GENDER DIFFERENCES IN THE USE OF NOUN CONCEPT CATEGORIES – A STATISTICAL STUDY BASED ON DATA FROM CHILD LANGUAGE ACQUISITION

Velina Slavova, Dimitar Atanasov, Filip Andonov

***Abstract**: We have analyzed data derived from large corpora of child language acquisition in the attempt to build a model of primary semantic categories of nouns. The statistical result found for noun frequencies, compared with results from brain imaging studies in adults suggests the possible existence of mental representations in early child development that shape the structure of the semantic space. We propose a set of primary categories of noun-concepts and investigate the influence of age and gender on the intensity of use of nouns from these categories. The statistical analysis of the progressive use of nouns from the proposed categories shows a coincidence with a wide range of theories of gender differences. From our data, it is not possible to distinguish between learned gender aptitudes and innate preferences. However, our statistical result supports the idea that gender distinctions may date to early human history.*

***Keywords:** cognitive modeling, mental representations, concept categories, gender differences*

***ACM Classification Keywords**: 1.2.7. Natural language processing, 1.2.0. Cognitive simulation, G.3 Correlation and regression analysis, H.2.8. Database applications.*

Introduction

This paper is based on the idea that the examination of child language acquisition gives scholars the opportunity to investigate the semantics and principles of human language.

One of the earliest scientific explanations of language acquisition was provided by [Skinner, 1957] accounting for language development by means of environmental influence. The behaviorist idea that infants imitate their environment reverses the view of Chomsky claiming that if the language acquisition mechanism was dependent on language input alone, children will never acquire to process an infinite number of sentences. The theory of Universal Grammar states the existence of innate, biological grammatical categories, such as a noun category and a verb category. Later some psycho-linguists argued that innate grammatical categories are biologically, evolutionarily and psychologically implausible. Researchers started to suggest that instead of having a language-specific mechanism for language processing, children utilize general cognitive and learning principles, returning to Piaget's view

that learning language is mainly focused around cognitive development (e.g. [Piaget, 1955]). For example, many works of Michael Tomasello claim the “language instinct” cannot explain how children learn language, but their linguistic ability is interwoven with other cognitive abilities (see for ex. [Tomasello M., 2003]).

Other studies concentrate on the social interaction and demonstrated its importance for the acquisition of language and for the overall cognitive development. For example, the analysis of vast observational data of 1- through 3-year-olds learning to talk during their everyday interactions with their parents, showed how crucial to development is the amount of children's language experience as partners in the social dances of conversation ([Hart & Risley, 1995], [Hart, 2000]).

Specialized studies suggest gender differences in first language acquisition. Fundamental works on verbal ability have shown that girls and women surpass boys and men in verbal fluency, correct language usage, sentence complexity, grammatical structure, spelling, and articulation [Eckert & McConnell-Ginet, 2003]. These general suggestions are regularly confirmed in statistical studies (e.g. [Slik, Hout & Schepens, 2015]). Although these observations are commonly accepted by the scientific community, the reasons for the existence of such differences remain unclear.

Several studies have been conducted in order to explain these gender related differences. For example, fMRI studies have suggested that girls rely on a supramodal language network, whereas boys process visual and auditory words differently [Burman et al., 2008]. Other researchers propose that gender differences are genetically determined. Researchers have found that the amount of FOXP2 protein (a gene related to language) in the brains of 4 to 5 years old boys and girls (in Brodmann area 44) is 30% more in girls’ brains as compared to boys’ [Balter, 2013].

Differences in brain areas related to language faculty were found in studies on gender differences (see [Northwestern University, 2008]). Analyses of these results have led researchers to the assumption that the observed distinctions may date to early human history.

Concerning the importance of social interaction mentioned above, a recent study of 268 children aged between 18 and 35 months has shown that the association between expressive language and social ability is significantly stronger in boys than in girls [Longobardi et al, 2016].

However, finding definitive answers to the questions of language acquisition and gender differences is an unresolved problem.

Our research does not provide an explanation for language-related gender differences. We present the reasoning which has led to the establishment of a set of noun-concept categories, proposed after analysis of brain imaging studies, models of cognition and our data. We performed a statistical analysis of the use of these noun concept categories and found statistically significant gender differences. The

results we have obtained from the statistical analysis of the use of noun-concept categories by children can help the deeper understanding of the problem.

