Baltic J. Modern Computing, Vol. 10 (2022), No. 3, pp. 382–391
https://doi.org/10.22364/bjmc.2022.10.3.11

How Masterly Are People at Playing with Their Vocabulary?

Matīss RIKTERS[1], Sanita REINSONE[2]

[1] University of Tartu, Estonia
[2] Institute of Literature, Folklore and Art, University of Latvia
matiss.rikters@ut.ee, sanita.reinsone@lulfmi.lv

Abstract. In this paper, we describe adaptation of a simple word guessing game that occupied the hearts and minds of people around the world. There are versions for all three Baltic countries and even several versions of each. We specifically pay attention to the Latvian version and look into how people form their guesses given any already uncovered hints. The paper analyses guess patterns, easy and difficult word characteristics, and player behaviour and response.

Keywords: linguistics, analysis, game, generation, Wordle, word game, word list

1. Introduction

Word guessing games are a phenomenon that represents the use of language in unusual socio-cultural contexts. Depending on the rules of a game, the meaning of a word can be completely irrelevant, whereas the structural elements of a word, such as its length and the composition of its letters, may play an important role. Despite words being used as game attributes without the need to know their meaning and context of use, it can be argued that such games contribute to vocabulary mastery and general language training.

One of the first computer games created in Latvia in the mid-1990s was the word scoring game Lingo. Inspired by a popular TV show, it was created by the language technology company Tilde. The game required guessing a five-letter word, the first letter of which was known to a player. As one of the few games available on almost all computers in Latvia at the time, it became very popular among players of all ages. The game's word corpus contained 999 words (Čudare, 2021). The game was required to be installed on a computer and could be played offline without restriction or any limit.

Almost thirty years later, in October 2021, an online word-puzzle game Wordle[3] was invented by a software engineer Josh Wardle. In a relatively short time, it became

[3] https://www.nytimes.com/games/wordle/index.html

globally popular, attracting more and more new players and spawning new language versions around the world. The principle of the game is fairly simple and similar to Lingo – a player is given six attempts to guess a five-letter word. After each guess, the letters are coloured in three colours (grey, orange, green), giving the player a hint on how to continue guessing the hidden word. Unlike other word games, the specifics of Wordle are that only one word can be guessed per day, which is the same for all players. For it to work, players must be disciplined not to reveal the word of the day to others.

However, to share results instantly without spoiling the enjoyment of the game for others, Wordle offers to create an abstract figure made up of emoji library squares in three colours. It contains a geometric pattern that shows the progress and result of a guess without revealing the word behind it. This figure that players share on social networks, is the most important representational attribute of the game, also acting in a symbolic way as a communication and interaction element within the community.

Although the original Wordle is in English, enthusiast developed open-source code behind the game enables it to be adapted for other languages. GitHub hosted ’Wordles of the World’[4] list contains links to Wordle games in more than 90 languages. For example, at least three versions of Wordle game have been currently developed for each – Estonian, Latvian and Lithuanian:

- Estonian versions:
  - `https://uudis.net/wordle`
  - `https://sonuk.subscribe.ee`
  - `https://sonar.ajad.ee`
- Latvian versions:
  - `https://wordle.lielakeda.lv`
  - `https://ralfulis.vercel.app`
  - `https://vardulis.lv`
- Lithuanian versions:
  - `https://jakut.is/vordl`
  - `https://dienos-zodis.lt`
  - `https://wordle.dario.cat`

While the Wordle developer admitted that the game is most appreciated precisely because of the fun it brings (Victor, 2022), Wordle users and re-designers have managed to add additional value to the game showing potential for promoting learning – it is being used in education for new language acquisition (Brown, 2022; Vincent, 2022), as well as to revitalise endangered languages (Schenck, 2022; CBC-News, 2022).

