Comparing Child and Adult Language: Exploring Semantic constraints

Ismaïl El Maarouf
Licorn, Valoria, Université de Bretagne Sud UEB
Rue de Saint Maudé, BP 92116
56321 Lorient, France
ismail.el-maarouf@univ-ubs.fr

Jeanne Villaneau
Valoria, Université de Bretagne Sud UEB
Rue de Saint Maudé, BP 92116
56321 Lorient, France
jeanne.villaneau@univ-ubs.fr

Farida Saïd
LMAM, Université de Bretagne Sud UEB
Rue de Saint Maudé, BP 92116
56321 Lorient, France
farida.said@univ-ubs.fr

Dominique Duhaut
Valoria, Université de Bretagne Sud UEB
Rue de Saint Maudé, BP 92116
56321 Lorient, France
dominique.duhaut@univ-ubs.fr

ABSTRACT

Actual research on child-machine interaction indicate that children are specific with respect to various acoustic, linguistic [7], psychological, cultural and social factors. We wish to address the linguistic factor, focusing on the semantic knowledge which needs to be mastered by a computer system designed to interact with children. Our work is intentionally usage-based and application-driven.

The research was conducted in the frame of the EmotiRob project, which aims at building a companion robot for children experiencing emotional difficulties. The robot is supposed to understand the emotional state of the child and respond (albeit non linguistically) adequately [1]. The interactional capacities are heavily dependent on the results of the comprehension module. The comprehension model incorporates semantic knowledge such as children-based ontologies and specific semantic associative rules.

Our study is based on a corpus of Fairy Tales, which will later be compared to an oral corpus when the latter is completed. We argue that lexical knowledge and semantic associations discovered in this corpus will not differ greatly between writing and speech. Fairy Tales constitute privileged material for teachers and psychologists who argue that they play a crucial role in child socialization and structuration of concepts.

To spot child language specificities, we provide a contrastive analysis of semantic preferences according to production (child VS adult authored text) and to reception (child VS adult destined text). We use a shallow ontology to compare verb constraints on specific syntactic positions in child VS adult texts. Preliminary results show, as expected, a significant difference in terms of reception, though questioning the idea that adult language is much more constraining, while differences in terms of production are less obvious and call for a detailed qualitative study.

ICMI-MLMI'09 Workshop on Child, Computer and Interaction November 5, 2009, Cambridge, MA, USA

Categories and Subject Descriptors

I.2.7 [**Artifical Intelligence**]: Natural Language Processing—*Language parsing and understanding, Text analysis*; J.5 [**Computer Applications**]: ARTS AND HUMANITIES—*Linguistics*; I.5.3 [**PATTERN RECOGNITION**]: Clustering—*Similarity measures*; H.5.2 [**Informations Interface and Presentations**]: User Interfaces—*Natural language*

General Terms

Language, Human Factors

Keywords

child language specificities, semantics, NLP, corpus analysis, contrastive analysis

1. SEMANTIC CONSTRAINTS

The linguistic feature studied here is the so-called selectional restrictions between verbs and nouns [11, 8]. It concerns semantic constraints imposed by verbs on the semantic type of their arguments so that violations of these constraints can be interpreted as clues of deviations, such as figurative language, errors, and so on. For example the verb "think" expects a [[HUMAN]] in its subject slot which is contradicted in (1):

(1) *The bird thinks we'd better take a shortcut.*

Semantic constraints have been studied in psycholinguistics, and used as a criterion to analyze how children develop and organize semantic knowledge.

The hypothesis is that children are not fully aware of semantic constraints [6] and produce more semantically incorrect sentences than adults. Duvignau and her colleagues [4] compared adult and children production of anomalous sentences in experiments and showed that the phenomenon they call "semantic approximations" is specific to children. In other words, children allow for semantic combinations such as (2), not observed in adult speech.

(2) *Mary is undressing the orange.*

The anomaly can be spotted here by a system which would only associate the verb "to undress" with words belonging to the semantic type [Human]. The shift of semantic type entails detecting the oddness and reinterpreting the verb ("to peel").

