Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 4729–4738
Marseille, 20-25 June 2022

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@gmail.com
[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

Abstract

The aim of this study was to compare the morphological complexity in a corpus representing the language production of younger and older children across different languages. The language samples were taken from the Frog Story subcorpus of the CHILDES corpora, which comprises oral narratives collected by various researchers between 1990 and 2005. We extracted narratives by typically developing, monolingual, middle-class children. Additionally, samples of Lithuanian language, collected according to the same principles, were added. The corpus comprises 249 narratives evenly distributed across eight languages: Croatian, English, French, German, Italian, Lithuanian, Russian and Spanish. Two subcorpora were formed for each language: a younger children corpus and an older children corpus. Four measures of morphological complexity were calculated for each subcorpus: Bane, Kolmogorov, Word entropy and Relative entropy of word structure. The results showed that younger children corpora had lower morphological complexity than older children corpora for all four measures for 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 show potential to demonstrate typological differences.

Keywords: language development, language sample analysis, morphological complexity, measurement

1. Linguistic complexity of languages

In general, complexity is the number and variety of elements and the elaborateness of their interrelational structure (Rescher, 1998, p. 1). Linguistic complexity or the complexity of a certain language has been the focus of linguistic interest for a long time, but until recently research into linguistic complexity was based on impressionistic views, personal intuitions, governing paradigms and fashions rather than being a serious, methodologically rigorous field of study. During the 19th century, linguistic complexity was viewed through the lens of supremacy differentiating between the more complex and culturally advanced (or superior) languages and the simple, primitive or underdeveloped languages (usually indigenous languages, pidgins, creoles etc.). With the rise of the idea of linguistic relativism (known as Whorfianism or Sapir-Whorf Hypothesis), which stated that language shapes and even determines the way its speakers think, this supremacist view was abandoned (more on this and the relationship between linguistic relativism and linguistic determinism in Despot Štrkalj (2021)). Despite the fact that this hypothesis has been heavily criticized and mostly rejected in its strong form, it is historically very important "as a reaction to the denigrating attitude toward unwritten languages that was fostered by the evolutionary view prevalent in anthropology in the 19th century" because it showed the so-called 'primitive languages' to be "as systematic and as logically rich as any European language" (Kay and Kempton, 1984, p. 65) leading to a generally accepted stance that all human languages are equally (and extremely) complex (the so-called equal complexity hypothesis). This influential view resulted in disagreement among linguists as to how to define objective complexity and in certain scepticism towards the idea that the notion of linguistic complexity can be defined or measured.

However, during the last ten years, a more systematic research of linguistic complexity has been undertaken, especially by finding ways of empirically measuring, operationalizing and approximating the complexity of its specific sub-parts: morphological, syntactic, morpho-syntactic, and typological complexity (Juola, 1998; Juola, 2008; Bane, 2008; Kettunen, 2009; Covington and McFall, 2010; Sinnemäki, 2011); for an overview of common methods of quantifying complexity in recent work, see Bane (2008).These quantitative approaches are based on the assumption that even if we adopt the equal complexity hypothesis, there should be some narrow range of complexity in which all human languages fall, and it is an important task to map the boundaries of that range according to some metric (Bane, 2008, p. 69). The new approaches to linguistic complexity have brought important methodolog-

ical advancement by clearly differentiating between: 1) complexity and difficulty, and 2) local and global complexity (for detailed explanations of these distinctions see Sinnemäki (2011)). It has been shown (Sinnemäki, 2011) that by focusing on particular types of complexity in their local contexts (local complexity), language complexity can be fruitfully measured, while studying global complexity is currently methodologically unattainable. Referring back to the equal complexity hypothesis, this means that different languages may vary as to the locus of complexity, for instance, one having complex morphology and another having many word order rules (Sinnemäki, 2011), which then balances out in typological comparison (Crystal, 2010, p. 6). According to Hawkins (2004, p. 9), complexity increases with the number of linguistic forms and the number of conventionally associated (syntactic and semantic) properties that are assigned to them when constructing syntactic and semantic representations for sentences (see Dressler (2011)). This corresponds to a building-block model of complexity (Zurek, 1990) in the sense of structural complexity (Dressler, 1999; Miestamo et al., 2008; Dressler, 2011). Morphological complexity of a language, simply put, refers to the richness of inflexion, i.e. a higher number of different nominal case forms, more structural units, and rules or representations indicate greater complexity (Kettunen, 2014). Inflectional morphology equips a language with the means to combine lexical and grammatical information: this typically refers to an inventory of forms and the contexts they are found in Baerman et al. (2017). The most important features of morphological complexity include: Case, Number, Person, Gender, Tense/Aspect/Mood (for the exhaustive list of the features of morphological complexity see Baerman et al. (2017)). As Dressler (1999) claims, in addition to the amount of morphological richness, complexity also includes all unproductive morphological patterns. This leads to a great difference between inflecting-fusional and strongly agglutinating languages (Dressler, 2011). Strongly inflecting-fusional languages have a sizeable amount of morphological richness, but also many unproductive patterns, i.e. additional morphological complexity. Strongly agglutinating languages have much more morphological richness, but ideally no unproductive morphological patterns, a situation nearly completely obtained by Turkish (Pöchtrager et al., 1998; Dressler et al., 2006; Dressler, 2011).

