Towards sophisticated writing

Krista Kerge*, Hille Pajupuu**, Pilvi Alp*, Halliki Põlda*, Anne Uusen*
* Tallinn University, ** Institute of the Estonian Language

Abstract. In the framework of the Natural Models1 approach to learning Estonian (see Kerge, Uusen, Põlda 2014; Pajupuu et al. 2010), examples of highly educated non-philologist native adult speech and writing, rather than ideal edited standard language use, are considered the benchmarks to be striven for. To improve both teaching and assessing writing, the vocabulary parameters of creative writings of Estonian L1 students from grades 5, 7, 9, and 11 and the vocabulary parameters of writings of L2 writers with certified A2-, B1-, B2-, and C1-level proficiencies are compared to this benchmark. It is found that, although most adults must have more experience than young students, even at the C1-level, the L2 lexicon of L2 writers is strikingly poorer than the lexicon of native educated adults or even upper secondary students.

Keywords: vocabulary; L1, L2 acquisition, educated L1 use, adults’ L2, CEFR proficiency levels, natural writing, vocabulary measures

Lexical proficiency is an important factor in mastering writing because it makes it possible to freely cope with different types of

---

1 The Natural Models approach is worked out in ESF grant projects “Assessing and Modelling of Speaking Naturalness” (2006–2009 held by Hille Pajupuu) and “Modelling and Assessment of Writing Naturalness” (2011–2014 held by Krista Kerge) – the latter, ETF8605, has funded the study reported here. The basic idea of Natural Models is that one should not idealize standard language use, which is typical of edited text, but rather base testing standards on how typical, educated native people speak and write in more formal situations. This language use is defined as the Natural Model.
contents and texts. This is a skill usually characteristic of a highly educated user of a language – let us call it a benchmark to strive for. In reverse of the prevalent understanding of this benchmark as nuanced standard language use, hardly attainable for people other than language editors or other L1 professionals, we call ‘natural writing’ – or the Natural Writing model – the way educated non-philologist native Estonians tend to write demanding texts, such as essays (see Pajupuu et al. 2010).

This natural benchmark must be achieved in both L1 and L2 because the modern objectives of any language learning activity are the same: to fulfil a person’s individual need to take part in social life as an active citizen and to get equal opportunities to compete in the labour market. Our interest is how this benchmark can be attained if a society is multilingual, that is, for many people, Estonian, as an official national language, is a second language.

Thus, we investigated how the benchmark of educated writing is approached, first by native (L1) students in grades 5, 7, 9, and 11 (see Kerge et al. 2014) and second by adult Russians with a certified Estonian (L2) proficiency of CEFR levels A2, B1, B2, and C1 (see Alp et al. 2013). We assumed that progress in the development of writing skills is related to vocabulary range and diversity (Verspoor et al. 2012).

It is expected that progress in the development of writing skills is related to vocabulary range and diversity (Verspoor et al. 2012). In general, a learner first acquires a basic lexicon and, thereafter, the more rare part of a vocabulary (cf. Milton 2010; Milton, Alexiou 2009; Šišková 2012); this position is also held by CEFR (2001: 112). Step by step, an indispensable basic lexicon is enriched by more sophisticated (advanced) vocabulary, which enables the selection of words and sentence structures that are more suitable for a given text type or genre. As a result, texts become denser, richer in content, and more nuanced in their lexical and stylistic choices.
The more rare words (advanced types or tokens) are met in a piece of writing, the higher the profile of the writer (Daller, Xue 2007). In our study, we first defined the basic lexicon as the 4,000 most frequent words, covering about 70–80% of the public corpora (Kaalep, Muischnek 2002). From this basis we looked at different aspects of individual writers’ vocabularies.

*Lexical sophistication* of a text is a measure of advanced vocabulary, no matter whether repeated or not (Laufer 1995), specifically, the relative rate of advanced tokens beyond a basic lexicon.

*General lexical diversity* is a measure of different words (types) in a text. The parameter is concomitant to language proficiency, leading to a more precise wording of messages (Verspoor et al. 2012).

