007), to suggest the development of two different peer cultures and socialization models (Maccoby, 1998; Underwood, 2007). In this regard, we observe that the lemma *to help* is more significantly associated with *friend* in boys than in girls (26.6% vs. 13.3%), which is consistent with the socialization attributed to boys: for example, ball games (such as *football*, more often mentioned by boys) value help and solidarity. In the model attributed to girls, friendship is more intense and intimate. This pattern does not reflect in our lexical study; however, the lemma *confiance* is more significantly associated with *friend* among girls whose home language is not French compared to boys in this category (22.9% vs. 4.8%).

3.2.2. Father *["père"]*

The childish word *"papa" [dad]* is the lemma most frequently associated with *father* and is significantly more used by younger children (56.9% in 6-7 vs. 25.4% in 9-11). In contrast, older children significantly associate

Lemma	*papa* *[dad]*	*aimer* *[to like]*	*aider* *[to help]*
Children	65	57	41
Lemma	*famille* *[family]*	*s'occuper* *[to take care]*	
Children	36	32	

Table 10: *Father*: number of children using most frequent lemmas.

the lemmas *family* (25.4% in 9-11 vs. 6.9% in 6-7) and *to help* (31.8% in 9-11 vs. 1.7% in 6-7) with the concept of father. On the other hand,*"s'occuper de" [to take care]* is significantly more used by children whose home language is French (23.4% vs 9.1%): while the verb *"occuper"* is not an unusual word, it is not very commonly used as a pronominal verb with the meaning of *to take care*.

> *"famille; s'occuper des enfants ; activités avec les enfants ; travailler ; s'acheter des voitures ; être gentil ; faire à manger ; lire des histoires" [family; to care for children; activities with children; to work; to buy oneself cars; to be nice; to cook; to read stories]* (girl, 9 years old, French home language)
> *"mon papa ; quelqu'un qui pense toujours à moi" [my dad; someone who always thinks of me]* (girl, 6 years old, French home language)

3.2.3. Mother *["mère"]*

Lemma	*maman* *[mom]*	*aimer* *[to like]*	*gentil* *[nice]*	*manger* *[to eat]*
Children	78	48	45	44
Lemma	*aider* *[to help]*	*s'occuper* *[to take care]*		*enfant* *[child]*
Children	42	33		30

Table 11: *Mother*: number of children using the most frequent lemmas.

The two most common lemmas are similar for *father* and *mother*, with *"maman" [mom]* instead of *"papa" [dad]*.

Once again, the childish word *mom* was used significantly more often by 6-7-year-olds than by 9-11-year-olds (63.8% vs. 32.5%). In contrast, older children used the lemma *"aider" [to help]* more significantly (28.6% in 9-11 vs. 6% in 6-7).

The following two lemmas, *"gentil" [nice]* and *"manger" [to eat]*, do not appear in the lemmas frequently associated with the concept of *father.*

> *"présente pour t'aider quand ça ne va pas, t'éduque, te nourrit, fait à manger, qui m'aide à faire mes devoirs, qui fait le linge" [present to help you when you are not well, educates you, feeds you, cooks, helps me with my homework, does the laundry]* (girl, 11 years old)
> *"elle prépare à manger ; elle met le linge à sécher ; elle lit une histoire aux enfants" [she cooks food; she puts the laundry to dry; she reads a story to the children]* (girl, 6 years old)

Craig (2006) states that "caregiving is a complicated mixture of work and love" . Her survey compares male and female care and indicates substantial differences between maternal and paternal caregiving, despite the progression of shared parenting. More recent surveys suggest that mothers often have a more important role than fathers in providing emotional support and organizing the daily lives of their children (Han and Jun, 2013). Further analysis is needed to determine whether our data demonstrate a different perception of father and mother roles among children.

3.2.4. Villain *["(un) méchant"]*

Lemma	*gentil* *[nice]*	*taper* *[to hit]*	*aimer* *[to like]*
Children	64	40	34
Lemma	*voler* *[to steal]*	*embêter* *[to bother]*	
Children	31	23	

Table 12: *Villain*: number of children using the most frequent lemmas.

