LEXICAL RICHNESS IN L1 AND L2 ORAL AND WRITTEN USE OF ESTONIAN

KRISTA KERGE - HILLE PAJUPUU
Tallinn University - Institute of the Estonian Language

ABSTRACT

Lexical richness is a marker of natural language use and a characteristic of the vocabulary range. Using the Uber index, we measured the lexical richness of the vocabulary of L1 and L2 users with higher education in two oral and one written genre. Although L2 users had passed a C1 level exam, being fluent in speech and writing, the lexical richness of the two groups varied considerably: L1 vocabulary was richer in dialogues, monologues, and especially in essays. The comparison of the vocabulary of L1 and L2 speakers with the so-called elementary vocabulary (3000 most frequent words) showed that the share of the elementary vocabulary for L1 users was relatively large: 64% in speech and 56% in written texts as compared to the L2 users’ 53% vs 64.5%. Poorer vocabulary of L2 speakers and a different use as compared to L1 speakers is not an obstacle to being fluent in speech and writing.

Keywords: Naturalness of language use, L1, L2 acquisition, C1, genre, lexical richness, frequent words, Uber index, Estonian.
1. Introduction

Within the research project “Assessing and modelling of speaking naturalness”, we attempt to describe the language that native Estonians consider acceptable, i.e. natural. Natural language occurs as well structured text characterized by fluent, spontaneous, flexible, and effective language use without much obvious searching of expressions (cf. CEFR: 24). Naturality cannot be equated with standardized form which cannot be the basis for L2 assessment. (See also Ratcliff et al. 2002.)

We consider spontaneous speech and self-controlled writing of a non-linguist working in a position requiring higher education to be the rolemodel of L1 naturalness. We have described several aspects of natural L1 use and compared the results with these of L2 users with similar background: accent (Meister & Meister 2007), intonation (Asu forthcoming), pausing (Kerge et al. forthcoming) and contextuality-formality (Kerge et al. 2007) in different genres. Various aspects of natural language use reveal rather striking genre differences.

This study focuses on vocabulary. The CEFR describes two aspects of proficient user’s (C1) lexical competence: 1) vocabulary control: “Occasional minor slips, but no significant vocabulary errors”; 2) vocabulary range: “Has a good command of a broad lexical repertoire allowing gaps to be readily overcome with circumlocutions; little obvious searching for expressions or avoidance strategies. Good command of idiomatic expressions and colloquialisms.” (Op. cit.: 112) Both vocabulary use and range are markers of linguistic competence and fluent speech (Little 2005). By subjective assessment it is possible to assess fluency. However, assessing vocabulary range is much harder, especially in terms of lexical richness, looking at the adequacy of the treatment of each topic with a view to vocabulary. It also raises the question whether it is necessary to deal with this aspect of L2 vocabulary range separately when assessing fluency, communicativeness and adequacy of language use at C1-level exams.
We aim to describe lexical richness as a marker of natural speech and consider the importance of this criterion in the subjective assessment of L2 skills. Our research questions:

1. How rich is the actual vocabulary of proficient L1-users?
2. How to describe natural lexical richness?
3. How rich is the C1-level vocabulary of L2-users?
4. Does lexical richness vary in different text-forms and -types?
5. Should lexical richness be assessed separately at language exams?

2. BACKGROUND, MATERIAL AND METHOD

Estonia’s population is 1.3 million; 921,800 people use Estonian (the official language) as their mother tongue. Due to the official language requirement in public service, Estonian language skills are tested on three levels. Command of the official language at this level (C1) is obligatory mainly in demanding positions requiring higher education. Every year around 1,000 mainly Russian-speaking people take C1-level exams.

The text material for this study was gathered in standardized situation: for non-natives at C1-level Estonian exams; for natives in an exam-like situation (same examiner, same time-limit, same three tasks). The tasks were: essay (250 words), oral dialogue between two subjects (5–7 min), short oral presentation (1–2 min). The topics were linked to environment and society.

