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 Models[1] 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[1]. 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=advanced tokens *100 / total number of tokens (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

	Grade/ level	Tokens/ types	Min	Q1	Median	Q3	Max
Estonian as L1 (by grade)	5	N	79	137	147	162	272
		V	49	74	93	103	164
	7	N	104	120	131	153	194
		V	61	74	90	101	130
	9	N	81	143	227	282	459
		V	67	106	157	182	298
	11	N	154	228	253	285	375
		V	121	169	189	214	269
Estonian as L2 (by proficiency level)	A2	N	31	36	41	50	85
		V	23	25	29	36	50
	B1	N	81	102	122	130	149
		V	48	59	71	76	93
	B2	N	160	191	221	276	301
		V	84	94	101	118	135
	C1	N	236	259	270	336	378
		V	110	133	138	158	180

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

$$G = \text{types} / \text{SQRTtokens} \quad (2)$$

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):

$$AG = \text{advanced types} / \text{SQRTtokens} \quad (3)$$

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).

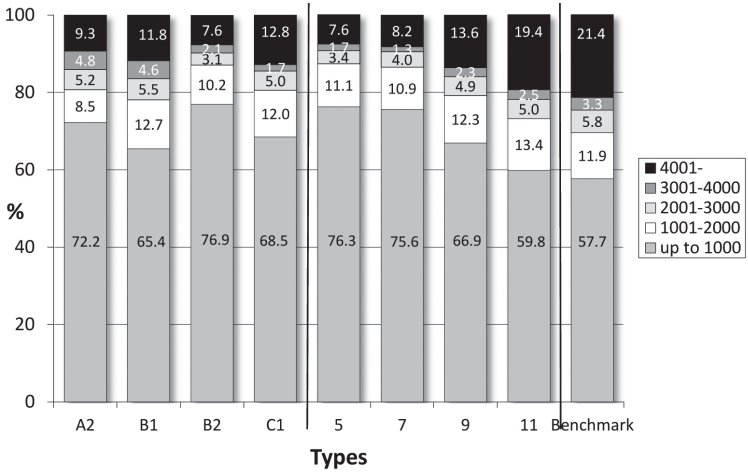

Figure 1. Basic and advanced types across L2 proficiency levels A2, B1, B2, and C1; L1 student groups of grades 5, 7, 9, and 11; and educated native writers (the benchmark)

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).

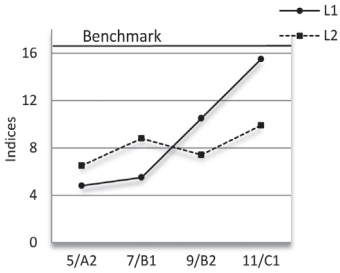

Figure 2. Lexical sophistication by grade (L1) and by proficiency level (L2)

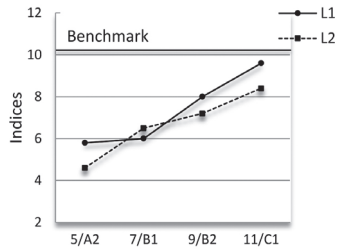

Figure 3. General lexical diversity by grade (L1) and by proficiency level (L2)

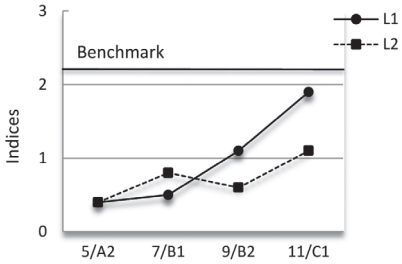

Figure 4. Diversity of advanced words by grade (L1) and by proficiency level (L2)

Table 2. Indices of lexical sophistication (LS), general lexical diversity (G), diversity of advanced words (AG)

Indices	LS				G				AG			
Benchmark	**Educated language user**											
M	**16.7**				**10.1**				**2.2**			
SD	3.9				0.8				0.6			
L1 grades	**5**	**7**	**9**	**11**	**5**	**7**	**9**	**11**	**5**	**7**	**9**	**11**
M	4.8	5.0	10.5	15.5	5.8	6.0	8.0	9.6	0.4	0.5	1.1	1.9
SD	0.3	0.2	0.4	0.5	0.8	0.8	1.1	0.8	0.2	0.2	0.4	0.4
L2 levels	**A2**	**B1**	**B2**	**C1**	**A2**	**B1**	**B2**	**C1**	**A2**	**B1**	**B2**	**C1**
M	6.5	8.8	7.4	9.9	4.9	6.5	7.2	8.5	0.4	0.8	0.7	1.2
SD	2.3	4.2	3.0	2.4	0.5	0.8	0.6	0.9	0.2	0.4	0.2	0.3

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

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

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

Alp, Pilvi; Krista Kerge, Hille Pajupuu 2013. Measuring lexical proficiency in L2 creative writing. – Jozef Colpaert, Mathea Simons, Ann Aerts, Margret Oberhofer (eds.), Language Testing in Europe: Time for a New Framework? Antwerpen: Linguapolis Universiteit Antwerpen, 274–286.

CEFR 2001 = Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Cambridge: CUP.

Daller, Helmut; Huijuan Xue 2007. Lexical richness and the oral proficiency of Chinese EFL students. – Helmut Daller, James Milton, Jeanine Treffers-Daller (eds.), Modelling and Assessing Vocabulary Knowledge. Cambridge: CUP, 150–164.

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

De Bastos Figueiredo, Sandra Andrade; Carlos Fernandes da Silva 2009. Cognitive differences in second language learners and the critical period effects. – L1–Educational Studies in Language and Literature[3], 9 (4), 157–178.

