Morphological Complexity of Children Narratives in Eight Languages

Gordana Hržica[1], Chaya Liebeskind[2], Kristina Š. Despot[3], Olga Dontcheva-Navratilova[4], Laura Kamandulytė-Merfeldienė[5], Sara Košutar[1], Matea Kramarić[1], Giedrė Valūnaitė Oleškevičienė[6]

[1] University of Zagreb, Zagreb, Croatia, gordana.hrzica@erf.unizg.hr
[2] Jerusalem College of Technology, Jerusalem, Israel, liebchaya@gmail.com
[3] Institute for the Croatian Language and Linguistics, Zagreb, Croatia, kdespot@ihjj.hr
[4] Masaryk University, Brno, Czech Republic, navratilova@ped.muni.cz
[5] Vytautas Magnus University, Kaunas, Lithuania, laura.kamandulyte-merfeldiene@vdu.lt
[6] Mykolas Romeris University, Vilnius, Lithuania, gentrygiedre@gmail.com

Join us on Zoom for the discussion.

This study is based upon work from COST Action NexusLinguarum - European network for Web-centered linguistic data science (CA18209), supported by COST (European Cooperation in Science and Technology, https://www.cost.eu/)

1. Introduction

The **morphological complexity (MC)** of a language refers to the richness of inflection, i.e., a higher number of different nominal case forms, more structural units and rules or representations indicate greater complexity (Kettunen, 2014). More recently, with the development of computer-based methods for assessing MC, a more systematic research of complexity has been undertaken (for an overview of common methods for quantifying complexity, see Bane, 2008). In this study, we applied four measures of MC to language samples from children speaking different languages to see if these measures reflected typological differences between languages and age differences.

1) **Word entropy** is a measure of the predictability and uncertainty of information conveyed by strings of symbols or words in a text or language (cf. Bentz et al., 2016). This reflects the average information content of words.
2) The **relative entropy of word structure** focuses on measuring the information content of the internal structure of words (Bentz et al., 2016). It measures the information stored in the words via morphological regularities (Dehouck, 2019).
3) **Kolmogorov complexity** is a measure of structural surface redundancy based on the repetition of orthographic strings in a text (Juola, 1998; Ehret and Szmrecsanyi, 2016). The measure is based on a compression technique, which means that the complexity of a given text is evaluated as the length of the ultimate shortest description of the text.
4) **Bane's** (2008) **measure** is used to separate word stems, affixes, and signatures and then calculate the description lengths of these strings. The morphological complexity of a language is calculated by dividing the description length of the affixes and stems by the total description length of the model.

2. Aim and hypotheses

To apply measures reflecting morphological complexity to language samples of children speaking different languages.

We compared corpora representing the language production of younger and older children to gain information about the morphological complexity of languages and to show morphological complexity from a typological perspective.

H Measures will show higher results for corpora of older children narratives than for corpora of younger children narratives, and higher results for morphologically more complex languages.

3. Methodology

Materials

The corpus used in this study is a selection from the Frog Story subcorpus of CHILDES corpora (MacWhinney, 2000).

- Oral narratives collected by various researchers between 1990 and 2005
- Based on *Frog, where are you?* → the 29-page wordless picture book by Mercer Mayer (1969)
- The narratives were collected and transcribed using the same procedure (for a full description, see Berman & Slobin, 1994)

The transcripts in seven languages were selected: Croatian, English, French, German, Italian, Russian, and Spanish. Additionally, samples of Lithuanian language, collected according to the same principles, were added. The languages differ in their inflectional morphology and represent four different language families: Slavic, Germanic and Romance, Baltic. Overall, the corpus included 249 narratives evenly distributed across eight languages.

Data analysis

Two subcorpora were formed for each language: a younger children corpus and an older children corpus. Available data was neither comparable in the number of transcripts nor in the age range of children. To form comparable corpora in each of the languages, transcripts were selected to form corpora similar in size within a language, with the maximum distance of the age range of children in younger vs. older corpora.