Subjects formed two well-comparable groups: 8 native Estonians and 8 native Russians (4 women, 4 men in both), all using fluent oral and written Estonian.

We measured lexical richness with the Uber index: $U = \frac{(\log N)^2}{(\log N - \log V)}$, where $N$ is the total number of word forms (tokens) and $V$ the number of different words (types). This formula is an algebraic transformation of TTR (type/token ratio), reducing somewhat the influence of text length on lexical richness assessment (see Vermeer 2000).

In assessing lexical richness, several authors have considered analysing the degree of difficulty of the words used, presuming that
more frequent words are more familiar (see e.g. discussion in Vermeer 2000: 79). Therefore, we found out the share of elementary vocabulary (3000 most frequent words) in texts produced by native and non-native Estonian-speakers’ group.

3. RESULTS

In our study we resorted to another vocabulary-related parameter: besides the lexical richness index, vocabulary can be measured by the formality index $F$ calculated on the basis of word type relations. The higher index $F$, the lower the share of context-specific vocabulary and the more formal and unambiguous the text. (Heylighen & Dewaele 2002.)

In our earlier research on the contextuality-formality of Estonian we compared the same L1-L2 material and reached two conclusions: (1) L1 and L2 users are not different in this dimension and (2) regardless of the subjects’ mother tongue, the most pronounced difference was found between text-types: oral dialogue is the most contextual and written essay the most formal type (see Figure 1), i.e. contextuality decreases and formality increases in the direction oral to written and dialogue to monologue. (Kerge et al. 2007.)

![Text types in contextuality/formality continuum](image)

Figure 1. L1 and L2 contextuality in different text-types. The higher F, the more formal the text.
Given the differences between text-types (genres in a narrower sense), we present measurements by text-types, indicating L1 and L2 user indices. (See Table 1)

<table>
<thead>
<tr>
<th>Text-types</th>
<th>types (V)</th>
<th>tokens (N)</th>
<th>Uber index (U)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>L1</td>
<td>L2</td>
<td>L1</td>
</tr>
<tr>
<td>Oral DIALOGUE</td>
<td>548</td>
<td>517</td>
<td>1752</td>
</tr>
<tr>
<td>Oral MONOLOGUE</td>
<td>477</td>
<td>402</td>
<td>1326</td>
</tr>
<tr>
<td>Oral speech (dialogue+monologue)</td>
<td>845</td>
<td>713</td>
<td>3078</td>
</tr>
<tr>
<td>Written ESSAY (MONOLOGUE)</td>
<td>736</td>
<td>666</td>
<td>1685</td>
</tr>
</tbody>
</table>

Table 1. L1 and L2 lexical richness in different text-types. The higher U, the richer the vocabulary.

Comparing the same groups in a column chart is visually more effective (see Figure 2).

The results show that text-types and forms of language use hold different lexical richness. An increasing lexical richness in the direction...
dialogue to monologue and oral production to written production can be considered natural. Both L1 and L2 users have the lowest lexical richness index in the most contextual text-type (dialogue) and the highest index in the most formal text-type (written essay).

L2 users still have considerably poorer vocabulary in every genre compared to L1 users; the difference is most striking in written essay.

Comparing the vocabulary of the two groups with Estonian elementary vocabulary gave quite a surprising discovery: the share of high-frequency words was 64% in L1 oral texts and 58% in L1 written texts. This means that the share of high-frequency words in the natural language use of educated speakers is unexpectedly high. L2 indicators are about the same, but in the reversed order: the share of high-frequency words in oral texts was 53% and in written texts 64.5% (see Figure 3).