2. Game Construction

Shortly after the swift rise in popularity of the original Wordle game several versions of its reconstruction started popping up on GitHub. Of these the most popular became React Wordle [5], which so far has over 1,700 forks and over 2,200 stars, and has been used as a base to create Wordle versions in 43 different languages (even Latin and Cornish),

[4] `https://rwmpelstilzchen.gitlab.io/wordles`

[5] `https://github.com/cwackerfuss/react-wordle`

32 thematic versions (such as birds, super heroes, airport codes), and even 20 mathematics, science, technology oriented ones (for example, gene symbols, JavaScript, prime numbers). The base code, which was made using React, Typescript, and Tailwind libraries, has been developed for easy adaption to new languages or themes. For example, to have a personal list of daily words and valid guesses only two files need to be updated, and to adapt the code to a new language 7 to 9 other files need changes, for which detailed instructions have are provided in the GitHub repository.

2.1. Adapting Wordle into Latvian and Audience Involvement

The first Latvian version of Wordle was created in mid January 2022. The game was named 'Vārdulis' – deriving its title from Latvian 'vārds' (word), but keeping a sonic resemblance to original title. Giving the game a unique, Latvian-specific name was a successful choice, as it was easy to find Vārdulis mentions on social media from day one of the game's launch. This, in turn, is essential for communication between players.

Even though Wordle is meant to be played in a single-player manor, an essential part of the game is sharing the result, i.e. game's auto-generated grids of emoji squares, and discussing the word of the day without revealing it on social media, such as Twitter[6], or internal communication tools, such as WhatsApp. The impact of social media on the popularity of the game is significant. Sharing a score is often a conversation starter with other, previously unknown players, it is also a micro-competition to compare who has the better score and more successful choice of words. The words of the day are discussed and evaluated mostly in terms of their game-specific difficulty.

When developing the Latvian version, the decision was made to also include person names and various inflections of the words instead of plain singular nominative forms of nouns, incorporating words that have four letters in the nominative case, but five when conjugated (e.g., flower: nom. 'puķe' – gen. 'puķes'), thus making Vārdulis much more of a challenge than its English counterpart. However, a decision to include such words was reached in order to highlight the diversity of the language and have a more abundant set of data for subsequent play analysis. In addition, to include the possibility to learn more about the meaning of words, a link to the word entry in the online dictionary and thesaurus[7] developed by the Institute of Mathematics and Informatics of the University of Latvia was included in the window that pops up when the game is finished.

In the public discussions on Twitter which is the most common public space for Vārdulis players to meet, it can be seen that the most topical issue regarding Vārdulis is the extended dictionary, i.e. the inclusion of inflected forms in the game. The criticism was particularly strong in the first months of the game. Players complained that the game's rules thus are not fair and that they should stick to the rules of the original version, that the Latvian version is too complicated, that there are too many conjugations in Latvian to win the game in six attempts. It is also joked that the title of the game should rather be "guess the correct conjugation".[8] Over time, the criticism decreased, players accepted the rules and the vocabulary used by players increased.

[6] https://twitter.com/search?q=vardulis

[7] https://tezaurs.lv

[8] https://twitter.com/DavisVilums/status/1489537145836609537

Vārdulis, just like its original Wordle is limited to one game per day. The average game sessions per day from January 28 to April 14, 2022 is 935, however it took around 2 weeks for popularity to rise from a few hundred plays to around a thousand per day.

By exploring the user statistics of tezaurs.lv in Google Analytics,[9] it can be seen that the daily word is one of the most frequent searches in the database on a given day. On average, 5.7% of players navigate to the thesaurus to explore a particular word.

Exploring which words are most frequently consulted in the tezaurs.lv, two tendencies can be observed: first, less known or unusual word, for example, ‘adobe’ (meaning air-dried clay brick) that many of the players have never heard in Latvian was searched for on tezaurs.lv by 61.97% of players. Secondly, words that were difficult to guess or that a large number of players failed to guess at all. For example, 40.81% of players failed to guess quite common word ‘šuves’ (stitches), accordingly, on the given day, 21.14% of players searched for this word on the tezaurs.lv.