According to Anderson (2015), morphological complexity typology works along two dimensions: overall system complexity and complexity of exponence. Overall system complexity is understood as (i) the number of elements in the system (e.g. morphosyntactic categories), (ii) the number of elements within a word, and (iii) the principles of their combination (e.g. morphological templates vs syntactic ordering). Complexity of exponence is manifested in (i) the realization of individual elements (e.g. distributed and multiple exponence), (ii) inter-word relations (paradigmatic complexity), and (iii) allomorphy. The most wide-known classification of languages is based on how those languages form words by combining morphemes, and it differentiates between analytic (isolating) languages (very little inflection; word order and auxiliary words are used to convey meaning) and synthetic languages (agglutinating, fusional, and polysynthetic languages) (Greenberg, 1960). Agglutinative languages rely primarily on discrete morphemes (prefixes, suffixes, and infixes) for inflection, and the original root is easily extracted. Fusional languages allow one morpheme to contain several categories, and the original root can be difficult to extract. A subcategory of agglutinative languages are polysynthetic languages, which may construct entire sentences as one word. The properties that distinguish these types are gradient rather than categorical. The analytic – synthetic continuum depends on the number of morphemes per word (an isolating language having one morpheme per word and a synthetic language having many). The fusional – agglutinating continuum focuses on the extent to which there are clear boundaries between morphemes within a word (a fusional language lacks clear boundaries, while an agglutinating language has them) (Garland, 2005).
Despite the fact that morphological complexity itself is a complex phenomenon, as may be deducted even from this brief overview,in agreement with (Sinnemäki, 2011), we believe that (morphological) complexity can be fruitfully measured focusing on particular types of complexity in their local contexts.

2. Measuring morphological complexity computationally

With the development of machine learning, the need arose for a computational method to reliably compare morphological complexity between corpora without relying on linguistic knowledge. Consequently, several methods have been developed for measuring morphological complexity. There have been several attempts to use entropy to estimate the morphological complexity of languages (overview: Ehret and Szmrecsanyi (2016)). 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. Languages with a wider range of word types that are providing more information within the word structure (compared to phrase or sentence structure) score higher. While typically focused on the relationship between word frequency, predictability, and length of words, entropy is also associated with cognitive cost, as "entropy reduction identifies the extent to which a word reduces uncertainty about what is being communicated" (Venhuizen et al., 2019).
The second method is relative entropy of word structure which focuses on measuring the information content of the internal structure of words (Bentz et al., 2016). It

measures the difference between the character entropy in the corpus and a scrambled version of the corpus. The scrambled corpus is assembled by replacing words with random strings of characters with the same probability and length as the original word. This measures the information stored in the words via morphological regularities (Dehouck, 2019).

Another approach was proposed by Bane (2008). The author uses the software Linguistica, which calculates the morphological features of a language from a given text sample. Linguistica analyses an unannotated text corpus to separate word stems, affixes, and signatures, with signatures identifying the possible relationships or distributions of affixes to stems. An example of a morphological signature in English might be a pattern of formation of the Past Simple Tense, Ø, ed, associated with word stems, e.g., walk-, jump- (which also means that the words walk, walked, or jump jumped are present in the data). Languages with a simple morphology should have many stems but few affixes and signatures, and morphologically complex languages have more affixes and signatures than stems.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.

The fourth approach is based on Kolmogorov complexity. It 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 relies 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. Orthographic variations affect the compression of a text and thus increase its complexity. Consider the following examples in (1) and (2). The strings match in number of characters, but differ in complexity. After compression, the length of the string in (1) is 5 times *ab*, while the string in (2) counts 10 different characters. According to the logic of Kolmogorov complexity, the string in (2) is more complex than the string in (1).

1. abababababab
2. klmhgnadst

Kolmogorov complexity is not limited to specific linguistic features, but is a holistic means of measuring linguistic complexity. Although compression algorithms are insensitive to form-function mappings at the deep structure level, they capture repetitions and (ir)regularities at the surface level of text(Ehret and Szmrecsanyi, 2019). Kolmogorov complexity thus refers to quantitative complexity (the number of rules in a grammatical system) and irregularity-based complexity (the number of irregular forms in a grammatical system) (Ehret, 2018). There are a variety of different Kolmogorov-based metrics and methods depending on the phenomenon one wishes to study, morphological Kolmogorov complexity being one of them. Indeed, there is evidence that structural redundancy is closely related to morphological complexity (Juola, 1998; Ehret, 2014). Large amounts of structural redundancy based on (ir)regularities of word forms increase morphological complexity. Complexity can be interpreted linguistically (Ehret and Szmrecsanyi, 2019). The more words of a given language have word forms, the more complex the language is.

3. Complexity of a language of an individual speaker

The language production of an individual speaker can be observed and analyzed to gain an understanding of his or her language skills. In order to do so, language samples are taken, usually in a particular context, e.g., by presenting a person with a wordless picture book, asking a particular series of questions, or providing similar material. Thus, a language sample (a written or spoken text produced by a person, usually as a result of a language task such as telling a story or writing an essay) provides information about the acquisition of or proficiency in the first and second language, i.e., it can be used to assess the language of an individual speaker. Language sample analysis can be used by teachers of a second language, speech and language pathologists, elementary school teachers, employers in certain fields, etc. However, it is mainly used in the fields of first and second language acquisition, i.e., by speech and language pathologists and teachers of a second language. The same or similar measures are used in both domains, but for first language acquisition the language samples are usually spoken, while for second language acquisition they are usually written. This type of analysis is commonly used in some countries, but researchers and professionals from many countries are unaware of its benefits.