*Diversity of advanced words* is measured by counting the number of advanced types shared by the square root of the total number of tokens per text (Daller et al. 2003; Tidball, Treffers-Daller 2007). (For those parameters, see section 2.)

We did not compare the vocabulary of L1 students with that of non-native adults. We compared the advancement within both groups towards the benchmark of educated adult writing. However, we considered that adults, with their experience (linguistic and other, e.g. Langacker 2000), are faster language learners than young students (de Bastos Figueiredo, da Silva 2009).

1. Method

1.1 Material

The research material consisted of creative writings on topics customized to age or CEFR level:

144 compositions written by L1 Estonian students from grade 5 (age $M=11.5$ years, $SD=0.56$), grade 7 ($M=13.6$ years, $SD=0.57$), grade 9 ($M=15.3$ years, $SD=0.47$), and grade 11 ($M=17.3$ years, $SD=0.47$);
64 compositions written by L2 Estonian speakers at a certain certified level (age $M=36.3$, $SD=11.6$).

L1 materials (see Table 1) came from a relatively spontaneous argumentative writing experiment\(^2\) with a topic of social values, specially designed for grades 5 and 7 (45 minutes, 150 words) and grades 9 and 11 (60 min, 250 words) (see also Kerge et al. 2014).

L2 materials were derived from official, job-related Estonian examinations. For papers of each level, we analysed texts written on a predetermined topic in a prescribed amount of time: A2 level, a 30-word basic description; B1 level, a 100-word detailed description (both 20 min); B2 level, a 180-word verbal reasoning task (50 min); and C1 level, a 250-word publicistic article (60 min). The sample size of each level was limited to 16 papers that had passed the examination with at least a 70% score (see Table 1, see also Alp et al. 2013).

**1.2 Procedure**

For the analysis, we counted all the types and tokens of individual writings and compared them against the frequency dictionary of Estonian (Kaalep, Muischnek 2002), considering (1) the 4,000 most frequently used words as the *basic lexicon* and (2) all other words – excluding names, numbers and abbreviations – as *advanced*.

The relative rate of advanced tokens in a text was taken to be indicative of lexical sophistication (Laufer, Nation 1995; see Equation 1):

\[
LS = \frac{\text{advanced tokens} \times 100}{\text{total number of tokens}} \tag{1}
\]

---

\(^2\) For L1 students, the argumentative type was predefined via the task; the process of writing was relatively spontaneous, meaning that there was no time for editing or rewriting the text.
Table 1. Material: tokens (N) and types (V) per group

<table>
<thead>
<tr>
<th>Grade/level</th>
<th>Tokens/types</th>
<th>Min</th>
<th>Q1</th>
<th>Median</th>
<th>Q3</th>
<th>Max</th>
</tr>
</thead>
<tbody>
<tr>
<td>Estonian as L1 (by grade)</td>
<td>N</td>
<td>79</td>
<td>137</td>
<td>147</td>
<td>162</td>
<td>272</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>49</td>
<td>74</td>
<td>93</td>
<td>103</td>
<td>164</td>
</tr>
<tr>
<td></td>
<td>7</td>
<td>104</td>
<td>120</td>
<td>131</td>
<td>153</td>
<td>194</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>61</td>
<td>74</td>
<td>90</td>
<td>101</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>9</td>
<td>81</td>
<td>143</td>
<td>227</td>
<td>282</td>
<td>459</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>67</td>
<td>106</td>
<td>157</td>
<td>182</td>
<td>298</td>
</tr>
<tr>
<td></td>
<td>11</td>
<td>154</td>
<td>228</td>
<td>253</td>
<td>285</td>
<td>375</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>121</td>
<td>169</td>
<td>189</td>
<td>214</td>
<td>269</td>
</tr>
<tr>
<td>Estonian as L2 (by proficiency level)</td>
<td>A2</td>
<td>N</td>
<td>31</td>
<td>36</td>
<td>41</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>23</td>
<td>25</td>
<td>29</td>
<td>36</td>
<td>50</td>
</tr>
<tr>
<td></td>
<td>B1</td>
<td>N</td>
<td>81</td>
<td>102</td>
<td>122</td>
<td>130</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>48</td>
<td>59</td>
<td>71</td>
<td>76</td>
<td>93</td>
</tr>
<tr>
<td></td>
<td>B2</td>
<td>N</td>
<td>160</td>
<td>191</td>
<td>221</td>
<td>276</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>84</td>
<td>94</td>
<td>101</td>
<td>118</td>
<td>135</td>
</tr>
<tr>
<td></td>
<td>C1</td>
<td>N</td>
<td>236</td>
<td>259</td>
<td>270</td>
<td>336</td>
</tr>
<tr>
<td></td>
<td>V</td>
<td>110</td>
<td>133</td>
<td>138</td>
<td>158</td>
<td>180</td>
</tr>
</tbody>
</table>