Overall, it can be concluded that the linking of tezaurs.lv with Vārdulis is successful and serves its purpose well, but it could also be used in a more targeted way by regularly including less known and used words in the list of daily words, which would provide additional opportunities for mastering vocabulary of players. However, as the game is to some extent competitive and players aim to complete the game in as few attempts as possible, players’ frustration and public complaining could be expected.

2.2. Word List Generation

There are two word lists necessary to play the game – a list of daily guesses (main list) and a list of all valid guesses (secondary list). Construction of both lists was performed semi-automatically. First, we acquired all monolingual Latvian corpora from Opus (Tiedemann, 2012), tokenised the data, filtered out only tokens consisting of 5 characters, and finally removed any tokens which had any character outside the 33 character Latvian alphabet. To make the game reasonably challenging, we ordered the remaining tokens by frequency of occurrence in the corpora and chose the 1,500 most frequent words for the main list and everything else for the secondary list.

To maintain purely words in the Latvian language, we cross-referenced the list with the Lexical Database for Latvian (Spektors et al., 2016) and manually reviewed each word. After this, 1,430 words remained in the main list while some very frequent foreign words such as “China” or “Apple” were removed. We selected the further 15,000 words from the list ordered by frequency for the secondary list, also cross-referencing with the Lexical Database, but without manually verifying.

The secondary list, however, was still at times falling short of it’s objective by failing to recognise perfectly valid Latvian words in specific inflections which may not have necessarily been among the 16,500 most frequent five-character words in the corpora. To improve the list, we once again turned to the Lexical Database and selected all words in lengths of 3 to 8 characters, automatically inflected them to all possible word forms using an inflection generator (Ņikiforovs, 2011), and filtered the results down to inflections of the words spanning exactly 5 characters. While still not fully exhausted, the secondary list grew to 22,341 words.

[9] For the purposes of the research, the Institute of Mathematics and Informatics of the University of Latvia granted access to the Google Analytics account of tezaurs.lv.

Table 1. Top 15 guesses at each turn. Words that were the actual answers within these days are marked in bold. English translations of the words can be found in Appendix B.

G1	$\sum$	**G2**	$\sum$	**G3**	$\sum$	**G4**	$\sum$	**G5**	$\sum$	**G6**	$\sum$
SAULE	5341	**LAIKS**	412	**LAIKS**	433	**TĒRPU**	432	**TĪRĪT**	382	**FLĪŽU**	345
SIENA	3179	SAULE	355	**TIESA**	334	**TĪRĪT**	382	**DARĀT**	364	**RAIŅA**	296
TIESA	1579	DIENA	337	**TAUKI**	295	**PUSEI**	382	**ILGĀK**	359	**BAUDU**	295
DIENA	1476	LIEPA	290	**LIETU**	273	**VĒLĀK**	380	**GROZĀ**	350	**MAIGI**	289
LAIME	1449	KAĶIS	284	**PUSEI**	271	**GARĀM**	364	**SAVĀM**	340	**BIEŽA**	278
PIENS	1237	PIENS	266	**LAIKU**	247	**LAIKS**	354	**TĒRPU**	337	**CELTA**	258
MAIZE	1217	ĀBOLS	262	**PRECE**	230	**KURSĀ**	353	**VĒRTS**	334	**SAVĀM**	250
LIEPA	1159	SAULĒ	207	**DIEVS**	225	**KRĀSU**	340	**ZEMĒM**	327	**JĀŅEM**	245
SAITE	958	SIENA	205	**LIKTS**	224	**LIKTS**	339	**IELEJ**	324	**PLAŠA**	243
KASTE	952	LIETA	204	**PUSES**	214	**PRECE**	338	**GARĀM**	322	**LABAS**	238
ĀBOLS	950	MAIZE	203	**TIRGU**	214	**GALDU**	335	**LABAS**	321	**ZEMĒM**	237
KAĶIS	869	RIEPA	192	**TĒRPU**	206	**DIEVS**	332	**VĒLĀK**	321	**ZINOT**	236
IELAS	676	LAIME	188	**MIERU**	201	**BLOKU**	331	**TĀPAT**	317	**IELEJ**	234
SKOLA	673	**TAUKI**	175	**REIZĒ**	200	**TIRGU**	328	**JĀŅEM**	317	**KAKLU**	226

3. Play Analysis

The design of our version of the game includes logging the array of guesses for each session played until the end (either correct guess or failed after six attempts). In this section we analyse game data of 77 daily words collected between January 28th and April 14th of 2022.