A number of measures have been introduced in various areas of language sample analysis (e.g., measures of productivity, measures of lexical diversity; for an overview of some measures see: MacWhinney (2000)). However, users often find transcribing and computing the measures time-consuming (Pavelko et al., 2016). In the last decades of the 20th century, computer programs were developed to support the analysis of speech samples (overview: Pezold et al. (2020)). Transcription, coding, and analysis are not user-friendly in these programs, so they tend to be used in the scientific community rather than by professionals.

Recently, new open-source web-based programs developed primarily within the scientific community have been introduced for various aspects of analysis. These web-based programs tend to focus on a specific area and are generally considerably more user-friendly than previously developed programs. For example, the Gramulator tool (McCarthy et al., 2012) calculates various measures of lexical diversity. Coh-Metrix (Graesser et al., 2004) is more elaborate and includes several domains, all of which are relevant to discourse analysis.

Such web services were developed for English and are mostly used for English, although there are some adaptations for other languages.

Traditionally, metrics in language sample analysis have been based primarily on basic calculations (e.g., type-token ratio, number of different words, mean length of utterance), especially for languages that are under-researched. There are some more advanced measurement methods based on language technologies, but they are only available for some languages. However, there is much room for progress here, especially considering that language technologies exist for a number of different languages. For example, a web-based language sample analysis application developed for a particular language and based on language technologies that are available such as open source could include, for instance, annotation of morphological and syntactic features, and recognition of connectives. Such annotation allows the implementation of metrics such as lemma-token ratio, lexical density (content words/number of words), clause density, or similar, which previously had to be computed by hand.

Although some tools rely on advanced computational methods for analyzing language samples, in general, tools and methods previously developed for more general language research can be adapted to measuring individual differences in the fields of first and second language acquisition using the method of language sample analysis.

Morphological complexity has rarely been investigated in studies of first language acquisition. Methods used in this area include observing the appearance of different morphemes based on the assumption that there is a universal acquisition order of morphemes (starting from Brown (1973); overview: Murakami and Alexopoulou (2015)), or counting the members of the same morphological paradigm (as proposed by Dressler (1999) with the notion of morphological pairs and mini paradigms). However, measures used to determine linguistic complexity in general, or morphological complexity in particular, can also be used as measures of individual performance. That is, measures of language morphological complexity can be used to reflect the richness of a language used by a speaker or writer. For instance, Kolmogorov's morphological complexity has been successfully used in a number of studies on the morphological complexity of typologically different languages (Juola, 1998; Kettunen et al., 2006). It has also been successfully applied to second language acquisition research (Ehret and Szmrecsanyi, 2019). However, the main question is whether such techniques (e.g., based on compression algorithms) are applicable to different types of data, including naturalistic corpus resources based on language samples.

There is evidence from second language acquisition (SLA) research that suggests that measures of morphological complexity may be suitable as a diagnostic tool for testing SLA language proficiency (Ehret and Szmrecsanyi, 2016; Ehret and Szmrecsanyi, 2019), although further research using larger data sets is needed to test the applicability of the measures. However, to the best of our knowledge, there are no studies using measures of morphological complexity to explore first language acquisition. Thus, this study is the first investigation using morphological complexity measures in the context of first language acquisition. Since it is known that language abilities generally increase with age, this factor should also be relevant to morphological complexity and therefore should be used to observe the applicability of using such measures in first language acquisition research and/or language assessment procedures.

4. Aims

In this paper, we aim to apply measures reflecting morphological complexity to language samples of children speaking different languages. We compare 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. We expect that the 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.

5. Methodology

5.1. Materials

The corpus used in this study is a selection from the Frog Story sub-corpus of the CHILDES corpora (MacWhinney, 2000), which has been used extensively in cross-linguistic work (Berman and Slobin, 1994b). It comprises oral narratives based on the Frog, where are you? 29-page wordless picture book by Mercer Mayer (1969), which have been collected by different researchers in the period 1990 - 2005. It tells the story of a boy and his dog who search for a missing pet frog and encounter different animals on their way. Since the story comprises a series of temporally sequenced events and requires inferencing of characters' relationships, it provides a rich context for language production. As such, it constitutes a convenient dataset for the purposes of this study focusing on morphological complexity across several languages.

For the purposes of cross-language comparison, the transcripts of oral narratives in seven languages were selected, i.e. Croatian, English, French, German, Italian, Russian, and Spanish. Additionally, samples of Lithuanian language, collected according to the same principles but not published on CHILDES, were added. The languages scrutinised in this study differ in their inflectional morphology and represent four different language families (Slavic, Germanic and Romance,Baltic); thus, they provide a sound basis for an analysis of cross-language variation in lexical diversity and morphological complexity.