A *villain* was often defined as not being nice, making *"gentil" [nice]* the lemma most commonly associated with the concept *villain*. Lemma *"aimer," [to like (love)]* was significantly more used by 9-11-year-olds (16.3% in 9-11 vs. 2.2% in 6-7) in different meanings: *"he doesn't like," "we don't like him"* or *"he enjoys [in French, 'il aime'] bothering people"*. Thus, children, especially in the 9-11-age group, often defined *villain* as a negation of *friend*, consistent with the concurrent development of emotional and social skills (Denham et al., 2002).

> *"quelqu'un qui est pas gentil et qui aime pas les autres" [someone who's not nice and doesn't like other people]* (boy, 9 years old)

The children used many lemmas to express the misdeeds of the *villain*. Besides *to hit*, *to steal*, *to bother*, the *villain* may *insult ["insulter"]* (21 children), *kill ["tuer"]* (18 children), or *harass ["harceler"]* (15 children). There was a significant effect of age on the verb *to steal* which was more used by 6-7-year-olds (29.3% in 6-7 vs. 11% in 9-11).

The definition of the concept *villain* given by the children through the lemmas seems more inspired by current events and real-life (Kingery et al., 1998; Hayden and Dlugosz, 2012) than by stories and literature (Spanothymiou et al., 2015).

3.2.5. Television *["télévision"]*

Lemma	*dessin animé* *[cartoon]*	*film* *[movie]*	*écran* *[screen]*
Children	86	84	52
Lemma	*série* *[serie]*	*jeu* *[game]*	*jouer* *[to play]*
Children	36	28	24

Table 13: Television: number of children using the most frequent lemmas.

Table 13 highlights that children associate *television* mainly with cartoons and movies. Nevertheless, the frequencies of *screen*, *game*, and *to play* show the importance of television as a game console or big screen. *Game* is significantly more common among boys whose home language is French compared to girls in this category (21.2% of boys vs. 3.6% of girls).
In addition, many children reported enjoying television and spending *time* (19 children) or having *fun* (20 children) watching television alone or with family or friends.

> *"profiter, prendre du bon temps et du bonheur avec sa famille" [enjoy, have a good time and be happy with your family]* (girl, 10 years old)
> *"la télévision ça sert à nous amuser en regardant des films rigolos" [TV is for fun watching funny movies]* (boy, 10 years old)

Some children expressed reservations:

> *"il ne faut regarder que parfois sinon on a les yeux qui fait mal"* [you have to look only sometimes otherwise your eyes will hurt] (boy, 6 years old)
> *"pas très bien mais franchement pour moi c'est un peu bien" [not very good but honestly for me it's a bit good] (boy, 10 years old)*

However, no children referred to issues associated with excessive screen time, such as behavioral problems (less sleep, violent behavior, Etc.) and health problems (obesity, low physical activity, Etc.), which are well documented (Paulich et al., 2021; Atabey, 2017; Ejaz et al., 2020; Villani, 2001).

3.2.6. Work *["travail"]*

"travailler" [to work] is by far the lemma most associated with the concept *"travail" [work].* It is significantly more common among 6-7-year-olds (76% in 6-7 vs. 54% in 9-11), for whom *travail [work]* is primarily defined by *travailler [to work].*

Lemma	*travailler* *[to work]*	*apprendre* *[to learn]*	*école* *[school]*
Children	112	79	50
Lemma	*argent* *[money]*	*devoirs* *[homework]*	
Children	49	29	

Table 14: *Work*: number of children using the most frequent lemmas.

> *"on est au bureau on travaille" [we're in the office we work]* (boy, 6 years old)
> *"bien écouter la maîtresse, bien travailler, bien lire et écrire" [listen to the teacher, work well, read and write well]* (boy, 6 years old)
> *c'est là où on travaille dur [it's where we work hard]* (girl, 6 years old)

Three of the most common lemmas (*learn, school, homework*) refer directly to schoolwork: for most children, *work* refers to their own work. The association between *work* and *school* is significantly more important among 9-11-year-olds (34.9% in 9-11 vs. 10.3% in 6-7).
The lemma *money* is also significantly more used among 9-11-year-olds (34.1% in 9-11 vs. 10.3% in 6-7); it never refers to an allowance given by parents, but to the possibility of having a job to earn a living.