![Degree of high-frequency words](image)

**Figure 3.** Share of high-frequency words by L1 and L2 use

4. DISCUSSION

As to lexical richness, which is a part of the natural language use of skilled communicators, the most important line seems to run between written and oral language use as well as between different text-forms (monologue, dialogue). This conclusion supports and enriches the variationist view on lexical richness (see Gijsel et al. forthcoming). Oral-written and monologue-dialogue are, first and foremost, parameters describing the genre (cf. Chafe, Tannen 1987: 385). It is natural that
Lexical richness is lowest in conversation, higher in presentation, and highest in essay. This agrees with, e.g., Chafe’s result according to which there is more repetition in oral text than in written text (see *op. cit.*).

L2 users have considerably poorer vocabulary than L1 users, irrespective of the genre. In interpreting this result, we need to remember that all L2 users had passed Estonian C1-level exam, which means that other aspects of language skills compensate for relatively poor vocabulary and we can still consider them skilled communicators. This questions the need for separate testing of vocabulary range due to the real outcome of the lexical richness parameter. Extensive testing of lexical richness would also disagree with CERF action and learner centred ideology which relates to the agents’ use of strategies (see CEFR: 9, 39, 15 etc.). Yet in Estonia, subjective assessment has been holistic – here we could consider adopting an analytical assessment scale to pay separate attention to the vocabulary-component. E.g. the extent of obvious compensation and avoidance strategies, the register and semantic adequacy of vocabulary, word position and functional properties of words, such as suitability and frequency, as referred to e.g. by Tuokko (2007: 101–103), but also characteristics facilitating understanding that are dependent on vocabulary (see *op. cit.*: 146) could be considered.

We are not sure how to interpret the fact that the share of elementary vocabulary in natural language use accounts to almost two thirds of words. As to Chafe and Tannen (1987: 385–86) in spoken language, the vocabulary is simpler; also parameters referring to high contextuality in Heylighen and Dewaele’s (2002) modern sense (like ratio of verbs, pronouns, adverbs) are more obvious by oral production. Our results agree with the conclusions referred to above, as the share of elementary vocabulary in oral texts is 6% higher than that in written texts. The results of our contextuality research also confirm those conclusions.

It is also difficult to interpret the reversed result of natural language use, according to which the share of less-frequent words in L2 written use is notably lower than in oral genres. According to CERF, production in C1-level can be influenced by a demanding discussion subject: „…. only a conceptually difficult subject can hinder a natural, smooth flow
of language“. However, even on C1-level avoidance strategies are considered common. (CEFR: 36.) Apparently proficient users are also afraid to make mistakes when producing a more demanding text in a foreign language, as writing reveals the real mastery of word. This may explain why a person uses more frequent words and avoids rarer ones in writing even if, according to the study, he does use the rarer ones in oral speech. Indirectly, the use of avoidance strategies is also confirmed by the fact that the Uber index of L2 users for written essay is especially low compared with L1 users – for oral production, it differs from that of native speakers significantly less (see Table 1).

5. CONCLUSION

The results add a new characteristic – lexical richness – to the model of natural language use. Natural language users (L1) resort to much richer vocabulary in written texts and monologue than in oral texts and dialogue. In written production, L1 use more less-frequent words than in oral production. The large share of elementary vocabulary both in L1 oral and written production is noteworthy. Although the lexical richness of L2 users follows the trends of L1 users by text-type, their vocabulary still remains considerably poorer than that of L1 users. This does not seem to curb fluent use of L2, however.

Although L2 and L1 texts contain a similar share of high-frequency words, L2 written texts still contain more such words than oral texts. This shows that even C1-level L2 users resort to avoidance strategies in more formal or demanding texts, i.e. use words that are completely familiar. Assessment criteria of L2 vocabulary should rely on and not be more demanding than the lexical richness parameters and share of elementary vocabulary characteristic of natural language use. The results should be taken into account in indirect testing of vocabulary range both in subjectively and objectively assessed tasks.

6. ACKNOWLEDGEMENTS

The study is supported by the Estonian Science Foundation Grant No 6742 and Alfred Kordelin Foundation Estonian Fund.
REFERENCES


Tuokko, E. 2007. Mille tasolle perusopetuksen Englannin opiskelussa