The language-specific data were extracted from the following ten Frog Story corpora: the Croatian data from the Croatian Frog Story Corpus (Trtanj and Kuvač Kraljević, 2017; Hržica and Trtanj, 2021) (Trtanj Ivana, Kuvač Kraljević Jelena and Hržica Gordana, 2021), the English data from the English-MiamiMono Corpus (Pearson, 2002) (Pearson Barbara Zurer, 2021), the German data from the German-Bamberg Corpus (Berman and Slobin, 1994a) (Bamberg Michael, 2021), the French data from the French-Lyon Corpus (Harriet Jisa, 2021), the French-Duguine Frogs Corpus (Duguine Isabelle, 2021) and the French-MTLN Corpus (Le Normand Marie-Thérèse, 2021), the Italian data from the Italian-Bologna Corpus (Cipriani Paola and Orsolini Margherita, 2021) and the Italian-Roma Corpus (Orsolini Margherita and Pizzuto Elena, 2021), and the Spanish data from the Spanish-Sebastian Corpus (Sebastián, 1991; Sebastián and Slobin, 1994; Sebastián and Slobin, 1995) (Sebastián Gascón Eugenia, 2021), the Spanish-Aguilar Corpus (Aguilar, 2001; Aguilar, 2003; Aguilar, 2007; Aguilar, 2015) (Aguilar César Antonio, 2021) and the Spanish-Ornat Corpus (López Ornat and del Castillo, 1994) (López Ornat Susana, 2021). In addition, Lithuanian data (50 files) were collected according to the same methodology. Overall, the corpus includes (249) narratives. 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 presents the age ranges of children in different corpora, while Table 2 presents the size of corpora in words.

Language	Age range: corpus of younger children	Age range: corpus of 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

Table 1: Age ranges of children in two subcorpora for each of the eight languages

Since the present study explores the language performance of typically developing populations, the selection of data was restricted to oral narratives produced by middle class monolingual young speakers of Croatian, English, French, German, Italian, Russian, Spanish and Lithuanian. We selected children of diverse age-range within these languages in order to show that older children who gradually extend their linguistic skills (e.g., vocabulary range, complexity of syntactic structures, planning and organizing their oral production) might also show increase in measures of morphological complexity.

Language	Size of the corpus in words: corpus of younger children	Size of the corpus in words: corpus of 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: The size of the corpora in words

5.2. Procedure

The oral narratives were recorded in controlled conditions, typically at the schools the respondents attended. While the respondents in the Croatian, Russian, German, French, English and Italian corpora come from the same or different regions in the same country, the Spanish corpora represents two national standards, i.e. Spain and Mexican Spanish. The language and general development of the respondents are regarded as typical of their age and none of them has had a history of speech, language therapy or special needs educational support as reported by parents and school.

Each child was interviewed individually in a quiet room while sitting side by side with the investigator (who was the only listener) and was receiving the same instructions. An attempt was made to minimize the burden on the children's memory, and to make them aware that they were going to tell a story. The children were explicitly oriented to the booklet in the initial instructions: "Here is a book. This book tells a story about a boy [point to picture on cover], a dog [point], and a frog [point]. First, I want you to look at all the pictures. Pay attention to each picture that you see and afterwards you will tell the story." (Berman and Slobin, 1994b, p. 22).

The children were asked to look through the entire booklet first and then tell the story as they looked at the pictures and turned the pages at their own pace. In order to allow the children to narrate independently, the researchers were instructed to limit their verbal feedback to neutral comments that were not intended to influence the children's chosen form of expression. The prompts that followed were silence or nod of head or neutral prompts like, uh-huh, okay, Anything else?, and...?, Go on, etc.

All narratives were audio-recorded and then transcribed according to the coding system Codes for the Human Analysis of Transcripts (CHAT) in the program Child Language Analysis (CLAN), both of which are part of the TalkBank project (MacWhinney, 2000). All language samples successfully passed the CHECK option in the CLAN program. Language samples and audio recordings are publicly available at Child Language Data Exchange System (CHILDES) (MacWhinney, 2000). We have retrieved language samples from CHILDES site and prepared them for the analyses. All nontextual information and coding was removed, as well as punctuation and white spaces. When needed, orthography was uniformed.

We have applied four measures of morphological complexity to our comparable corpora: Word entrophy, Relative entropy of word structure, Kolmogorov complexity and Bane measure. These measures evaluate the text samples according to: 1) quantitative complexity: the number of grammatical contrasts, markers or rules; 2) irregularity-based complexity: the number of irregular grammatical markers.

Word entropy We computed the word entropy in the manner outlined by Bentz and Alikaniotis (2016), using the Shannon's (1948) measure:

$$H(Text) = -\sum_{i=1}^{V} p(w_i) log_2(p(w_i)) \quad (1)$$

where V is the vocabulary of word types and the word type probability $p(w_i)$ is computed using *James-Stein shrinkage* estimator (Hausser and Strimmer, 2009) as

$$\hat{p_{w_i}}^{shrink} = \lambda \hat{p_{w_i}}^{target} + (1-\lambda)\hat{p_{w_i}}^{ML} \quad (2)$$

where $\hat{p_{w_i}}^{ML}$ is the maximum likelihood word probability, $\lambda \in [0,1]$ ($\lambda = 0.7$ in our experiments) is the "shrinkage intensity", and $\hat{p_{w_i}}^{target}$ is the "shrinkage target", the maximum entropy case of a uniform $p_{w_i} = \frac{1}{V}$.

Relative entropy of word structure We calculated the per character entropy of the text using the Kontoyiannis et al. (1998) estimation measure:

$$\hat{H}(Text) = \Big[\frac{1}{n}\sum_{i=1}^{n}\frac{l_i}{log_2(i+1)}\Big]^{-1} \quad (3)$$

where n is the total number of characters in the text and l_i denotes the length of the largest non-repeating substring from position i forward.