> *"pour gagner de l'argent ; pas non plus pour être riche mais pour vivre bien" [to earn money; not to be rich but to live well]* (girl, 10 years old)
> *"c'est recevoir de l'argent" [it's receiving money]* (girl, 10 years old)
> *"les gens travaillent pour gagner de l'argent pour prendre soin de leurs familles" [people work to earn money to take care of their families]* (girl, 10 years old)
> *"tu travailles tu trouves des amis ; l'école c'est plus pour apprendre ; tu travailles pour gagner de l'argent ; être heureux." [you work you find friends ; school is more for learning ; you work to earn money; be happy]* (boy, 10 years old)

In a survey of 13- to 14-year-olds and their parents, Dixon et al. (2014) identify four categories related to beliefs about the goals of schooling: learning and gaining self-awareness; developing social and life skills; optimizing life chances and quality of life; and enabling future employment and economic well-being. The 9-11-year-old pupils in our survey generally express the same goals.

3.2.7. Dog *["chien"]*

Family pets are reported in the literature to provide complementary friendship (Davis and Juhasz, 1995;

Lemma	*jouer* *[to play]*	*aimer* *[to like]*	*manger* *[to eat]*
Children	54	44	31
Lemma	*gentil* *[nice]*	*promener* *[to walk]*	*ami* *[friend]*
Children	30	24	23

Table 15: *Dog*: number of children using the most frequent lemmas.

Hawkins et al., 2017), so it is not surprising to find in the concept *dog* some of the lemmas most frequently associated with the concept *friend*, namely *to play, to like, friend*:

> *"il te tient compagnie ; tu vas jouer avec lui" [he keeps you company; you will play him]* (girl, 10 years old)]

The lemmas *to eat* and *to walk* emphasize the children's role as caregivers in the relationship with a pet (Muldoon et al., 2015).

> *"c'est un animal de compagnie que tu vas promener tous les jours même s'il pleut" [it's a pet you walk every day even if it rains]* (girl, 10 years old)

Older children associate more significantly the lemmas *to play* (37.3% in 9-11 vs. 12.1% in 6-7) and *friend* (17.5% in 9-11 vs. 1.7% in 6-7) with the *dog* concept. Age of pet attachment is controversial in the literature: the survey of Hawkins (2017) gives no significant difference in pet attachment between younger (6-9 years) and older (10-13 years) children, whereas Melson, Peet, and Sparks (1991) found that pet attachment was strongest among 9-10-year-olds. Nevertheless, we observed that some 6-7-year-olds expressed fear of dogs:

> *"peut faire du mal" [it can hurt]* (girl, 6 years old)
> *"mordre, méchant" [to bite, mean]* (boy, 6 years old)
> *"un animal qui mord les gens" [an animal that bites people]* (boy, 6 years old)

The lemma *to like* as an expression of affection to the dog is significantly more frequent in French native girls compared to non-French native girls (79.2% vs. 20.8%).

4. Conclusion

The COSMOS data are transcripts of oral responses from 184 schoolchildren to open-ended questions about seven concepts: *friend, father, mother, villain, work, television* and *dog*. The collected data includes about 46 000 French words and will soon be freely available to the scientific community.

We presented the results of a word-level lexical study: assessing language proficiency and lexical richness in terms of lexical diversity and density and analyzing the most common associations between each concept and the collected lemmas. We used mixed-effects models to explore the effects of age, gender, and home language on lexical richness while considering variability across schools. We found that language proficiency and lexical diversity were significantly higher among older children. We also observed an interaction effect between home language and gender on lexical diversity as measured by MLR. The analysis of lemma-concept associations revealed a significant effect of age for all concepts except television, a gender effect on the concept *friend* (lemma *buddy*), a home language effect for the concept *father* (lemma *to care*), and a gender*language interaction for the concepts *friend* (lemma *to trust*), *television* (lemma *to play*), and *dog* (lemma *to like*). Many of the revealed effects are documented in the literature, with the exception of language and gender*language effects that may be related to the composition of the "other" home language group and deserve further study.

Research is currently in progress to determine children's representations of each concept by categorizing their associated lemmas: for example, the lemmas *food, cooking, eating, nourishing* associated with the concept *mother* could as well be grouped in the "nourishing function" category as in the "food" category.

Other research directions include semantic analysis at chunk and clause levels (as defined in 2.2): taking into account multiword expressions (MWEs) (Baldwin and Kim, 2010; Laporte, 2018) such as *"faire tout pour quelqu'un" [do everything for someone], "passer du temps ensemble" [spend time together]*, Etc.; disambiguation of certain words based on their context: for example, the meaning of *"aimer" [to like, to love]* used in *"mon papa c'est quelqu'un qui nous aime" [my dad is someone who loves us]* is not the same as in *"il aime dormir sur le canapé" [he likes sleeping on the couch]* ; taking negations into account; Etc.