We wish to test whether semantic constraints can be used to characterize children language. We constituted 4 corpora. We first compare adults and children production with respect to an equivalent task: Fairy Tales writing. We then contrast adult with children reception by selecting a corpus representing a sample of each of their universe, namely Press and Fairy Tales.

2. DATA AND METHOD

2.1 Size and Productivity

For the purpose of analyzing the constraining power of verbs in a corpus, we argue that **frequency** is not as important a criterion as **productivity**. Frequency corresponds to the number of times the same event is observed, while productivity designates the number of different events observed in a specific configuration, regardless of their respective frequency [2]. As we are more interested in how many different semantic types may occur in a specific linguistic context, productivity becomes primary.

For example, the number of different words in object position with respect to the verb "prendre" ("to take") is 142 in the adult-Audience corpus and 137 in the children-Audience corpus, while their cumulated frequency is respectively of 1021 and 284.

Our analysis of Reception is based on a Fairy Tales corpus of 170,000 words contrasted with a Press corpus of 1,200,000 words. These two corpora should not share anything in common since they represent completely different linguistic worlds which are representative of Children universe and Adult universe respectively. We however found a common vocabulary of 1800 words: 120 verbs and 1680 arguments (17% of Press arguments and 58% of Fairy Tales arguments).

Our analysis of Production is based on two Fairy Tales sub-corpora which are much more balanced in terms of frequency (roughly 60 000 running words for each). What is more, they share an equivalent text genre and should therefore show a broader similarity of vocabulary. We noticed that the main difference concerns text length: children productions are usually shorter.

2.2 Corpus Processing

The corpora were parsed semi-automatically with the help of Syntex [3] so as to draw lists of lexical units occurring in specific syntactic positions (subject, object,...) with respect to a given verb (called lexical distributional databases). The lexical units are then matched with a general-purpose shallow ontology (containing 40 semantic types) in order to obtain the list of semantic types which fit in a specific verbal position.

For example, the verb "voler" in the meaning of "to fly", cooccurs with arguments such as "avion" ("plane") in the adult-Speaker corpus, which belongs to the type [Vehicle], whereas it strongly collocates with arguments such as "oiseau" ("birds") in the child-Speaker corpus ([Flying Animal]). We however do not deal with the phenomenon of polysemy and associate every word to its main semantic type. A proper treatment of polysemy implies indexing every verb and noun according to one of its senses.

2.3 Method : comparing verb classes

In order to have a comparable basis, we extracted from the distributional databases only the couples which featured common vocabulary. We here wish to study how adults and children combine verbs with nouns, considering that they possess the knowledge of each of them. Thus, we only selected the verb-noun couples for which each of the verb and noun were present in both corpus, independently of their co-occurrence. One couple may be authorized in one corpus while it may not be observed in the other.

To compare verbs accross corpora, we analyzed productivity and semantic constraints in the following way:

To measure each verb's constraining power, we simply compute the number of different semantic types observed in a specific position, a measure we call semantic productivity.

We then evaluated how semantic constraints differed from one corpus to another by clustering semantic types occurring in the same syntactic position with respect to each verb. We defined the similarity index between two semantic types as the ratio of their shared verbal contexts to the total verbal contexts in the study. The dissimilarity between semantic types i and j is then computed as:

$$d_{ij} = 1 - \frac{n_{ij}}{n} \tag{1}$$

where n_{ij} is the number of shared verbal contexts and n is the total number of verbal contexts in consideration.

We used Ward's hierarchical clustering method to build clusters of semantic types using the dissimilarity measure above. Each semantic type is initially assigned to its own singleton cluster. The analysis then proceeds iteratively, at each stage joining in a new cluster the two clusters whose fusion results in minimum increase in 'inertia loss', continuing until there is one overall cluster.

The inertia loss resulting from the fusion of singleton clusters $[i]$ and $[j]$ is given by :

$$\Delta_{ij} = \frac{m_i m_j}{m_i + m_j} d_{ij}^2$$

where d_{ij} stands for the dissimilarity measure between i and j and m_i, m_j their respective masses.