To determine the degree of redundancy/predictability added by within-word structure, we used Koplenig et al. (2017) method and substituted each word token in the text with a token of the same length but composed entirely of letters randomly picked from the alphabet. The original text's entropy is then subtracted from the masked text to get:

$$\hat{D} = \hat{H}(Text^{masked} - \hat{H}(Text^{original}) \quad (4)$$

The greater the value of $\hat{D}$, the more information is kept inside words, such as in morphological regularities.

Kolmogorov We concatenated the entire corpus into a single string separated by spaces, transformed it to a byte string in Python, and then compressed the string using gzip. Finally, we computed the ratio of the compressed string's length to the original, which yields a complexity score between 0 and 1.

$$\frac{length(compressed)}{length(original)} \quad (5)$$

Bane We parsed the affixes, signatures, and stems from the corpus using Bane's research tool Linguistica (Lee and Goldsmith, 2016) and then computed the description lengths of those strings.

$$\frac{DL(Affixes) + DL(Signatures)}{DL(Affixes) + DL(Signatures) + DL(Stems)} \quad (6)$$

where $DL(x)$ is the description length of x.

6. Results and Discussion

6.1. Word entropy

Results of the word entropy measure show higher results for the older children narratives corpora in four out of eight languages: German, Spanish, Russian and Italian. For four languages higher results were obtained for the younger children narratives corpora: French, English, Lithuanian and Croatian. Figure 1 shows results of two corpora per language.

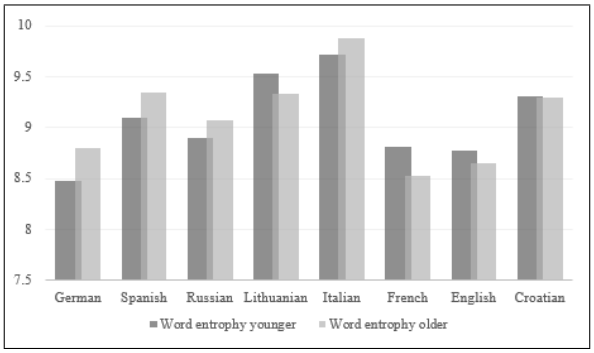

Figure 1: Word entropy results

Despite the fact that the results of the word entropy measure do not reveal clear age differences in morphological complexity, Figure 3 shows lower results for both groups of children in French and English. This confirms the Koplenig's et al. (2017) claim that languages that rely more strongly on word order information tend to rely less on word structure information and vice versa. It is known that analytic languages, as English, show lower level of word structure information than synthetic languages, and we see similar results for French which has many analytic tendencies.

6.2. Relative entropy of word structure

The results of the relative entropy of word structure measure show higher values for the older children narratives corpora in 6 out of 8 languages: Spanish, Russian, Lithuanian, French, English, and Italian. For two languages, German and Croatian, higher results were obtained for the younger children narratives corpora. Figure 2 shows the results of two corpora per language.

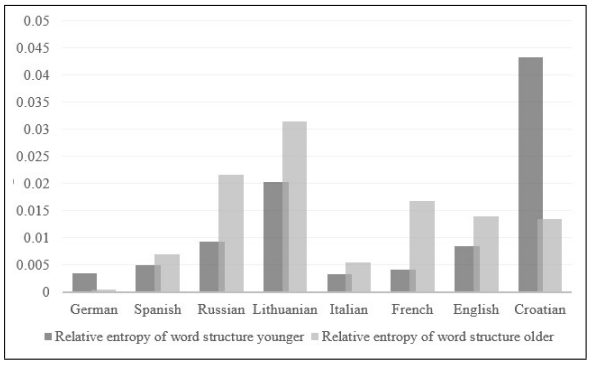

Figure 2: Relative entropy of word structure

The analysis of the relative entropy of word structure reveals differences in morphological complexity between corpora of older and younger children narratives in 6 of 8 languages, but it did not confirm our expectations for the German and Croatian subcorpora. From the typological point of view, it is interesting to note that higher average results were obtained for Croatian, Lithuanian and Russian language, considered to be morphological more complex, when compared to English, German, French and Spanish.

6.3. Kolmogorov

Results of the Kolmogorov measure show higher results for older groups of children in four out of eight languages: German, Spanish, Russian and Croatian. For four languages higher results were obtained for the corpora of younger children: Italian, French, English and Lithuanian. However, it is worth noting that the observed differences are very small. Figure 3 shows *that differences between two age groups are insignificant in all eight languages.*

The analysis of Kolmogorov complexity from a typological perspective reveals that higher results are typical for morphologically rich and highly inflected languages: Russian, Croatian, and Lithuanian. The results for English, German, Spanish, and Italian present the least morphological complexity (see Fig 3). It means that English, German, Spanish, and Italian narratives contain little word form variation. The results for French are between these two groups of languages.