Another direction of research is the analysis of the orientation of the concepts by studying the polarity of the sentiment (positive or negative) and the strength of the sentiment of the lemmas, clauses, and texts collected. A fine-grained analysis of sentiment at the clause level involves the analysis of words, phrases (n-grams), and texts. All levels will have to be considered for good modeling of compositional sentiment (children's collective sentiment).

The COSMOS data can be used for a variety of studies, and we hope it will be helpful to the language science research community.

5. Acknowledgements

This research was partly supported by the MSHB (Human Sciences Institute in Brittany).

6. Bibliographical References

Atabey, D. (2017). Cartoons: A profound outlook within the scope of children and media. *International Journal of Research in Education and Science*, 7(1):93–111.

Baayen, R., Davidson, D., and Bates, D. (2008). Mixed-effects modelling with crossed random effects for subjects and items. *Journal of Memory and Language*, 59(4):390–412.

Babakr, Z. H., Mohamedamin, P., and Kakamad, K. (2015). Piaget's cognitive developmental theory: Critical review. *Education Quarterly Reviews, Vol.2, No.3, 517-524*, 2(3):517–524.

Baldwin, T. and Kim, S. N., (2010). *Multiword Expressions*, pages 267–292. Nitin Indurkhya and Fred J. Damerau (eds.), CRC Press, Boca Raton, USA.

Bates, D., Mächler, M., and Bolker, B. (2012). *lme4: Linear Mixed-Effects Models Using S4 Classes (R Package Version 0.999999-0).* http://cran.r-project.org/web/packages/lme4/index.html.

Berman, R. and Slobin, D. I., (1994). *Relating events in narrative: A cross-linguistic developmental study*, pages 660–663. Hillsdale, NJ: Erlbaum.

Borgers, N., de Leeuw, E., and Ho, J. (2000). Children as respondents in survey research: Cognitive development and response quality. *Bulletin de Méthodologie Sociologique*, 66:60–75.

Coelho, L., Torres, N., Fernandes, C., and Santos, A. J. (2017). Quality of play, social acceptance and reciprocal friendship in preschool children. *European Early Childhood Education Research Journal*.

Craig, L. (2006). Does father care mean fathers share? a comparison of how mothers and fathers in intact families spend time with children. *Gender & Society*, 20(2):259–281.

Davis, J. and Juhasz, A. M. (1995). The preadolescent pet friendship bond. *Anthrozoös*, 8:78–82.

de Leeuw, E. D. (2011). Improving data quality when surveying children and adolescents: Cognitive and social development and its role in questionnaire construction and pretesting. Annual Meeting of the Academy of Finland, Naantali Finland.

Denham, S. A., von Salisch, M., Olthof, T., and Caverly, S. (2002). Emotional and social development in childhood. In *Blackwell Handbook of Childhood Social Development*. Peter K. Smith and Craig H. Hart.

Dixon, R., Peterson, E., Rubie-Davies, C., and Irving, S. (2014). Why go to school? student, parent and teacher beliefs about the purposes of schooling. *Asia Pacific Journal of Education*, 1(14).

Ejaz, Z., Rafiq, N., Zubair, R., and Saleem, H. (2020). Preadolescent perception of television violence. *European Journal of Special Education Research*, 6(1).

Fisher, K. W. and Bullock, D., (1984). *Cognitive Development In School-Age Children: Conclusions And New Directions*, chapter 3. National Academy Press (US).

Galatanu, O., (2018). *Analyse sémantico-discursive et sémantique expérimentale : méthodologies alternatives et croisées*, chapter 5, pages 261–309. P.I.E, Peter Lang.

Han, Y. S. and Jun, W. P. (2013). Parental involvement in child's development: Father vs. mother. *Open Journal of Medical Psychology*, 2:1–6.

Hawkins, R. D., Williams, J. M., and for the Prevention of Cruelty to Animals (Scottish SPCA), S. S. (2017). Childhood attachment to pets: Associations between pet attachment, attitudes to animals, compassion, and humane behaviour. *Int. J. Environmental Research and Public Health*, 14(5).