Table 1. The size of the two subcorpora in words for each of the eight languages

Language	Younger children	Older children
German	4090	4214
Spanish	6348	6708
Russian	2111	2182
Lithuanian	2649	3132
Italian	10947	11715
French	3050	2971
English	5385	5021
Croatian	3476	3497

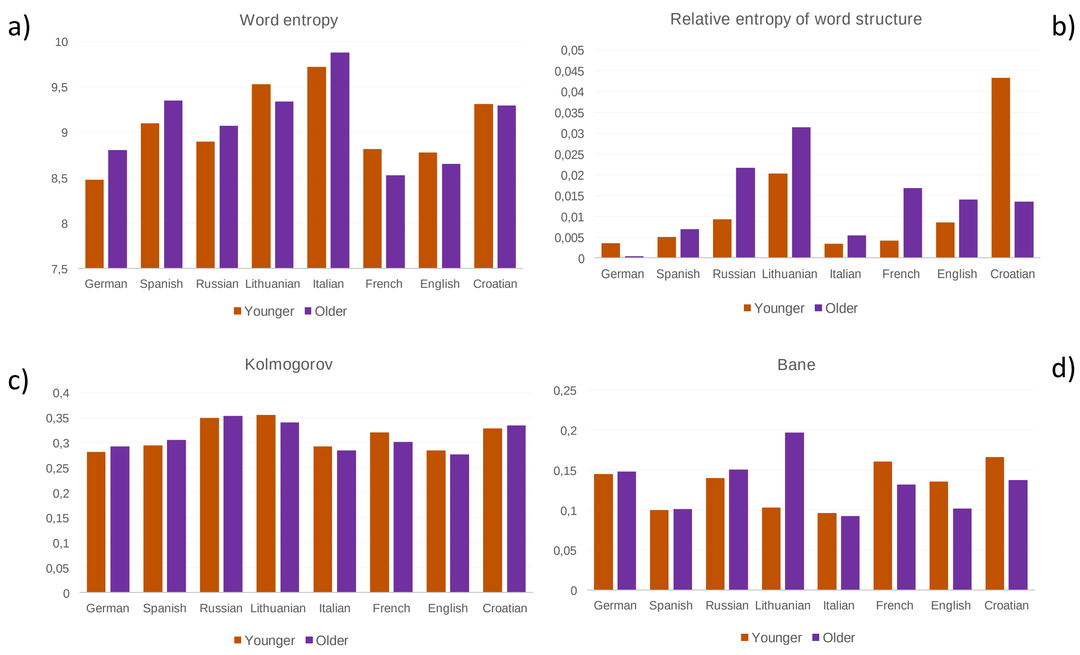
Table 2. Age ranges of children in two subcorpora for each of the eight languages

Language	Younger children	Older children
German	5;0 – 5;11	9;0 – 9;11
Spanish	5;0 – 5;11	8;0 – 9;11
Russian	5;0 – 7;11	9;0 – 9;11
Lithuanian	5;0 – 6;11	9;0 – 9;11
Italian	5;0 – 7;11	8;0 – 10;11
French	5;0 – 5;11	8;0 – 9;11
English	7;0 – 8;11	10;0 – 11;11
Croatian	7;0 – 8;11	10;0 – 11;11

Four measures of morphological complexity were calculated for each subcorpus: Word entropy, Relative entropy of word structure, Kolmogorov complexity and Bane measure. These measures evaluate the text samples according to: 1) quantitative complexity, i.e., the number of grammatical contrasts, markers or rules; 2) irregularity-based complexity, i.e., the number of irregular grammatical markers. Scores were compared between corpora of younger and older children.

4. Results

Figure 1. Results on four measures of morphological complexity obtained in corpora of younger and older children in eight languages: a) Word entropy, b) Relative entropy of word structure, c) Kolmogorov, and d) Bane measure.



5. Conclusion

Younger children corpora had lower morphological complexity than older children corpora on all four measures in Spanish and Russian. Reversed results were obtained for English and French, and the results for the remaining four languages showed variation. Relative entropy of word structure proved to be indicative of age differences. Word entropy and relative entropy of word structure showed potential to demonstrate typological differences.

References: **(1)** Bane, M. (2008). Quantifying and measuring morphological complexity. In *Proceedings of the 26th west coast conference on formal linguistics*, pages 69–76. Citeseer. **(2)** Bentz, C., Ruzsics, T., Koplenig, A., & Samardzic, T. (2016). A comparison between morphological complexity measures: Typological data vs. language corpora. In *CL4LC@COLING 2016*, pages 142–153. **(3)** Berman, R. & Slobin, D. I. (1994). *Relating events in narrative: A crosslinguistic developmental study*. Lawrence Erlbaum Associates. **(4)** Dehouck, M. (2019). *Multi-Lingual Dependency Parsing: Word Representation and Joint Training for Syntactic Analysis*. Ph.D. thesis, École Doctorale Sciences Pour L'Ingénieur. **(5)** Ehret, K. & Szmrecsanyi, B. (2016). *An information-theoretic approach to assess linguistic complexity*, pages 71–94. **(6)** Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213. **(7)** Kettunen, K. (2014). Can type–token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245. **(8)** MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs* (3rd ed.). Lawrence Erlbaum Associates Publishers. **(9)** Mayer, M. (1969). *Frog, where are you?* Dial Press.