Basic Assumptions and Data Description

The assumptions in this paper follow the model of language faculty proposed in [Slavova & Soschen 2015]. The analysis provided there of contemporary findings in brain imaging, neuroscience, cognitive science, psychology, linguistics etc. lead to the conclusion that the building of a mental representation of the world starts on the basis of genetically determined information treatment processes. Following the proposed *Self-centered model of language faculty*, the primary mental representations are obtained using neuron networks present and functioning at birth - multimodal perception, proprioception, interoception, mirror neuron system and default mode network. The central suggestion derived from the analysis of the specialized sources was that a new-born has a biologically underwritten notion about the existence of his Self as actor in the environment (in compliance with [Barsalou, 2003]). It was suggested that semantic categories are further shaped from the point of view of personal, situational experience. In big, the proposed there model focuses on the creation of Meaning as internal mental representation, which establishment relies initially on inborn mechanisms, on the role of an actor in the environment and on the innate "knowledge" about the existence the Self and the self-similar.

The statistical study presented here concerns the questions of Meaning. We assume that the development of the semantic representation of the world starts on the basis of primary conceptualization mechanisms realized by biologically underwritten processing of "automatic" classification of the information into semantic categories, necessary to guide actions of importance for the individual's survival. We suppose that the initial concepts arise as internal information units which creation relies on inborn brain processing that organizes the information flow coming from the environment and the flow from the "inside" of the biological system, insuring its functioning.

Our study is based on the assumption that language faculty is a *result* of the capacity for conceptualization. We concentrate on the analysis of nouns as the use of nouns starts first in language production, apparently being the most natural vehicle of meaning, and strongly dominates child speech during the period of language acquisition considered in our data.

Data from 30 corpora containing dialogues of child speech in English, annotated with parts of speech and grammar, were extracted from CHILDES [MacWhinney, B. & C. Snow (1985)], [MacWhinney, 2000]) and stored and organized in a relational database [Slavova, 2016]. The obtained data collection contains 125,584 speech utterances of different children aged between 9 and 62 months, produced in free dialogues and collected by researchers in child language acquisition during several decades. Our study is based on the linguistic annotation for parts of speech (POS) taken from CHILDES which

respects the developed MORPH system [Hausser, R. (1989)]. Our previous statistical results based on the ratios of use of POS [Atanasov et al., 2016] confirmed the "special" role of nouns in language acquisition.

Proposed Model of Noun Concept Categories

We try to elaborate a set of noun-concept categories which are "primary" as they emerge using inborn mechanisms for semantic assessment of the perceived Entities. For the purpose, we have looked into the results coming from brain studies in adults, searching for reported "particular features".