6.4. Bane

Results of the Bane measure show higher results for older groups of children in four out of eight languages: German, Spanish, Russian and Lithuanian. However, for four languages higher results were obtained for the

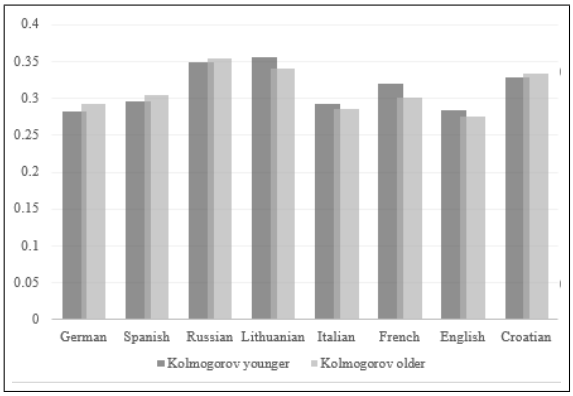

Figure 3: Kolmogorov results

corpora of younger children: Italian, French, English and Croatian. Figure 4 shows

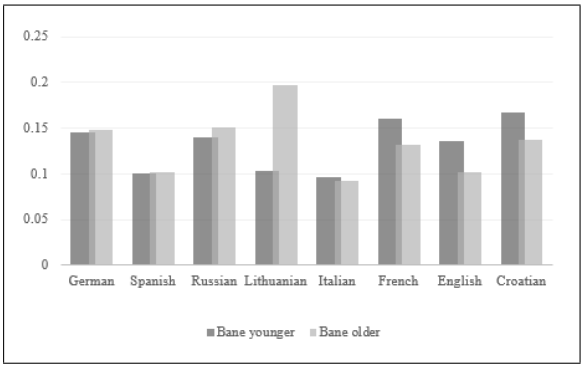

Figure 4: Bane results

The results of the Bane measure do not show clear trends in terms of language typology. It seems to reflect a similar level of morphological complexity between languages, which could be attributed to the nature of the task or the characteristics of the child language.

7. Discussion

The goal of this work was to apply measures that have been shown to reflect morphological complexity to language samples obtained from children speaking different languages. For each of the 8 languages, we formed two corpora, one with texts produced by younger children and one with texts produced by older children. We compared the scores for various measures of morphological complexity to see if they differed in corpora of younger and older children and to consider whether they reflect typological differences.

Our results show that in some cases younger children narrative corpora had lower morphological complexity scores than older children narratives corpora. For all four measures this was true for two languages: Spanish and Russian, and for three measures this was true also for German. However, three measures showed reversed results for two languages: English and French. The second measure, relative entropy of word structure,

proved to be the most convenient for showing differences between corpora of younger and older children. The observed differences in the results can be attributed to three main factors.

First of all, the measures we use differ considerably in terms of the reflected aspect of morphological complexity. For example, entropy measures do not distinguish between effects due to the breadth of the base lexicon and morphological processes such as derivation, inflection or compounding (Bentz et al., 2016), and thus do not discriminate between regular and irregular word formation processes. On the other hand, Kolmogorov complexity tends to favour morphological complexity, meaning that high structural redundancy can lead to lower complexity. Both entropy-based measures and Kolmogorov complexity are not based on linguistic features or do not contain information about morphological structure, so identifying what they measure is not as easy to determine. This is different for the Bane measure, which is based on an automated linguistic analysis prior to the measure calculation. However, this measure did not yield the expected pattern of typological differences (Slavic and Baltic languages tend to have higher scores) that was obtained by other measures. It should be noted, however, that the results of computational morphological complexity measures depend on the size of the corpora used for the study. For example, Bane's morphological complexity measure may not show the expected typological differences because it could not extract enough signatures from small corpora.

Second, there are a number of methodological reasons that may have led to the mixed results. Previous studies have mostly used the measures mentioned here in parallel corpora research. The results obtained could be due to the fact that the corpora were comparable rather than parallel. Since we worked with spoken language samples, the corpora under analysis are smaller compared to other studies (for example, Ehret (2016) suggests at least 10000 words of text for Kolmogorov complexity analysis).

Third, the results could be related to the specificity of our sample. In other words, they might simply reflect that language samples of children of different ages in different languages are characterised by a different kind of complexity. While older children use more diverse morphological forms in some languages, they might use more sophisticated syntactic structures than younger children in other languages, and this is not reflected in measures of morphological complexity. This could be confirmed by the fact that certain languages show the same patterns for the majority of measures (e.g., consistently higher scores for older or younger groups of children).

Higher morphological complexity could also be the result of the fact that some younger children produce more grammatical errors, neologisms, or overgeneralizations and this might be the reason for the higher values of certain morphological complexity measures. We should investigate these hypotheses in future studies.

Additional factors (e.g., individual differences in children's language abilities, the influence of grammatical errors on word form variation) may have masked the age difference. In future studies, it might be useful to focus on individual samples of children and calculate measures for each sample on the limited number of words or utterances (as has been suggested for other measures, such as mean length of utterances or type-token ratio). Future research should also include more information about an individual speaker's language (e.g., more general language measures such as mean length of utterance) and should attempt to examine the relationship between measures of morphological complexity and other measures of language development (e.g., vocabulary diversity, mean length of utterance). Finally, some measures could be used for different purposes in future studies. For example, the measure of Bane morphological complexity, which depends on the signatures extracted from the corpora, could be used to track the number of signatures in child language development.

To conclude, to the best of our knowledge, this study was the first to apply the Bane measure, the Kolmogorov measure, word entropy, and relative entropy of word structure to spoken language samples organized in small corpora of individual speakers. All things considered, the results seem to provide a good basis for future research into morphological complexity in children narratives across different languages.

8. Acknowledgements

This work has been supported by the project NexusLinguarum – European network for Web-centred linguistic data science (COST Action CA18209) and it has been supported in part by the Croatian Science Foundation under the project Multilevel approach to spoken discourse in language development (UIP-2017-05-6603).

9. Bibliographical References

Aguilar, C. A. (2001). Procesos de coherencia en una narración infantil a partir de relaciones entre acciones y personajes. In *Actas del VII Simposio Internacional de Comunicación Social*, pages 343–345.