The resulting cluster $I = [i, j]$ is of mass $m_I = m_i + m_j$ and grouping it with another cluster $[k]$ results in the following inertia loss :

$$\Delta_{Ik} = \frac{1}{m_I + m_k} \left\{ (m_i + m_k)\Delta_{ik} + (m_j + m_k)\Delta_{jk} - m_k \Delta_{ij} \right\}.$$

Figure 1: Children-Reception dendrogramme of semantic types using Ward criterion.

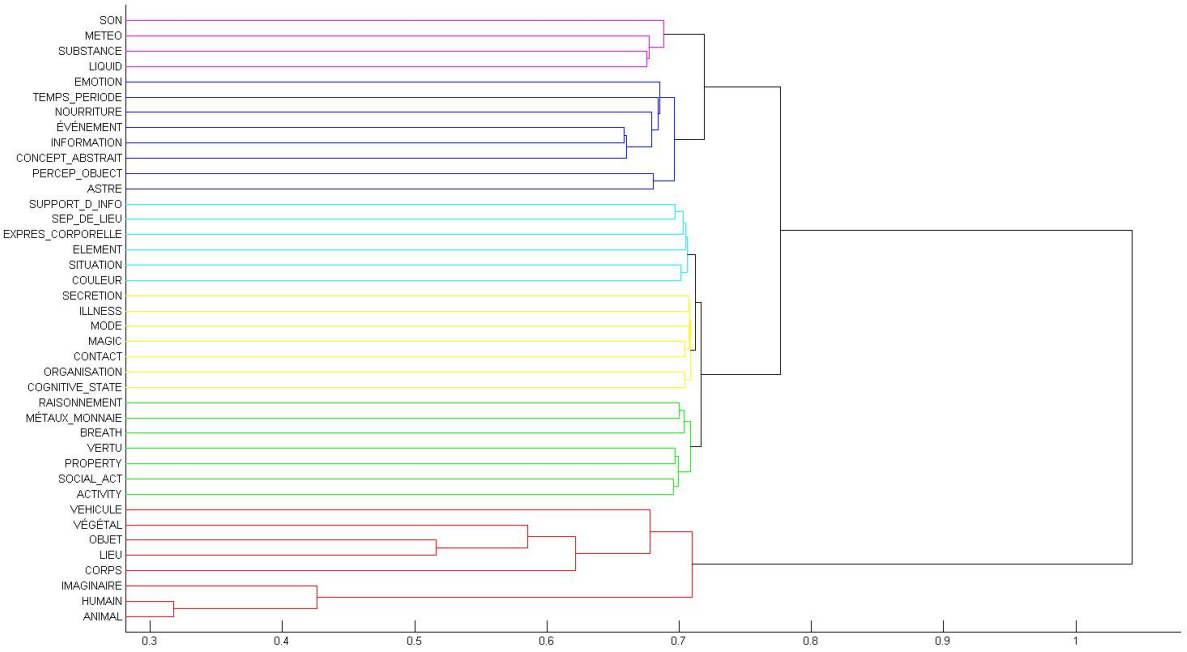
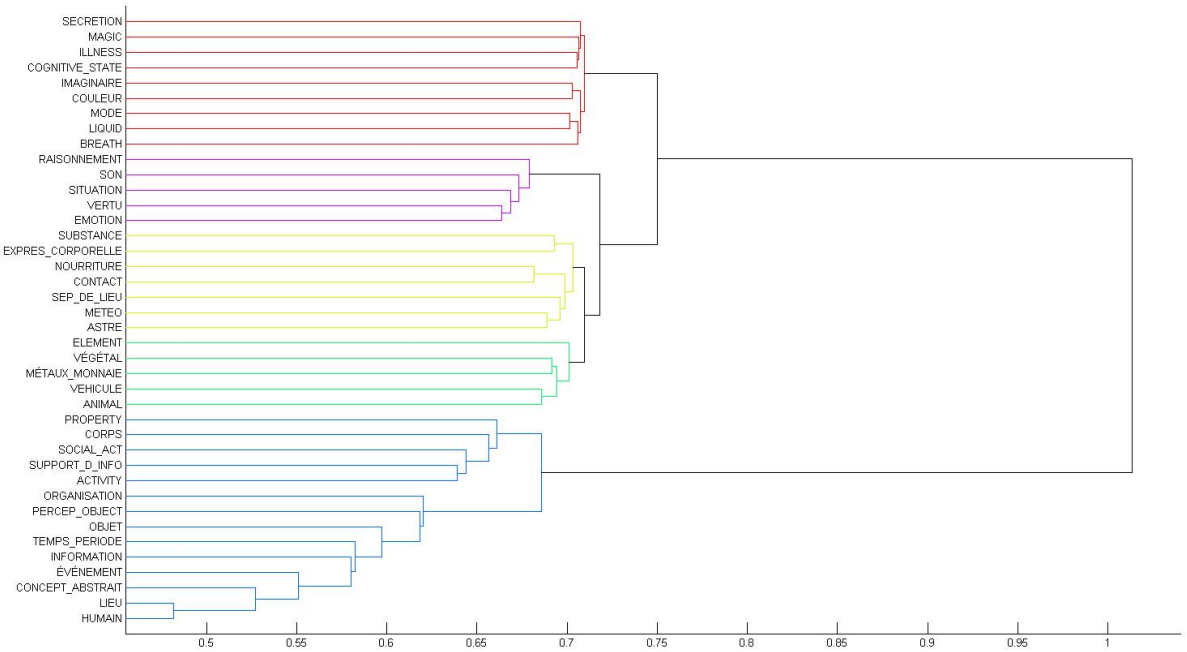