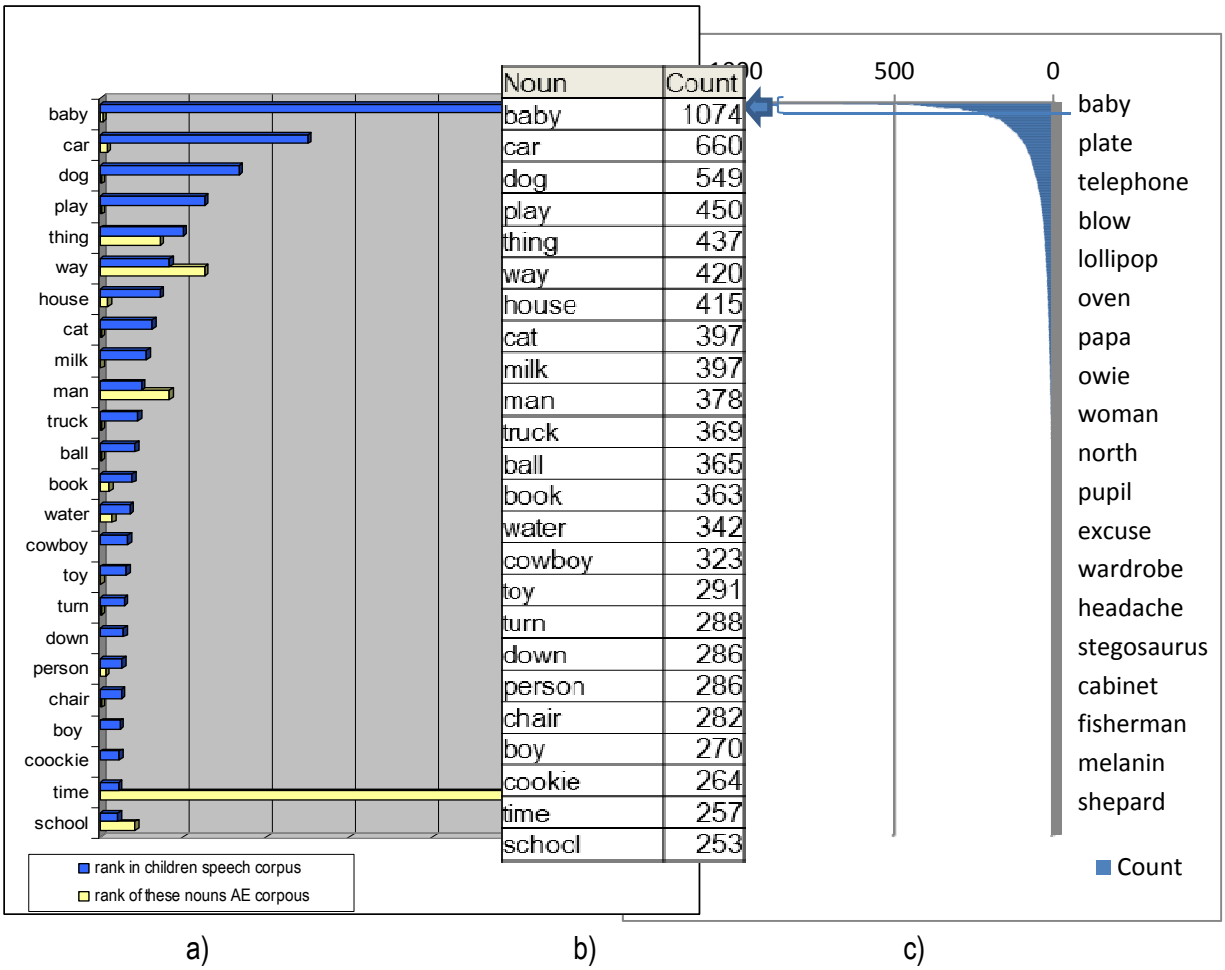
Studies based on application of voxel-wise models and huge fMRI data [Huth et al., 2012, 2016] show that the thousands of distinct object and action categories that humans see and name are represented as locations in a continuous semantic space mapped smoothly across the overall cortical surface. In [Huth et al., 2012] authors used movies to examine the cortical representation of 1,705 object and action categories. The first few dimensions of the underlying semantic space were recovered from the fit models by principal components analysis (PCA). The results suggest that the overall brain activation related to semantics can be represented in a 4-dimensional space which is common across individuals. It is seen from the plots that this fMRI-data derived semantic space is organized across some distant points (meaning that the concepts evoke very specific brain activation) such as "*car*", "*man*", "*face*", "*room*", "*text*" and forms clusters such as "*animals*", "*body parts*", "*humans*", "*communication*", "*structures*", "moving *objects*", "*indoor category*", "*outdoor category*" etc. The finding has been confirmed and unreached recently [Huth et al. 2016]. To model brain responses elicited by *naturally spoken narrative stories* presented to 7 individuals, the authors applied again PCA of voxel-wise model weights. The results for the overall brain activation evoked by words were projected into 985-dimensional word embedding space constructed using word co-occurrence statistics from a large text-corpus. The statistical analyses lead to a space structured around four main axes. After evaluation of the predictive capacity of the proposed model, the authors suggest (again) that most brain areas within the semantic system represent information about specific semantic domains and the obtained brain atlas is common for the individuals (see http://www.nature.com/nature/journal/v532/n7600/full/nature17637.html).

According to our reasoning, if there is a brain atlas common for the individuals, its initial structure would be available at birth and detectable in the first period of language production. Our sample contains speech from the period when children have just started saying recognizable words. In our data from free dialogues, children say a word in order to communicate some idea, so they have a semantic representation for its meaning.

We compared the nouns used often in children speech with the structure of the semantic space derived by Huth and colleagues [Huth et al., 2012]. The examination yielded intriguing results. The Entities

expressed frequently by small children correspond surprisingly well to distant points of the adult fMRI-space (see http://www.cell.com/cms/attachment/2007952467/2030515204/gr5_lrg.jpg).

The nouns observed with significantly higher frequency in child speech are listed in Figure 1.b. A trivial check shows that only five of the first 25 frequently used by children nouns (time, way, man, thing and school) are among the first 25 nouns with high frequencies in the adult's speech corpus. Four of the first 25 "children-preferred" nouns (boy, down, cookie and cowboy) are not among the first 5000 words of the adults' speech (see http://www.wordfrequency.info). The other nouns from the "children list" rank in adults' speech in a quite different way (Figure 1.a.). Obviously there exist differences between children's and adults' vocabulary and experience. Our data suggests that the expressed by children noun-concepts are not a simple consequence of the content of language to which children are exposed in their everyday activities. We assume that the concepts expressed in the children's speech are related to their experience, but through the conceptualization ability available at the corresponding age.



a) b) c)

Figure 1. a) Nouns used the most frequently (from 250 to 1,074 times) observed in our data from child speech, compared with the frequencies in the adult's speech. b) List of the first 25 most frequently used by children nouns, c) general plot of all the frequencies (from 1 to 1,100) of the used by the children nouns

We supposed that the observed coincidence between the Entities expressed frequently by children and the distant points of the adult fMRI-space shows that that the semantic structure is gradually developed around some axes and points, which are available very early (our assumption being that the principles of conceptualization are based on an inborn apparatus for information categorization and insure the basis of the gradual development of the semantic description of the world).

