Syntactic complexity of spontaneous spoken language of adult Croatian speakers

Abstract

The aim of this study was to explore syntactic complexity in the spontaneous language of 139 adult speakers of Croatian, as well as examine the interrelation between two syntactic complexity measures: length of the production unit, here measured as the mean length of C-units; and syntactic sophistication, here measured as the proportion of C-units that were relative clauses. The analysis included syntactic structures from the Croatian Adult Spoken Language Corpus (Kuvač Kraljević and Hržica 2016). Our results indicate a significant positive correlation between these two measures, confirming that speakers who produce longer utterances also produce less frequent and more complex syntactic structures.

Key words: syntactic complexity, length of the production unit, syntactic sophistication, Croatian Adult Spoken Language Corpus

1. Introduction

Linguistic complexity is a measure of the number of formal characteristics in language: the more characteristics, the greater the complexity. Different measures are used to assess complexity in absolute or relative terms (for an overview, see Kortmann and Szmrecsanyi 2012). Absolute complexity is defined quantitatively, e.g. as numbers of connections, clauses, or lexical items. Most theoretical linguists agree that all natural languages, because they play identical or similar functions in diverse language-based societies, must be equally complex overall, even though they may differ in the types and numbers of complex features. In other words, languages may differ in absolute numbers of each type of complex element, but all languages have approximately the same overall absolute complexity. In contrast to absolute complexity, relative complexity is defined from the perspective of the language user: it refers to the cost of processing, or difficulty of acquiring, a given language element or feature (Miestamo 2008). The more it costs to acquire or process a language element, the more complex that element is.

Complexity can be analysed globally across an entire language, or locally within a specific language domain, such as syntax. Measuring local or domain-specific language complexity,
such as syntactic complexity, is usually more feasible than determining global language complexity because the latter requires considering many elements in different subdomains (Miestamo 2008). For example, measuring absolute syntactic complexity means counting the number of syntax rules in a language, as well as determining how much clausal embedding and recursion the language permits (Kortmann and Szmrecsanyi 2012). Measuring relative syntactic complexity means evaluating the cost of processing, or difficulty of acquiring, a given syntactic element. Some syntactic elements are more complex than others, which can be verified in empirical studies. If a given element is empirically demonstrated to be more complex, its appropriate use by a speaker or writer can serve as a measure of language competence.

Following this approach, Hawkins and Buttery (2010) used the Cambridge Learner Corpus (CLC) to identify linguistic descriptors associated with the proficiency levels described in the Common European Framework of Reference for Languages (CEFR) (Council of Europe 2001), extending from A1 (Breakthrough) to C2 (Mastery). If an individual has mastered the linguistic descriptors associated with a certain proficiency level, then that individual can be considered to have that language proficiency. In other words, corpus-based linguistic descriptors can assess language knowledge.

Using this approach, researchers have developed many measures of syntactic complexity for assessing individual speakers. Such measures are useful for addressing various types of research question, as well as for performing normative comparisons across multiple samples, populations and contexts (Ortega 2003). These measures have already contributed extensively to research on acquisition of both a first language (for an overview, see Cheung and Kemper 1992) and a second language (e.g. Lu 2010). They have also proven useful in other research areas, such as first language assessment in adults (Cheung and Kemper 1992) and comparisons between populations showing typical or impaired language development (e.g. Thordardottir, Chapman and Wagner 2002).

1.1. Measures of syntactic complexity

Numerous features can be quantified when assessing syntactic complexity of an individual speaker: length of the production unit, amount of embedding, range of structural types and
degree of sophistication of certain structures distributed in production (for an overview, see Iwashita 2006). The simplest measure of syntactic complexity is the mean length of utterances, but this measure is insensitive to structural differences within sentences. More complex measures, such as Developmental Sentence Scoring (DSS) (Lee 1974), the Index of Productive Syntax (IPSyn) (Scarborough 1990), and the Developmental Level (D-Level) (Rosenberg and Abbeduto 1987), have been derived from various linguistic theories or models of syntactic processing, and empirically developed in studies on language acquisition and sentence processing (Covington et al. 2006). More recent studies have used corpus data to determine the frequencies of syntactic constructions, which then serve as a basis for complexity calculations in which lower frequency corresponds to greater complexity (e.g. Wiersma, Nerbonne and Lauttamus 2011). Measures that assess the frequency or processing cost of linguistic structures are typically referred to as measures of syntactic sophistication. All these measures are based on spoken or written English or on English language corpora. Even the most basic types of measures of the length of the production unit have been applied predominantly to English. These measures have been applied to Croatian, but exclusively in studies of child language (e.g. Kelić, Hržica and Kuvač Kraljević 2012).