Figure 2: Adult-Reception dendrogramme of semantic types using Ward criterion.



3. RESULTS

3.1 Analysing verb constraining power

To have a global view on verbs' constraining power we computed the verb mean productivity of each sub-corpus:

Table 1: Verb Mean Productivity for Adults and children according to Reception and Production.

Audience	Feature	
	Reception	Production
Children	4.8	3
Adult	10.7	3.4

We observe that, on a similar task (Production), Adults and Children do not significantly differ in terms of their use of semantic constraint. This result questions the hypothesis that children allow for loose semantic associations. We have looked up each verb and could not identify similar phenomena to semantic approximation.

On the contrary, the verb mean productivity difference is much more important regarding Reception: the adult corpus shows a wider flexibility of semantic types. Again, this result goes against the idea that adults associate specific semantic types to specific verb positions.

In the face of such results, we conducted a qualitative verb-by-verb analysis to see whether the mean productivity criterion hid specific uses of verbs in Production. We observed that movement ("to jump, to run") and emotional ("to like") verbs seem to be more frequent and more productive in children productions.

3.2 Clustering semantic types

There are two main findings which stem from the cluster analysis.

The most productive class is [Human]; it combines with most verbs: a [Human] talks, moves, likes, thinks, uses, etc. It occurs at least with 50% of verbs in every corpus.

1. The Reception corpora both show regular semantic violations for several verbs. As expected, the text genre influences semantic constraints and therefore, the clustering process:

   For Fairy Tales (see Fig.1), the [Animals], [Plants and Trees] and [Imaginary Creatures] classes share strong similarity with the [Human] class, since they put on characteristics which are usually attributed to humans (speaking, thinking, etc.). We call this process humanization. [Objects] and [Locations] are also clustered with [Humans] but this also happens in the Press corpus.

   In the Press corpus (see Fig.2), we observed the same semantic extension with classes like [Organization], [Abstract Concept], [Event] and [Information], which is explained by the typical metonymical language used and destined to Adults, as in (3).

   (3) *The government declared a state of emergency.*

2. No significant difference of this kind was observed regarding Production, since both Adults and Children allow for talking trees and thinking rabbits when it comes to writing a Fairy Tale. However, semantic violations were helpful to spot idiomatic expressions and metaphorical language, which, according to the data, are much more common in Adult productions, as attested by French examples (4) and (5)

   (4) *Les blessures qui déchirent vos coeurs.*

   trans.: The wounds which tear your hearts out.