We have discussed the technical aspects related to the data annotation and proposed a model of primary noun concept categories in previous works [Slavova et al., 2016]. Here we give the three main groups of arguments which have led to this set of categories:

1. *Brain studies* [Huth et al., 2012, 2016] - the activation of the overall brain related to semantics has an underlying structure, common across individuals. We have taken into account particularities discussed in this section as well as other suggestions, for example, following [Huth et al., 2016], the perceptual and physical categories (tactile, locational) are separated from human-related categories (social, emotional, violent) by the axis which lies along the first dimension of the common semantic space derived by means of PCA.
2. The *self-centered model of language faculty*, following which the semantic categories are primarily shaped from the point of view of personal proprioceptive, interoceptive and perceptual, situational experience, as explained in the section “basic assumptions”.
3. The observation and the analysis of the child speech collected in our database.

Based on these points, we assembled a set of 14 categories, shown in Figure 2.

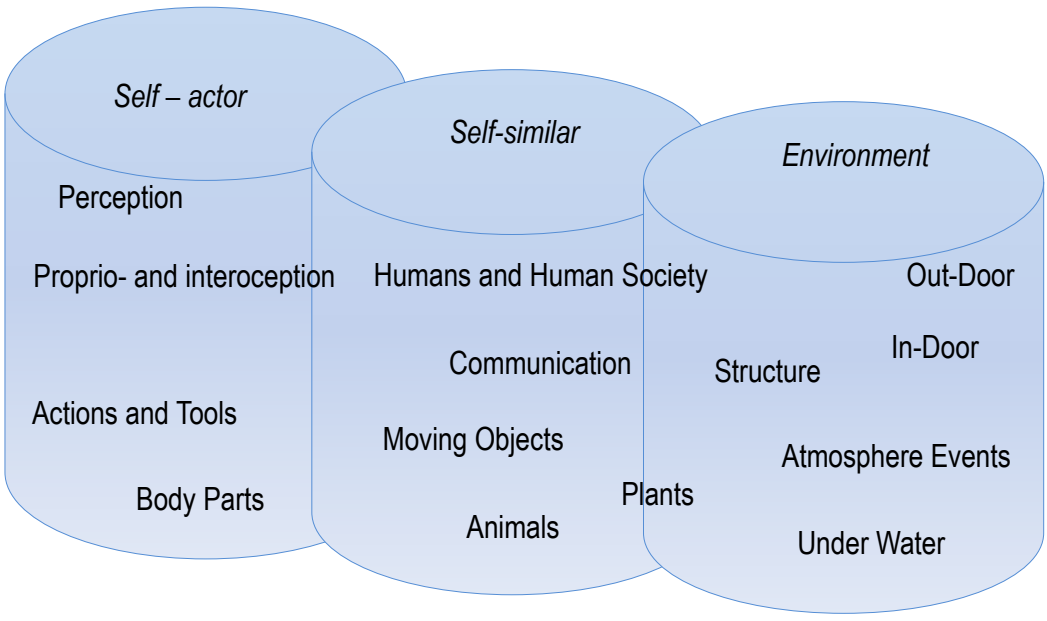


Figure 2. Proposed set of noun concept categories [Slavova et al., 2016]

Investigation of the Use of the Proposed Noun Concept Categories

In order to study the use of nouns from the proposed set, we extracted and treated the nouns from the child speech database, obtaining a big sample of utterances - Nouns (47,777 records) and Proper nouns (12,673 records).

We annotated manually with categories of the proposed semantic set all the common (3026) and proper (1490) nouns extracted from the child speech [Slavova et al. 2016]. Examples for the annotated common nouns for each category are given in table 1.

Table 1. Examples of nouns, extracted from the child speech data

1	Self, proprioception; interoception	baby	hurt	anger	pain	sleep	dress
2	Perception	beauty	light	dark	black	color	voice
3	Actions and tools	action	move	swim	cry	drag	lego
4	Humans and society	friend	guest	family	sport	anger	artist
5	Body parts	arm	beak	beaker	beard	behind	mind
6	Animals	bat	bird	bunny	cat	chicken	mouse
7	Plants	beet	bran	branch	bush	flower	mulberry
8	Moving objects	aeroplane	automobile	baby	barge	beetle	locomotive
9	Indoor	bath	bed	blanket	bowl	candle	room
10	Outdoor	alley	avalanche	avenue	barbecue	forest	river
11	Communication	letter	library	magazine	newspaper	paper	joke
12	Atmospheric	air	tornado	candle	comet	dust	plane
13	Underwater	aquarium	corral	dive	noun	octopus	water
14	Structure	cube	edge	machine	block	count	evening

The "Self-proprioception and interoception" category contains nouns for pain, hunger (including foods), emotional states, feelings (i.e. coldness, warmth) and clothes. The "Perception category" contains nouns for incoming multimodal perceptual information. The "Structure" category contains materials, machines, spatial forms, time and time-periods, quantities, and socially determined structures such as hospitals and cities. The content of the other categories is clear from the given examples.