Hayden, C. and Dlugosz, G. (2012). Secondary school children and the experience of robbery: A survey in three south london schools. *Crime Prevention and Community Safety*, 14(2):122–139.

Jarvis, S. (2013). Capturing the diversity in lexical diversity. *Language Learning*, 63(s1):87–106.

Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: a developmental perspective. Technical Report 53, Lund University, Dept. of Linguistics and Phonetics. 3rd. edn.

Johansson, V. (2009). *Developmental Aspects of Text Production in Writing and Speech*. Ph.D. thesis, Lund University.

Kang, O. and Yan, X. (2018). Linguistic features distinguishing examinees' speaking performances at different proficiency levels. *Journal of Language Testing & Assessment*, 1:24–39.

Kingery, P. M., Coggeshall, M. B., and Alford, A. A. (1998). Violence at school: Recent evidence from fournational surveys. *Psychology in the Schools*, 35(3):247–258.

Koizumi, R. (2012). Relationships between text length and lexical diversity measures: Can we use short texts of less than 100 tokens ? *Vocabulary Learning and Instruction*, 1(1):60–69.

Laporte, E., (2018). *Choosing features for classifying multiword expressions.*, pages 143–186. Manfred Sailer; Stella Markan.

Liberman, Z. and Shaw, A. (2018). Secret to friendship: Children make inferences about friendship based on secret sharing. *Developmental Psychology*, 54(11):2139–2151.

Lété, B., Sprenger-Charolles, L., and Colé, P. (2004). Manulex: A grade-level lexical database from french elementary-school readers. *Behavior Research Methods, Instruments & Computers*, 36:156–166.

Maccoby, E. (1998). *The two sexes: Growing up apart, coming together.* Harvard University Press, Cambridge.

Malvern, D., Richards, B., Chipere, N., and Duran, P. (2004). *Lexical Diversity and Language Devel-*

opment. Quantification and Assessment. Palgrave Macmillan, New York.

McCarthy, P. and Jarvis, S. (2010). Mtld, vocd-d, and hd-d: a validation study of sophisticated approaches to lexical diversity assessment. *Behav Res Methods.*, 42(2):381–392.

Melson, G., Peet, S., and Sparks, C. (1991). Children's attachment to their pets: links to socio-emotional development. *Children's Environments Quarterly*, 8(2):55–65.

Muldoon, J. C., Williams, J. M., and Lawrence, A. (2015). 'mum cleaned it and i just played with it': Children's perceptions of their roles and responsibilities in the care of family pets. *Childhood*, 22(2):201–216.

Paulich, K. N., Ross, J. M., Lessem, J. M., and Hewitt, J. K. (2021). Screen time and early adolescent mental health, academic, and social outcomes in 9-and 10-year old children: Utilizing the adolescent brain cognitive development (abcd) study. *PLOS ONE*, 16(9).

R Core Team, (2021). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Ruth, A. B. and Bracha, N. (2010). The lexicon in writing–speech-differentiation developmental perspectives. *Written Language & Literacy*, 13(2):183—-205.

Spanothymiou, P., Kyridis, A., Christodoulou, A., and Kanatsouli, M. (2015). Children's evaluations of villains in children's the heroes and literature. *Studies in Media and Communication*, 3(1).

Underwood, M. (2007). Introduction to the special issue: Gender and children's friendships: Do girls' and boys' friendships constitute different peer cultures, and what are the trade-offs for development? *Merrill-Palmer Quarterly*, 53(3):319—-324.

Van Hout, R. and Vermeer, A. (2010). Comparing measures of lexical richness. In *Modelling and assessing vocabulary knowledge*. H. Daller, J. Milton J. Treffers-Daller (eds.), Cambridge University Press.

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in dutch l1 and l2 children. In *Vocabulary in a Second Language: Selection, acquisition, and testing*. John Benjamins: : Paul Bogaards and Batia Laufer.

Villani, S. (2001). Impact of media on children and adolescents: a 10-year review of the research. *Journal of the American Academy of Child and Adolescent Psychiatry*, 40(4):392–401.

Zaidman, C. (2007). Jeux de filles, jeux de garçons. *les cahiers du CEDREF*, 15:283–292.

Zhang, Y. and Wu, W. (2021). How effective are lexical richness measures for differentiations of vocabulary proficiency? a comprehensive examination with clustering analysis. *Lang Test Asia*, 11(15).