Guiraud, Pierre 1954. Les Caracteres Statistiques du Vocabulaire. Paris: Presses Universitaires de France.

Kaalep, Heiki-Jaan; Kadri Muischnek 2002. Eesti kirjakeele sagedussõnastik. [The Estonian Frequency Dictionary.] Tartu: Tartu Ülikooli kirjastus.

[3] The journals *L1–Educational Studies in Language and Literature* and *Estonian Papers in Applied Linguistics* (referred in this volume many times) have free open access; see http://www.l1research.org/ and http://www.rakenduslingvistika.ee/ajakirjad/index.php/aastaraamat (28.10.2014).

Kerge, Krista; Anne Uusen, Halliki Põlda 2014. Teismeea loovkirjutiste sõnavara ja selle hindamine / Teenage vocabulary and its assessment in creative writing. – Eesti Rakenduslingvistika Ühingu aastaraamat / Estonian Papers in Applied Linguistics, 10, 157–175. doi:10.5128/erya.v0i10.260

Laufer, Batia 1995. Beyond 2000: A measure of productive lexicon in a second language. – Lynn Eubank, Larry Selinker, Michael Sharwood Smith (eds.), The Current State of Interlanguage. Amsterdam: John Benjamins, 265–272.

Laufer, Batia; Paul Nation 1995. Vocabulary size and use: Lexical richness in L2 written production. – Applied Linguistics, 16 (3), 307–322.

Langacker, Ronald W. 2000. A dynamic usage-based model. – Michael Barlow, Suzanne Kemmer (eds.), Usage-Based Models of Language. Stanford: CA, CSLI Publications, 1–63.

Milton, James 2010. The development of vocabulary breadth across the CEFR levels. – Inge Bartning, Maisa Martin, Ineke Vedder (eds.), Communicative Proficiency and Language Development: Intersections between SLA and Language Testing Research. (=EUROSLA Monographs Series 1.) European Second Language Association, 211–232.

Milton, James; Thomaï Alexiou 2009. Vocabulary size and the Common European Framework of Reference for Language. – Brian Richards, Michael H. Daller, David D. Malvern, Paul Meara, James Milton, Jeanine Treffers-Daller (eds.), Vocabulary Studies in First and Second Language Acquisition. Basingstoke: Palgrave, 194–211.

Pajupuu, Hille; Krista Kerge, Pilvi Alp 2009. Sõnavara loomulik rikkus haritud keeleoskaja tekstides / Natural lexical richness in educated language use. – Eesti Rakenduslingvistika Ühingu aastaraamat / Estonian Papers in Applied Linguistics, 5, 187–196. doi:10.5128/ERYa5.12

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

Tidball, Francoise; Jeanine Treffers-Daller 2007. Exploring measures of vocabulary richness in semi-spontaneous French speech. – Helmut

Daller, James Milton, Jeanine Treffers-Daller (eds.), Modelling and Assessing Vocabulary Knowledge. Cambridge: CUP, 133–149.
Šišková, Zdislava 2012. Lexical richness in EFL students' narratives. – Language Studies Working Papers, 4, 26–36.
Verspoor, Marjolijn; Monika S. Schmid, Xiaoyan Xu 2012. A dynamic usage based perspective on L2 writing. – Journal of Second Language Writing, 21 (3), 239–263. doi:10.1016/j.jslw.2012.03.007

Teel arenenud kirjaoskuse poole

Krista Kerge*, Hille Pajupuu**,
Pilvi Alp*, Halliki Põlda*, Anne Uusen*
* Tallinna Ülikool, ** Eesti Keele Instituut

Loomulike eesti keele kasutusmudelite käsitluses (vt Kerge, Uusen, Põlda 2014, Pajupuu, Kerge, Meister, Asu, Alp 2010) ei peeta ideaaliks normitud keelekasutust, nagu seda valdavad peamiselt keeletoimetajad, vaid tavalist haritud standardkeelt, nagu seda asjalikes olukohastes žanrites kasutavad haritud mittefiloloogid – nimetagem sellist haritud keelt etaloniks, mille poole püüelda. Et parandada nii kirjutamise õpetamist kui ka kirjutusoskuse hindamist, on artiklis jälgitud, kuidas saavutavad haritud täiskasvanule omase sõnavara eesti (emakeelsed) õpilased ja eesti keelt omandavad muu emakeelega täiskasvanud. Õpilaste L1 sõnavara on uuritud 5., 7., 9. ja 11. klassi õpilaste katsekirjutistes, teise keele omandajate sõnavara aga tasemeeksamite sellistes kirjutistes, mis on hinnatud taotletava A2-, B1-, B2-, and C1-taseme väärilisteks. Kõigi rühmade kirjutiste üldise ja keeruka sõnavara ulatust ja variatiivsust on võrreldud haritlaskirjutiste korpuse näitajatega, kasutades kahe valimi t-testi.

Tulemused näitavad, et kirjaliku emakeele sõnavaras saavutatakse haritlastele statistiliselt lähedane tase hiljemalt 11. klassis, mis näitab noorte valmisolekut edasi õppida või töötada jõukohastes ametites. Teise keele kirjalik sõnavara on täiskasvanud eksaminandide suuremale üld- ja keelekogemusele vaatamata ka veel tasemel C1 statistiliselt olulisel määral vaesem kui haritud emakeelekõnelejal, ehkki selle taseme eksameid sooritavad kõrgharidust nõudvate ametite taotlejad. Selline olukord võib alandada teise keele omandajate konkurentsivõimet Eesti tööturul. Ilmne on vajadus tegelda keeleõppes mitmekesisema tekstivalikuga, mis tagab avarama ainekäsitluse toel ka suurema sõnavara.

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