General diversity was measured by means of Guiraud’s index (1954) (see Equation 2):

\[ G = \text{types} / \sqrt{\text{tokens}} \] (2)

The higher the index value, the more diversified the vocabulary.

The diversity of advanced words was measured by means of Advanced Guiraud (Daller et al. 2003; see Equation 3):

\[ AG = \frac{\text{advanced types}}{\sqrt{\text{tokens}}} \] (3)

The reference values (benchmark) have been obtained from social essays written by non-philologist employees (L1 Estonian) working in positions requiring higher education (Pajupuu et al. 2010).
A two-sample t-test was used to compare L1 students from grades 5, 7, 9, and 11 to the benchmark and L2 writers with certified A2-, B1-, B2-, and C1-level proficiency to the benchmark.

2. Results and Discussion

Based on the analysis of individual writings, the vocabulary range of the studied groups was characterized by the relative proportion of words from the basic vocabulary (up to 4,000 most frequent Estonian words) and advanced words (see Figure 1, Table 3).

![Figure 1](image.png)

**Figure 1.** Basic and advanced types across L2 proficiency levels A2, B1, B2, and C1; L1 student groups of grades 5, 7, 9, and 11; and educated native writers (the benchmark)

A two-sample t-test indicated that, in L1 student groups of grades 5, 7, 9, and 11, there were no statistically significant differences from the benchmark in the basic vocabulary range, and the 11th grade advanced vocabulary range showed no significant difference from
the benchmark either. All L2 proficiency levels showed no statistically significant difference from the benchmark in basic vocabulary, but even C1 level vocabulary did not reach the benchmark’s advanced vocabulary range. As for lexical sophistication, general diversity, and the diversity of advanced words, these parameters differentiated between age- or level-dependent proficiency groups more or less significantly both for L1 (see Kerge et al. 2014) and for L2 (see Alp et al. 2013). Though here, the progress towards the benchmark is mostly relevant.

The results demonstrated a gradual approach to the benchmark in the lexical skills of both L1 and L2 test groups (see Figures 2–4, Table 2 and Table 3).

Figure 2. Lexical sophistication by grade (L1) and by proficiency level (L2)

Figure 3. General lexical diversity by grade (L1) and by proficiency level (L2)

Figure 4. Diversity of advanced words by grade (L1) and by proficiency level (L2)
### Table 2. Indices of lexical sophistication (LS), general lexical diversity (G), diversity of advanced words (AG)

<table>
<thead>
<tr>
<th>Indices</th>
<th>LS</th>
<th>G</th>
<th>AG</th>
</tr>
</thead>
<tbody>
<tr>
<td>Benchmark</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Educated language user</td>
<td>16.7</td>
<td>10.1</td>
<td>2.2</td>
</tr>
<tr>
<td>Benchmark</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>L1 grades</td>
<td>5</td>
<td>7</td>
<td>9</td>
</tr>
<tr>
<td>M</td>
<td>4.8</td>
<td>5.0</td>
<td>10.5</td>
</tr>
<tr>
<td>SD</td>
<td>0.3</td>
<td>0.2</td>
<td>0.4</td>
</tr>
<tr>
<td>L2 levels</td>
<td>A2</td>
<td>B1</td>
<td>B2</td>
</tr>
<tr>
<td>M</td>
<td>6.5</td>
<td>8.8</td>
<td>7.4</td>
</tr>
<tr>
<td>SD</td>
<td>2.3</td>
<td>4.2</td>
<td>3.0</td>
</tr>
</tbody>
</table>