The present study applied two measures of syntactic complexity to spontaneous spoken language of adult speakers of Croatian. One measure assessed the length of the production unit; the other, syntactic sophistication. Our research goal was to determine whether adult speakers who produce longer utterances also produce less frequent and more complex syntactic structures. These relationships have typically been studied in other populations (e.g. children), in other language modalities (e.g. written) and in other languages (e.g. English).

1.1.1. Measure of length of the production unit

Discourse analysis of spoken language requires identifying appropriate units into which the corpus can be divided (Crookes 1990). These units can then be analysed according to structural and functional systems developed to address different research objectives (Chaudron 1988). Crookes (1990) classifies basic units of analysis into T-units (Hunt 1965), its variations (e.g. communication units [C-units]; Loban 1976), turns, tone units and utterances. Tone units and utterances are defined by prosodic features; turns, by the communicational exchange between or among speakers. Texts are divided into T- and C-units based on syntactic features.
A T-unit is an independent clause and its modifiers. It cannot be further subdivided without losing its essential meaning. Main clauses can stand by themselves and can be segmented into one T-unit; therefore, a T-unit will consist of either a main clause or a main clause with its subordinating clause(s) (Miller, Andriacchi and Nockerts 2015). C-units are closely related to T-units, but they also include common utterances that consist of isolated phrases without a verb, such as answers to a question. Therefore, using C-units as the basic text unit allows the inclusion of crucial communicational elements while keeping syntax as the main criterion in text analysis. The mean length of C-units reflects the syntactic complexity of a text more accurately than the mean length of other possible text units because it takes into account not only the number of words but also the hierarchy of syntactic organisation in the text.

1.1.2. Measure of syntactic sophistication

Syntactic sophistication refers to the use of syntactic elements that occur with low frequency or incur a high processing cost. Empirical studies of frequencies of use and processing costs confirm that these criteria are reasonable indices of structural grammatical complexity (for an overview, see Bulté and Housen 2012). An individual’s use of such syntactic structures is considered an indicator of greater language proficiency.

Studies have analysed how the use of several constructions of different complexity varies with age or education. For example, Cheung and Kemper (1992) discovered that as adult speakers age, they embed sentences less often, and this decline is more noticeable for left-branching structures than for right-branching ones. Left-branching structures include subordinate clauses at the start of sentences, that clauses and wh- clauses that function as subjects, and relative clauses that modify the sentence subject. Right-branching structures include subordinate clauses that end sentences, verb phrase infinitive complements, and relative clauses that modify the sentence predicate. Biber and Gray (2011) showed that as writers develop, they use more complex noun phrases, while Kyle (2016) showed that the frequency of verb-constructions (VAC) influences acquisition of first and second languages. These findings laid the basis for automated measurement of syntactic sophistication (Kyle 2016; Lu 2010).
1.2. Interrelation between measures of length of the production unit and syntactic sophistication

Measures of length of the production unit are insensitive to structural differences within sentences because they depend on the number of words. This is true even if the production unit is defined based on syntactic criteria, since the measure does not take into account the internal structure of the unit. These measures are relatively fast and easy to calculate. In contrast, measures of syntactic sophistication are based on available theoretical and empirical data allowing for the selection of precisely defined elements (syntactic structures). This analysis is rather demanding, especially for languages other than English.

The current study aimed to examine the interrelation between two syntactic complexity measures: length of the production unit, as measured by the mean length of communication unit (MLCU); and syntactic sophistication, as measured by the ratio of relative clauses (RRC) in the total number of C-units. These measures were applied to spontaneous language spoken by adult speakers of Croatian in order to examine whether speakers who produce longer C-units also produce less frequent and more complex syntactic structures.

2. Methods

2.1. Corpus analysis

Language samples analysed in this paper were extracted from the Croatian Adult Spoken Language Corpus (HrAL; Kuvač Kraljević and Hržica 2016), which currently consists of 165 language transcripts and 617 participants. All language samples in HrAL are based on interactive spontaneous conversations lasting an average of 15 minutes between peers or family members. Data in the corpus were transcribed, coded and segmented using the Codes for Human Analysis of Transcripts (CHAT) transcription format and the Computerised Language Analysis (CLAN) suite of programmes, all within the Child Language Data Exchange System (CHILDES; MacWhinney 2000). Transcripts in the corpus are linked to the corresponding audio files.

The present study drew on corpus transcripts of 139 monolingual Croatian-speaking adults (101 women, 38 men) with ages ranging from 17 to 86 years (M=42.71; SD=17.86). The
education of these participants was distributed roughly evenly across five levels, from elementary and high school (40%) to undergraduate and postgraduate education (60%). Participants came from across Croatia, including the central counties, Dalmatia, Slavonia and Istria. To be included in the present study, participants had to produce at least 50 C-units per sample, since this is considered the minimal size of language sample appropriate for this type of research (Miller, Andriacchi and Nockerts 2015). Participants were selected based on this inclusion criterion using the CLAN programme MLU (Mean Length of Utterance).