In order to evaluate how the proposed categories are used by children with relation of their age and gender, we calculated for each dialogue taken from CHILDES the Ratio per Utterance (RU) of the 14 categories:

$$\mathrm{RU}\left(\mathrm{Cat_{ij}}\right) = \frac{\mathrm{NCat_{ij}}}{\mathrm{N_i}} \qquad (1)$$

Where :

RU(Cat$_{ij}$) is the ratio of use in the dialogue *i* the category *j;*

NCat$_{ij}$ is the number of the nouns from category *j* in the dialogue *i;*

N$_i$ is the number of recognizable word-forms pronounced by the child in the dialogue *i*.

Statistical model

Based on the derived Ratio per Utterance we have realized a linear logistic model given with expression (2). For predictor parameters we have used the Age, the Gender and the Type of the noun (proper noun or common noun):

$$logit(\frac{Pr}{1-Pr}) = I + A * age + G * gender + T * type + AG * (age \times gender) + TG * (type \times gender) \quad (2)$$

Where:

A, G, T, AG, TG are the unknown parameters of the model;

the *Cartesian products* note the mixed effect of the two parameters involved;

Pr is the probability for performing (using in the speech) given concept category.

The results we have obtained are shown in Table 2, where with bold we give the statistically significant parameters (for which the estimated *p-value* is less than 0.1). As it is seen from the results, the parameters Gender, Age and Type influence each the categories except one – Plants.

The observations of use of nouns in the categories "Atmospheric events" and "Underwater" are further discarded because of their insufficient number in our sample. They are well-pronounced as distant "special points" in the adult's fMRI semantic space obtained by Huth and colleagues [Huth et al. 2012], but according to our data they are not used early in childhood.

Table 2. Estimation of the model parameters

model categories	A	G	A*G	T*G	T
01 Self, proprioeption, interoception	**0.0309**	**3.4932**	***-0.0079***	***-2.7225***	**6.4785**
02 Perception	***-0.0306***	1.2405	**0.0114**	***-1.8828***	**6.3894**
03 Actions and tools	***-0.0156***	0.8979	**0.0175**	***-1.8882***	**7.3912**
04 Humans and society	**0.0090**	0.0109	***-0.0159***	**0.8203**	***-5.4844***
05 Body parts	**0.0260**	0.9772	***-0.0192***	*-0.3763*	25.1459
06 Animals	**0.0248**	**0.9663**	***-0.0090***	***-0.3648***	**2.9320**
07 Plants	***-0.0096***	18.0716	*-0.0040*	18.4748	13.9519
08 Moving objects	**0.0169**	*-0.7577*	**0.0071**	0.4570	**5.0394**
09 Indoor	**0.0196**	*-0.8133*	***-0.0085***	1.4555	22.0610
10 Outdoor	***-0.0303***	*-1.8982*	**0.0165**	0.9104	23.0149
11 Communication	***-0.0136***	*-3.6582*	**0.0111**	3.2120	20.7665
12 Atmospheric	***-0.0345***	*-0.5586*	**0.0229**	*-0.6173*	24.6671
13 Underwater	***-0.0447***	*-2.0171*	**0.0271**	0.5412	21.9022
14 Structure	***-0.0138***	***-0.5373***	0.0013	**0.3736**	**2.0220**

As the ratios of use all the 14 categories are not independent in the speech, the results in table 2 cannot be seen as a hall, but has to be analyzed category by category. Examples of plots are shown in figure 3.