### Table 3. Two-sample t-test for vocabulary range, lexical sophistication (LS), general lexical diversity (G), diversity of advanced words (AG)

<table>
<thead>
<tr>
<th>Benchmark vocabulary range</th>
<th>5</th>
<th>7</th>
<th>9</th>
<th>11</th>
<th>A2</th>
<th>B1</th>
<th>B2</th>
<th>C1</th>
<th>4001-</th>
<th>.001</th>
<th>.001</th>
<th>.001</th>
<th>1.000</th>
<th>.001</th>
<th>.001</th>
<th>.001</th>
</tr>
</thead>
<tbody>
<tr>
<td>3001-4000</td>
<td>.406</td>
<td>.075</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.656</td>
<td>1.000</td>
<td>1.000</td>
<td>.556</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>2001-3000</td>
<td>.238</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.234</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1001-2000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td>.737</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>up to 1000</td>
<td>.001</td>
<td>.001</td>
<td>.002</td>
<td>1.000</td>
<td>.001</td>
<td>.048</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>LS</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.989</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>G</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.339</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AG</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.213</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td>.001</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note. p < .05 indicates that the groups differ statistically significantly from the benchmark (grey background)*
As for sophisticated vocabulary, there is quite a striking difference between the upper secondary students (L1, grade 11) and proficient L2 users (C1). For the L2 group, the AG and LS indices were far lower than the benchmark. Now the why-question arises.

Based on L1 language experience, which is inseparable from general experience (Langacker 2000), adult L2 learning should be faster than the learning of students, native or not. For example, de Bastos Figueiredo and da Silva (2009: 173) argue that adults, as ‘experts’, learn L2 more successfully than children because, from puberty onwards, a child’s competencies are surpassed by the speed with which an adult reaches L2 sensitivity. In our case, the C1-certified adult-writers may have had insufficient practice in authentic written Estonian usage, while the assessors of their writings were not – and could not be – capable of following their vocabulary parameters as objective measurement of this requires not only precise computations but also knowledge of the average usage frequency of words (Pajupuu et al. 2009: 188).

3. Conclusion

Before graduating from upper secondary school, the vocabulary of L1 Estonian speakers is sufficiently close to educated language use, enabling them to go to work or further their education.

However, L2 Estonian speakers lack sophisticated vocabulary. This may prevent them from competing on an equal footing for positions requiring higher education. To improve this, language teaching should focus on lexical diversity by approaching a wider variety of topics.

As for the research design, there might have been some influence from the L2 test tasks of the lower proficiency levels on the results of the study as the writing tasks of C1 and grades 9 and 11 were similar.
Acknowledgements
The study was supported by Projects ETF8605 and SF0050023s09 and the Alfred Kordelin Foundation.

References


Daller, Helmut; Roeland van Hout, Jeanine Treffers-Daller 2003. Lexical richness in the spontaneous speech of bilinguals. – Applied Linguistics, 24 (2), 197–222. doi:10.1093/applin/24.2.197


---


Pajupuu, Hille; Krista Kerge, Lya Meister, Eva Liina Asu, Pilvi Alp 2010. Natural speaking and how to assess it. – Trames: Journal of the Humanities and Social Sciences, 59 (2), 120–140. doi:10.3176/tr.2010.2.02

Teel arenenud kirjaoskuse poole

Krista Kerge*, Hille Pajupuu**, Pilvi Alp*, Halliki Põlda*, Anne Uusen*

* Tallinna Ülikool, ** Eesti Keele Instituut


Võtmesõnad: sõnavara, L1 ja L2 omandamine, haritud emakeelekasutus, täiskasvanute teine keel, CEFR keeleoskustasemed, loomulik kirjutamine, sõnavaramõõdikud

115