   (5) *Il s'y est cassé les dents.*

   trans.: He broke his teeth. [He tried hard but could not make it]

   This would indicate that adults are much more creative in terms of metaphors than children and that they also master idioms better than them.

4. APPLICATIONS - PERSPECTIVES

It is clear from our research that the method using semantic restrictions cannot help us to identify potential differences between adult and children productions.

Our study seems to indicate that children do not differ significantly from adults in terms of semantic constraints when involved in a similar task (Fairy Tales writing). Such differences may remain at the syntactic level in the mastering of specific complex constructions ([9, 10]). However, the semantic universes sampled here can be contrasted thanks to a cluster analysis of semantic types applied to each of the reception corpora. Concerning the semantic level, it appears that the situational context in which an interaction is engaged has a greater impact on verb-noun combinations than the kinds of Speaker involved in it.

We are actually working on the possibility of automatizing this method, basing ourselves on larger structures, called semantic patterns [5], which incorporate syntactic as well as semantic information in order to evaluate to which extent situational context can be linguistically characterized. One important aspect of semantic patterns is that each pattern correspond to a single verb meaning.

If we accept the premises of this research, then our contrastive analysis of Reception corpora justifies the fact that children and adults evolve in different semantic universes. This, in turn, entails that the robot's knowledge should be based on children-targeted data so that the robot's comprehension module makes the right semantic associations out of children's input. We have thus, for every verb, created a database of all and only the semantic types which were found associated in each syntactic position in the Fairy Tales corpus. The corpus helped us tuning the robot's knowledge and constituting a first version of the system. This work now needs to be validated and completed with corpora collected in real situation, such as children playing and telling stories to their plush robot.

We wish to insist on the fact that such corpus analyses, based on real data, provide authentic information to build cognitive models, which will later be applicable to real children-machine interactions.

5. REFERENCES

[1] A. Achour, J. Villaneau, D. Duhaut, and F. Said. Cognitive and emotional linguistic interaction. In *Child, Computer and Interaction (ICMI'08 post-conference workshop)*, Chania, Crete, Greece, October 2008.

[2] D. Bourigault. Upery : un outil d'analyse distributionnelle étendue pour la construction d'ontologies à partir de corpus. In *Actes de la 9ème conférence annuelle sur le Traitement Automatique des Langues*, pages 75–84, 2002.

[3] D. Bourigault. *Syntex, un analyseur syntaxique opérationnel*. Habilitation à diriger les recherches, Toulouse, Université Toulouse II Le Mirail, 2007.

[4] K. Duvignau, M. Fossard, B. Gaume, M.-A. Pimenta, and J. Elie. Semantic approximations and flexibility in the dynamic construction and deconstruction of meaning. In N. Moura, Vieira, editor, *Metafora e contexto / Metaphor and Context*, volume 7, pages 371–389, 2007.

[5] P. Hanks. Lexical patterns: From hornby to hunston and beyond. In *Proceedings of the XIII EURALEX International Congress*, pages 89–129, 2008.

[6] S. L. James and J. F. Miller. Children's awareness of semantic constraints in sentences. In *Child Development*, volume 44, pages 69–76. 1973.

[7] A. Potamianos and S. S. Narayanan. A review of the acoustic and linguistic properties of children's speech. In *Proceedings of IEEE Multimedia Signal Processing Workshop*, Chania, Crete, Greece, 2007.

[8] J. Pustejovsky and E. Jezek. Semantic coercion in language: Beyond distributional analysis. In *Distributional Models of the Lexicon in Linguistics and Cognitive Science*. 2008.

[9] G. Sampson. The structure of children's writing: moving from spoken to adult written norms. In *Extending the Scope of Corpus-Based Research*, pages 177–193. 2003.

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

[11] A. Wagner and M. Mastropietro. Collecting and employing selectional restrictions. In *Papers of the First Swiss-Estonian Student Workshop on Computational and Theoretical Linguistics*, 1996.