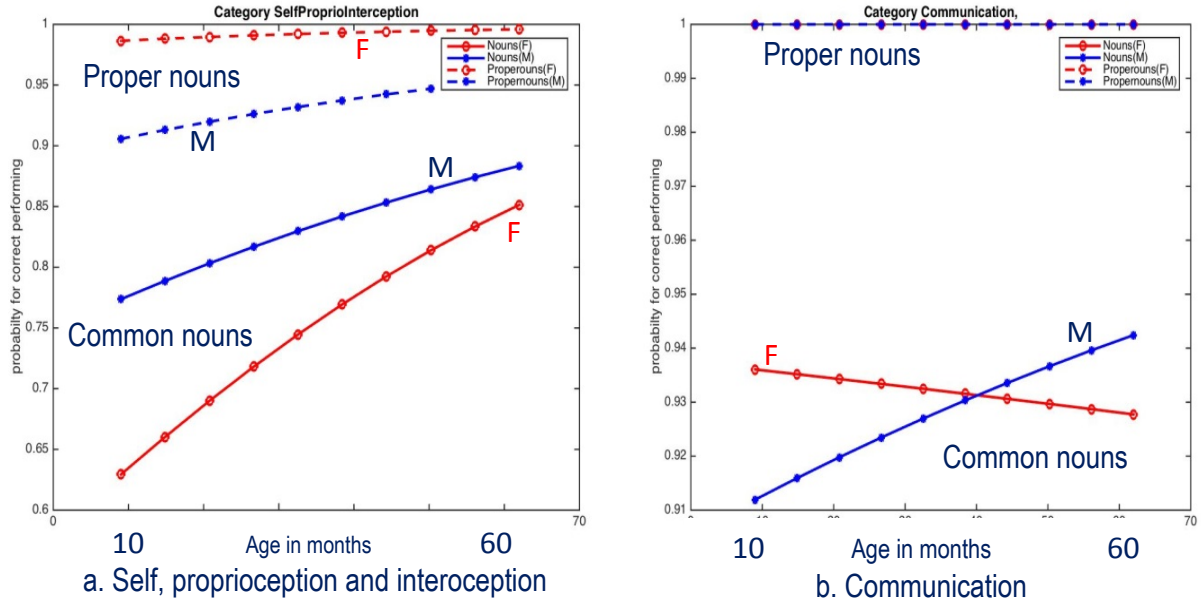


Figure 3. Examples of use of nouns and proper nouns in two of the categories.

Figure 3.a illustrates how develops with the age the probability for performance of gills (red) and boys (blue) of use of the category “Self-proprioception and interoception”, for the two types of nouns. As seen, for both genders this probability of use increases. The same dependencies are shown in figure 3.b. for the category Communication, where Proper nouns are not observed and are displayed as a constant. As it is seen, the performance of boys gradually increases, whereas girls tend to decrease the relative use of nouns from this category.

The results for all the other categories are obtained in the same way.

Analysis of the Results with regard of Gender Differences

The graphics in figure 4 expresses the overall result, where the categories are grouped depending on the influencing parameters. As shown, the gender influences the use of three of the categories – Self - proprioception and interoception, Structure and Animals (their plots are given in figure 3.a, figure 5.a and figure 5.b respectively).

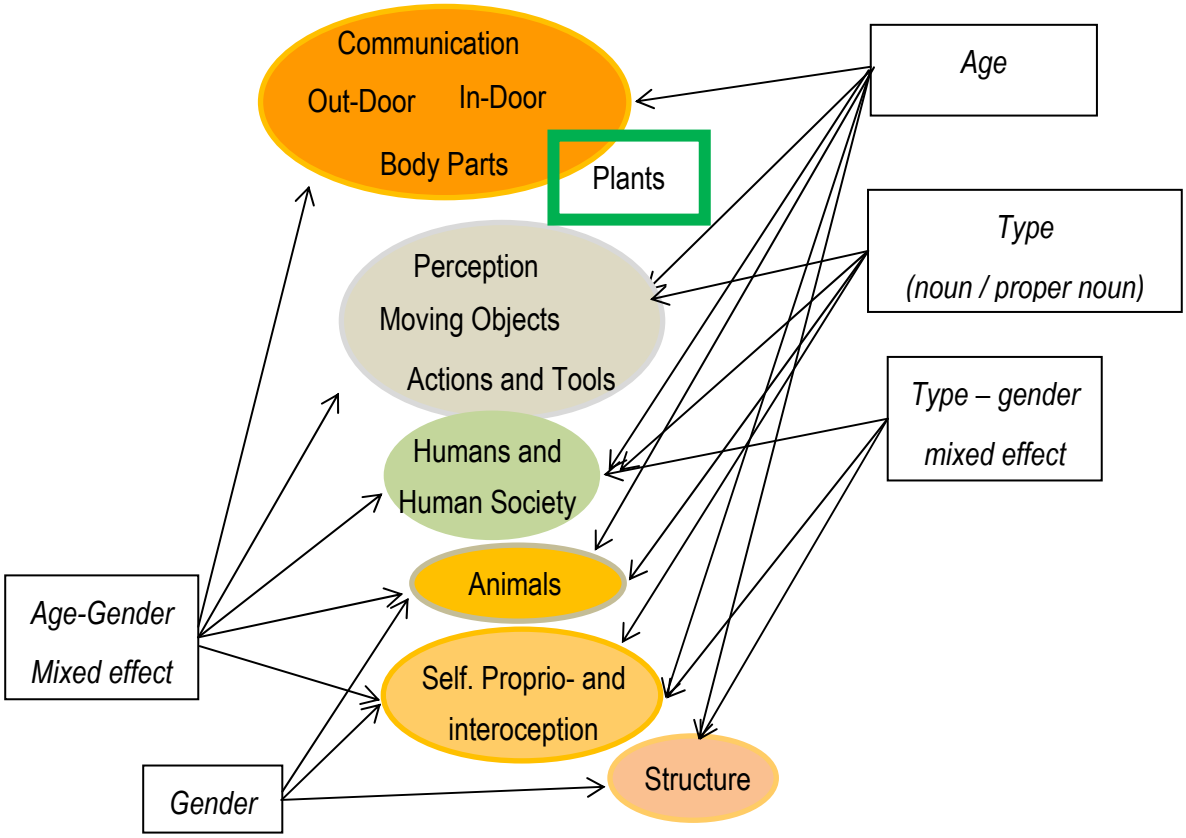
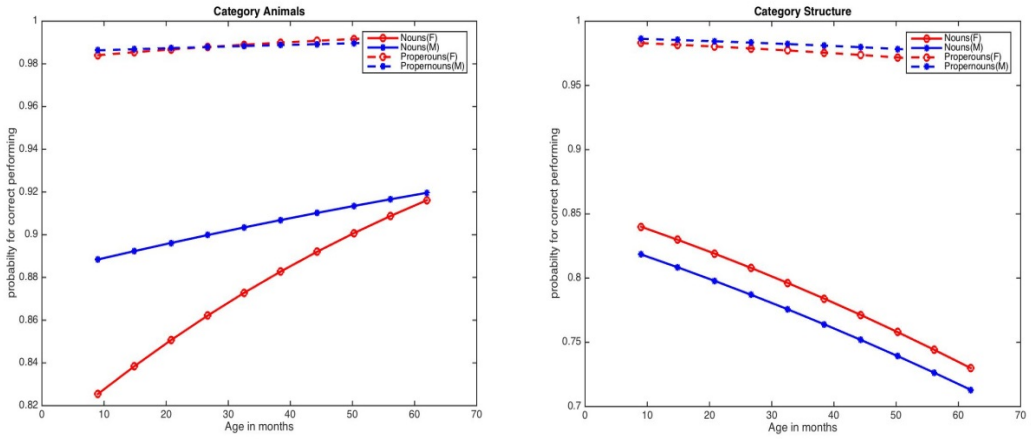