Aguilar, C. A. (2003). Más allá del aquí y el ahora: el desarrollo de los marcadores temporales en el discurso narrativo en español. *Estudios de Lingüística Aplicada*, 22(38):33–43.

Aguilar, C. A., (2007). *¿Te enseño cómo te cuento un cuento? Habilidades narrativas y desarrollo lingüístico*, pages 51–66. UNAM.

Aguilar, C. A., (2015). *Había una vez un niño que tenía su sapo en un frasco. Relaciones entre referencia y cohesión en narraciones infantiles*, pages 283–309. El Colegio de México.

Anderson, S. R. (2015). Dimensions of morphological complexity. *Understanding and measuring morphological complexity*, pages 11–26.

Baerman, M., Brown, D., and Corbett, G. G. (2017). *Morphological complexity*, volume 153. Cambridge University Press.

Bane, M. (2008). Quantifying and measuring morphological complexity. In *Proceedings of the 26th west coast conference on formal linguistics*, pages 69–76. Citeseer.

Bentz, C. and Alikaniotis, D. (2016). The word entropy of natural languages. *arXiv preprint arXiv:1606.06996*.

Bentz, C., Ruzsics, T., Koplenig, A., and Samardzic, T. (2016). A comparison between morphological complexity measures: Typological data vs. language corpora. In *CL4LC@COLING 2016*, pages 142–153.

Berman, R. and Slobin, D. I., (1994a). *Different ways of relating events in narrative: Introduction to the study*, pages 1–16. Lawrence Erlbaum Associates. In collaboration with Ayhan Aksu, Michael Bamberg, Virginia Marchman, Tanya Renner, Eugenia Sebastian, and Christiane von Stutterheim.

Berman, R. and Slobin, D. I. (1994b). *Relating events in narrative: A crosslinguistic developmental study*. Lawrence Erlbaum Associates.

Brown, R. (1973). *A first language: The early stages*. George Allen & Unwin.

Covington, M. A. and McFall, J. D. (2010). Cutting the gordian knot: The moving-average type–token ratio (mattr). *Journal of quantitative linguistics*, 17(2):94–100.

Crystal, D. (2010). *The Cambridge encyclopedia of language*. Cambridge University Press Cambridge.

Dehouck, M. (2019). *Multi-Lingual Dependency Parsing : Word Representation and Joint Training for Syntactic Analysis*. Ph.D. thesis, École Doctorale Sciences Pour L'Ingénieur.

Despot Štrkalj, K. (2021). How language influences conceptualization: From whorfianism to neo-whorfianism. 45(4):373–380.

Dressler, W., Kilani-Schoch, M., Gagarina, N., Pestal, L., and Pöchtrager, M. (2006). On the typology of inflection class systems. *Folia Linguistica*, 40(1-2):51–74.

Dressler, W. (1999). Ricchezza e complessità morfologica. In *Fonologia e morfologia dell'Italiano e dei dialetti d'Italia. Atti del 21. congresso SLI. Roma: Bulzoni*, pages 587–597.

Dressler, W. (2011). The rise of complexity in inflectional morphology. *Poznań Studies in Contemporary Linguistics*, 47(2):159–176.

Ehret, K. and Szmrecsanyi, B., (2016). *An information-theoretic approach to assess linguistic complexity*, pages 71–94.

Ehret, K. and Szmrecsanyi, B. (2019). Compressing learner language: an information-theoretic measure of complexity in sla. *Second Language Research*, 35(1):23–45.

Ehret, K. (2014). Kolmogorov complexity of morphs and constructions in english. *Language Issues in Linguistic Technology*, 11(2):43–71.

Ehret, K. (2018). Kolmogorov complexity as a universal measure of language complexity. *Proceedings of the First Shared Task on Measuring Language Complexity*, pages 8–14.

Garland, J. (2005). Morphological typology and the complexity of nominal morphology in sinhala. *Santa Barbara Papers in Linguistics*, (17):1–19.

Graesser, A. C., McNamara, D. S., Louwerse, M. M., and Cai, Z. (2004). Coh-metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, 36(2):193–202.

Greenberg, J. H. (1960). A quantitative approach to the morphological typology of language. *International journal of American linguistics*, 26(3):178–194.

Hausser, J. and Strimmer, K. (2009). Entropy inference and the james-stein estimator, with application to nonlinear gene association networks. *Journal of Machine Learning Research*, 10(7).

Hawkins, J. A. (2004). *Efficiency and complexity in grammars*. Oxford University Press.

Hržica, G. and Trtanj, I. (2021). Mjere rječničke raznolikosti u pričama djece predškolske i rane školske dobi. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 47(1):8–22.

Juola, P. (1998). Measuring linguistic complexity: The morphological tier. *Journal of Quantitative Linguistics*, 5(3):206–213.

Juola, P. (2008). Assessing linguistic complexity. *Language complexity: Typology, contact, change*, pages 89–108.

Kay, P. and Kempton, W. (1984). What is the sapir-whorf hypothesis? *American anthropologist*, 86(1):65–79.

Kettunen, K., Sadeniemi, M., Lindh-Knuutila, T., and Honkela, T. (2006). Analysis of eu languages through text compression. In *FinTAL*.

Kettunen, K. (2009). Reductive and generative approaches to management of morphological variation of keywords in monolingual information retrieval: an overview. *Journal of Documentation*.