2.2. Measures of syntactic complexity

2.2.1. MLCU

All language streams were segmented into C-units (Loban 1976), which include syntactic as well as communicational elements. The MLCU takes into account not only the number of words but also the hierarchy of syntactic organisation in the text. All transcripts had passed internal controls by the CHECK programme in the CLAN suite, and the segmentation of each transcript into C-units was checked by experienced researchers.

2.2.2. RRC

Relative clauses are embedded structures that emerge later in language acquisition (Byrnes and Sinicrope 2008; Diessel 2004). For the present study, relative clauses were defined as less frequent, more complex constructions. The number of relative clauses in the total number of C-units was calculated as an index of syntactic sophistication.

Comprehension and production of relative clauses is based on semantic and syntactic adjustment between the main clause and relative clause, which are connected by the relative pronouns who, what and which (Croatian koji, koja, koje). These pronouns replace the subject or object of the main clause. The adjustment between clauses is grammatically complex because of the several roles that pronouns can play, and because of semantics that depend on whether the relative clause is restricted or non-restricted (see Kordić 1995; Kuvač Kraljević et al. 2016). The grammatical complexity of relative clauses makes their processing more demanding than that of other complex structures, so mastering this syntactic structure is a long-term process (see Balija, Hržica and Kuvač Kraljević 2012). The subclass of relative
clauses relevant for this study is termed *attributive relative clauses* in the nomenclature of Croatian philology (e.g. Barić et al. 2005). All language samples in the present study were coded for production of attributive relative clauses.

3. Results and discussion

The aim of this study was to assess the level of syntactic complexity in spontaneous spoken language of adult Croatian speakers in terms of length and sophistication, and to examine the interrelation between the two complexity measures. Length was measured using MLCU and sophistication was measured using RRC. MLCU was calculated in an entirely automated fashion using the *CLAN* programme MLU, while RRC was calculated in a semi-automated fashion using the *CLAN* programmes KWAL and FREQ. Descriptive statistics and tests for normality and correlations were calculated in SPSS for Windows (version 22.0; IBM, Armonk, NY).

Table 1 shows descriptive statistics about the number of utterances, length of the production unit and syntactic sophistication of adult speakers of Croatian in informal contexts.

<table>
<thead>
<tr>
<th></th>
<th>Utterances</th>
<th>MLCU</th>
<th>RRC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Minimum</td>
<td>50</td>
<td>2.76</td>
<td>0.29</td>
</tr>
<tr>
<td>Maximum</td>
<td>340</td>
<td>7.38</td>
<td>3.48</td>
</tr>
<tr>
<td>Mean (M)</td>
<td>131.29</td>
<td>4.69</td>
<td>1.43</td>
</tr>
<tr>
<td>Standard deviation (SD)</td>
<td>66.31</td>
<td>0.88</td>
<td>0.88</td>
</tr>
<tr>
<td>Median (C)</td>
<td>111.00</td>
<td>4.55</td>
<td>1.25</td>
</tr>
<tr>
<td>Semi-interquartile range (Q)</td>
<td>47.00</td>
<td>0.56</td>
<td>0.58</td>
</tr>
<tr>
<td>Skewness (S)</td>
<td>0.98</td>
<td>0.67</td>
<td>0.97</td>
</tr>
</tbody>
</table>
Kurtosis (K) | 0.47 | 0.74 | 0.19
--- | --- | --- | ---
Kolmogorov-Smirnov Z (K-S) | 1.51* | 0.96 | 0.97
N | 139 | 139 | 139

*p<0.05; **p<0.01

Adult native speakers of Croatian language produced an average of 131 utterances (M=131.29; SD=66.31) in the informal, spontaneous contexts of the corpus. The distribution of data is skewed to the right (Figure 1), which was confirmed by the Kolmogorov-Smirnov Z test showing significant deviation from normality (Z=1.51; p<0.05). Thus, the median (C) may describe the results more reliably than mean (M) because it is insensitive to distribution asymmetry. The median of 111.00 utterances (Q=47.00) means that 50% of participants produced up to 111 utterances and 75% produced up to 171 utterances (Figure 1). In other words, speakers tended to produce fewer than 111 utterances.

Figure 1. Frequency distribution for the number of utterances
Next we examined the interrelation between the two measures of syntactic complexity. First we used the K-S test to show that MLCU followed a normal distribution ($Z=0.96; p>0.05$; Figure 2), as did RRC ($Z=0.97; p>0.05$; Figure 3) (Table 1). These results allowed us to structure subsequent analyses in the present study, and they provide potentially useful insights into how to use these less commonly used measures in future work.