Figure 4. Influence of the parameters Age, Gender, Type of noun and their mixed effect on the proposed categories.

The relative use of the category “Structure” decreases with age for both genders, in the same way (figure 5.b). We have determined this category as a semantically large set (machines, spatial forms, time and time-periods, quantities, materials, and socially determined structures), so its split into sub-categories could give a more precise picture of its decreasing tendency of use.



a. Animals b. Structure

Figure 5. Use of nouns and proper nouns in two of the “gender dependent” categories by boys and girls.

Next, we examine how develops over the age the categorical content of boys’ and girls’ vocabulary. The value found for the mixed effect of Age and Gender (A*G) given in Table 2 shows how progresses the use of noun categories by the two genders.

Table 3. Gender differences in the use of noun concepts over the course of language acquisition

Categories in which ***girls*** show relatively increasing use of nouns		Categories in which ***boys*** show relatively increasing use of nouns	
model categories	**A*G**	**model categories**	**A*G**
Body parts	***-0,019***	**Actions and tools**	**0,018**
Humans and society	***-0,016***	**Outdoor**	**0,017**
Animals	*-0,009*	**Perception**	**0,011**
Indoor	*-0,009*	**Communication**	**0,011**
Self, proprioeption, interoception	*-0,008*	Moving objects	0,007

We have annotated the two genders with 1 for the girls and with 2 for the boys, so the positive value of A*G means that the use of the corresponding category by boys increases at a faster rate than that of girls. (Or, that the use of the category by boys decreases at a slower rate than that of girls.) A negative value obtained for A*G means that girls develop more intensively the use of nouns from this category that boys do.

After having taken into account only the statistically significant A*G (for *p-values* less than 0.1), the reorganization of the results from Table 2 provides the two lists given in Table 3. The lists show the concept-related differences in the "gender dependent development of the vocabulary" (in bold are the categories for which the differences are considerable, threshold taken at 0.01).

These "gender-distinctive" lists of categories showed up after the concluding analysis of the statistical result obtained from the vast corpus of child speech. As it can be seen, the result corresponds to the classical view about the social function of "men-hunters" and "women-gatherers".

It is not possible from our data to distinguish between abilities which are learned and abilities which are innate. However, our statistical finding supports in general the supposal that the observable gender distinctions may date to early human history.

Conclusion

We proposed a set of "primary" noun-concept categories, supposing that the conceptualization mechanisms are functional at birth and taking into account results obtained in studies of the semantic brain activation in adults.