Kettunen, K. (2014). Can type–token ratio be used to show morphological complexity of languages? *Journal of Quantitative Linguistics*, 21(3):223–245.

Kontoyiannis, I., Algoet, P. H., Suhov, Y. M., and Wyner, A. J. (1998). Nonparametric entropy estimation for stationary processes and random fields, with applications to english text. *IEEE Transactions on Information Theory*, 44(3):1319–1327.

Koplenig, A., Meyer, P., Wolfer, S., and Müller-Spitzer, C. (2017). The statistical trade-off between word order and word structure–large-scale evidence for the principle of least effort. *PloS one*, 12(3):e0173614.

Lee, J. and Goldsmith, J. (2016). Linguistica 5: Unsupervised learning of linguistic structure. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 22–26.

López Ornat, S. and del Castillo, J., (1994). *De la adquisición narrativa en L1 a la adquisición narrativa en L2*, pages 155–161. Siglo XXI.

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk: Transcription format and programs (3rd ed.)*. Lawrence Erlbaum Associates Publishers.

Mayer, M. (1969). *Frog, where are you?* Dial Press.

McCarthy, P. M., Watanabe, S., and Lamkin, T. A. (2012). The gramulator: A tool to identify differential linguistic features of correlative text types. In *Applied Natural Language Processing: Identification, Investigation and Resolution*, pages 312–333. IGI Global.

Miestamo, M., Sinnemäki, K., and Karlsson, F. (2008). *Language complexity: Typology, contact, change*. John Benjamins.

Murakami, A. and Alexopoulou, T. (2015). L1 influence on the acquisition order of english grammatical morphemes: A learner corpus study. 38(3):365–401.

Pavelko, S. L., Owens, R. E., Ireland, M., and Hahs-Vaughn, D. L. (2016). Use of language sample analysis by school-based slps: Results of a nationwide survey. *Language, Speech, and Hearing Services in Schools*, 47(3):246–258.

Pearson, B. Z., (2002). *Narrative competence among monolingual and bilingual school children in Miami*, pages 135–174. Multilingual Matters.

Pezold, M. J., Imgrund, C. M., and Storkel, H. L. (2020). Using computer programs for language sample analysis. *Language, Speech, and Hearing Services in Schools*, 51(1):103–114.

Pöchtrager, M., Bodó, C., Dressler, W., and Schweiger, T. (1998). On some inflectional properties of the agglutinating type illustrated from finnish, hungarian and turkish inflection. *Wiener linguistische Gazette*, 62(63):57–92.

Rescher, N. (1998). *Complexity: A Philosophical Overview*. Transaction Publishers.

Sebastián, E. and Slobin, D. I., (1994). *Development of linguistic forms: German*, pages 239–284. Lawrence Erlbaum Associates.

Sebastián, E. and Slobin, D. I. (1995). Más allá del aquí y el ahora: el desarrollo de los marcadores temporales en el discurso narrativo en español. *Substratum*, (5):41–48.

Sebastián, E. (1991). El desarrollo del sistema de referencia temporal en español: Un paseo por la morfología verbal. *Anales de Psicología/Annals of Psychology*, 7(2):181–196.

Shannon, C. E. (1948). A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Sinnemäki, K. (2011). *Language universals and linguistic complexity: Three case studies in core argument marking*. Ph.D. thesis, University of Helsinki.

Trtanj, I. and Kuvač Kraljević, J. (2017). Language and speech characteristics of children's narratives: the analysis of microstructure. *Govor: časopis za fonetiku*, 34(1):53–69.

Venhuizen, N. J., Crocker, M. W., and Brouwer, H. (2019). Semantic entropy in language comprehension. *Entropy*, 21(12):1159.

Zurek, W. H. (1990). *Complexity, entropy, and the physics of information*. Addison Wesley.

10. Language Resource References

Aguilar César Antonio. (2021). *CHILDES Frogs Spanish Aguilar Corpus*. doi:10.21415/T5BW2W.

Bamberg Michael. (2021). *CHILDES Frogs German Bamberg Corpus*. doi:10.21415/T51S4T.

Cipriani Paola and Orsolini Margherita. (2021). *CHILDES Frogs Italian Bologna Corpus*. doi:10.21415/T5T01W.

Duguine Isabelle. (2021). *CHILDES Frogs French Duguine Corpus*. https://childes.talkbank.org/access/Frogs/French-Duguine.html.

Harriet Jisa. (2021). *CHILDES Frogs French Lyon Corpus*. doi: 10.21415/T5X30G.

Le Normand Marie-Thérèse. (2021). *CHILDES Frogs French MTLN Corpus*. doi:10.21415/T5H02G.

López Ornat Susana. (2021). *CHILDES Frogs Spanish López Ornat Corpus*. doi:10.21415/T5JW2R.

Orsolini Margherita and Pizzuto Elena. (2021). *CHILDES Frogs Italian Roma Corpus*. doi:10.21415/T5860M.

Pearson Barbara Zurer. (2021). *CHILDES Frogs English Miami Corpus*. doi:10.21415/T5T30J.

Sebastián Gascón Eugenia. (2021). *CHILDES Frogs Spanish Sebastián Gascón Corpus*. doi:10.21415/T5KS4S.

Trtanj Ivana, Kuvač Kraljević Jelena and Hržica Gordana. (2021). *CHILDES Frogs Croatian TKH Corpus*. doi:10.21415/8ARR-NH60.