**Figure 2.** Normal Q-Q Plot of MLCU data

**Figure 3.** Normal Q-Q Plot of RRC data
To measure syntactic sophistication, we divided the number of relative clauses by the total number of C-units to obtain the RRC, which depends on the number of relative clauses that are not frequently used. In the present study, 56% of participants did not produce relative clauses, giving an RRC=0. The extremely low syntactic sophistication for this group, as least insofar as relative clauses are concerned, led us to analyse them separately from the remaining 44% of participants with non-zero RRCs.

In order to examine whether MLCU differed significantly between participants who produced or did not produce relative clauses, we compared the results for the two groups using the independent-samples t test. C-units uttered by participants who produced relative clauses (M=4.85; SD=0.86) were significantly longer than those uttered by participants who did not produce such clauses (M=4.56; SD=0.81) (t=-2.01; df=137.00; p<0.05).

To examine the interrelation between the two measures of syntactic complexity, we performed Pearson’s correlation between them, first across all participants and then after excluding participants with RRC=0. The correlation between the two measures was stronger after excluding participants with RRC=0 (Pearson’s r = 0.46; p<0.01; Table 2). In other words, including all participants led to a significant but weak positive correlation between length of the production unit and syntactic sophistication; excluding participants with RRC=0 led to a significant and moderate positive correlation. These results provide a clear answer to our original research question: adult speakers of Croatian who produce longer production units are more likely to use less frequent and more complex syntactic structures.

**Table 2.** Pearson’s correlation between MLCU and RRC

<table>
<thead>
<tr>
<th></th>
<th>Pearson’s r</th>
</tr>
</thead>
<tbody>
<tr>
<td>MLCU * RRC (without RRC=0)</td>
<td>0.46**</td>
</tr>
<tr>
<td>MLCU * RRC (with RRC=0)</td>
<td>0.33**</td>
</tr>
<tr>
<td>Range of utterances</td>
<td>50 – 340</td>
</tr>
<tr>
<td>N</td>
<td>139</td>
</tr>
</tbody>
</table>
4. Conclusion

This corpus-based study provides some of the first insights into measures of syntactic complexity in spontaneous spoken language of adult speakers of a language other than English. The results indicate a significant positive correlation between MLCU and RRC regardless of whether the speaker produces relative clauses, confirming our hypothesis that speakers who produce longer utterances also produce less frequent and more complex syntactic structures. Our finding that MLCU and RRC are interrelated under these conditions of spontaneous spoken language may help researchers select the least costly approach to assessing syntactic complexity in individual speakers. This is especially useful when dealing with large amounts of data or time-limited procedures.

The present study applied only two measures of syntactic complexity: one measure of length (MLCU) and one measure of sophistication (RRC). Language production by individuals should be assessed using more measures of syntactic complexity, in order to gain reliable insights into syntactic complexity in a spontaneous conversation. More studies are also needed to describe syntactic complexity in different genres and modalities. In particular, measures of syntactic sophistication should take into account rare and complex structures other than relative clauses. Such future work may lead to more reliable methods for assessing an individual’s language performance, which may be useful for guiding research as well as clinical interventions.

Acknowledgments
This research was conducted within the “Adult language processing” project (2421) funded by the Croatian Science Foundation.
5. References


Kuvač Kraljević, Jelena and Hržica, Gordana. 2016. Hrvatski korpus govornog jezika odraslih ‘Croatian Adult Spoken Language Corpus’.


**Sintaktička složenost spontanog govornog jezika odraslih govornika hrvatskoga**

**Sažetak**

Cilj je ovoga rada ispitati sintaktičku složenost spontanog govornog jezika odraslih govornika hrvatskoga i utvrditi međupovezanost dviju mjera sintaktičke složenosti: duljine jedinice jezične proizvodnje i sintaktičke sofisticiranosti. Analizirane su sintaktičke strukture 139 odraslih govornika koje su preuzete iz Hrvatskog korpusa govornog jezika odraslih (Kuvač Kraljević i Hržica 2016). Kao mjera duljine uzeta je prosječna duljina C-jedinica, dok je mjera sintaktičke sofisticiranosti izračunata kao omjer odnosnih rečenica i ukupnog broja C-jedinica. Podatci pokazuju da su ove dvije mjere značajno i pozitivno povezane, što potvrđuje da govornici koji proizvode dulje iskaze ujedno proizvode i više rijetkih i složenih sintaktičkih struktura.

**Ključne riječi:** sintaktička složenost, duljina jedinice jezične proizvodnje, sintaktička sofisticiranost, Hrvatski korpus govornog jezika odraslih.