We classified the nouns from a huge database of child speech to the proposed set. The linear logistic model we have applied gave statistically reliable results concerning the influence of Age, Gender and Type of noun on the use of these categories.

The obtained statistical result is in compliance with suggestions and general theories concerning social, cognitive, anthropological etc. gender differences.

As we don't treat purely language-related parameters such as richness of vocabulary, correctness, grammatical parameters etc., we accord the result to the establishment of initial mental representations obtained in interaction with the environment.

At this point, the proposed set of noun-concept categories gives a meaningful picture and the statistical result indicates directions for further investigation and adjustment of the noun-concept's model.

Acknowledgement

This paper is published with partial support by the ITHEA ISS (www.ithea.org) and the Central Fund for Strategic development, New Bulgarian University.

References

[Atanasov et al, 2016] Atanasov, D., Slavova V. & Andonov F. A statistical study of first language acquisition: no gender differences in the use of parts of speech, in proc. of the 12th Annual International Conference on CSECS 2016, Germany, in print

[Balter, 2013] Balter, M. Language Gene More Active in Young Girls than Boys. ScienceMag.org, Balter, M. (2013, February 19). http://www.sciencemag.org/news/2013/02/language-gene-more-active-young-girls-boys

[Barsalou, 2003] Barsalou, L. W. Situated simulation of human cognitive processes, Language and cognitive processes, 2003, 18 (516), 513-562

[Burman el al, 2008] Burman, D. D., Bitan, T., & Booth, J. R. Sex differences in neural processing of language among children. Neuropsychologia, 46(5), 2008. 1349-1362.

[Eckert & McConnell-Ginet, 2003] Eckert, P., & McConnell-Ginet, S. Language and gender. Cambridge University Press, 2003.

[Hart & Risley, 1995] Hart, B., & Risley, T. R. Meaningful differences in the everyday experience of young American children. Paul H Brookes Publishing, 1995.

[Hart, 2000] Hart, B. A natural history of early language experience. Topics in Early Childhood Special Education, 20(1), 2000, 28-32

[Huth et al, 2012] Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. Neuron, 76(6), 2012, 1210-1224.

[Huth et al, 2016] Huth A.G., de Heer W.A., Griffiths T.L., Theunissen F. E. & Gallant J. L. Natural speech reveals the semantic maps that tile human cerebral cortex, Nature, 2016 .

[Longobardi et al, 2016] Longobardi, E., Spataro, P., Frigerio, A., & Rescorla, L. Gender differences in the relationship between language and social competence in preschool children. Infant Behavior and Development, 43, 2016, 1-4.

[Northwestern University, 2008] Northwestern University. Boys' And Girls' Brains Are Different: Gender Differences In Language Appear Biological. ScienceDaily. 2008, March 5. www.sciencedaily.com/releases/2008/03/080303120346.htm

[Piaget, 1955] Piaget, Jean. The Language and Thought of the Child, New York: The World Publishing, 1955, 1924/55

[Skinner, 1953] Skinner, B. F. Science and human behavior. Simon and Schuster. 1953

[Slavova & Soschen, 2015] Slavova, V., A. Soschen, On mental representations: Language structure and meaning revised, International Journal Information theories & applications. 2 (4), 2015, 316-325.

[Slavova, 2016] Slavova V. Data collection for studying language acquisition, in proc. of the 12th Annual International Conference on CSECS 2016, Germany, in print

[Slavova et al, 2016] Slavova V., Andonov F. & D. Atanasov (2016), A study of noun concept categories using data from child language acquisition – an outline, in proc. of the 12th Annual International Conference on CSECS 2016, Germany, in print

[Slik, Hout & Schepens, 2015] van der Slik FWP, van Hout RWNM & JJ Schepens (2015) The Gender Gap in Second Language Acquisition: Gender Differences in the Acquisition of Dutch among Immigrants from 88 Countries with 49 Mother Tongues. PLoS ONE 10(11): e0142056. doi:10.1371/journal.pone.0142056

[Tomasello, 2003] Tomasello, M. Constructing a Language: A Usage-Based Theory of Language Acquisition. Cambridge, MA: Harvard University Press, 2003.

Authors' Information

***Velina SLAVOVA**, New Bulgarian University, department of Computer Science, vslavova@nbu.bg*

***Major Fields of Scientific Research**: AI, Cognitive Science*

***Dimitar ATANASOV**, New Bulgarian University, department of Computer Science, datanasov@nbu.bg*

***Major Fields of Scientific Research**: Probability, Statistics and related fields, Psychometrics.*

***Filip ANDONOV**, New Bulgarian University, department of Computer Science, fandonov@nbu.bg.*

***Major Fields of Scientific Research**: multicriteria optimization, data mining, text processing, Python language*