Table 2 shows the top 10 most difficult words to guess ordered by the amount of plays where the player was unable to guess the word after six guesses, and top 10 easiest words ordered where only very few players were unable to find the correct word while most were successful after the third or fourth guess. Here it is visible that a good deal of the easy words are nouns in singular nominative form, most of them do not contain diacritics, and have almost no repetition of characters within the word. On the other hand, most of the difficult words contain at least one or two diacritics, have repeating characters within the word, and none of the words are in singular nominative nouns.

The total number of tokens used by Vārdulis players in 77 days is 12,705. As it can be seen in Figure 1, the vocabulary used by players tends to expand. Table 1 shows the most popular word choices at each stage of the game. All words in columns G3-G6 have been the correct word of the day at some point. From the opening guess column G1 we clearly see that most players start with a singular nominative noun without diacritics, and with no overlapping characters within the word to make use of uncovering hints for future guesses. An interesting observation in Table 1 is that the most popular opening word by far is "Saule" (the Sun), followed by "siena" (wall), and "tiesa" (court or truth).

We look in detail at the most challenging word so far in the game and depict most common guess paths taken by players in Figure 2. The different arrows show at which of the six attempts to guess players were at. It is visible here that the vast majority of guesses at the last stages had already uncovered the ending of the correct answer "AS", and some had other critical characters uncovered like "Ī", "C" or "Ņ".

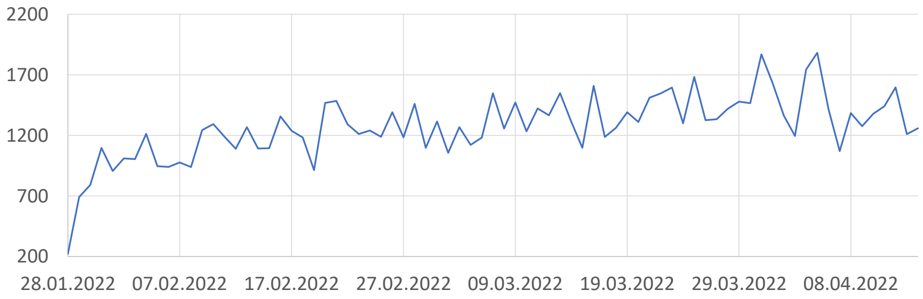


Fig. 1. Change of unique word forms used for guessing over time.

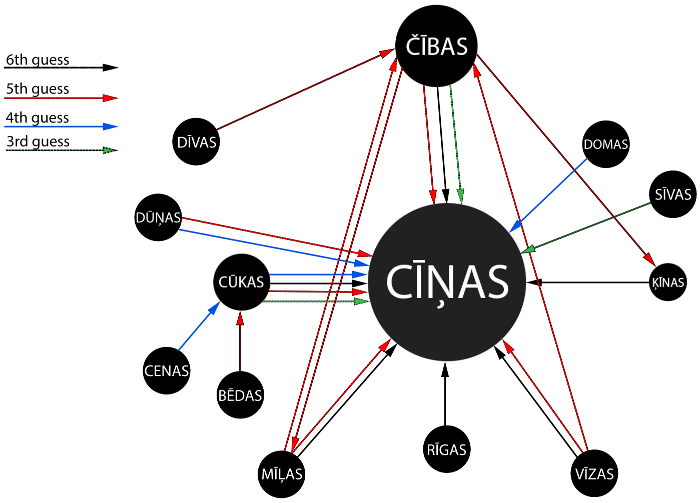


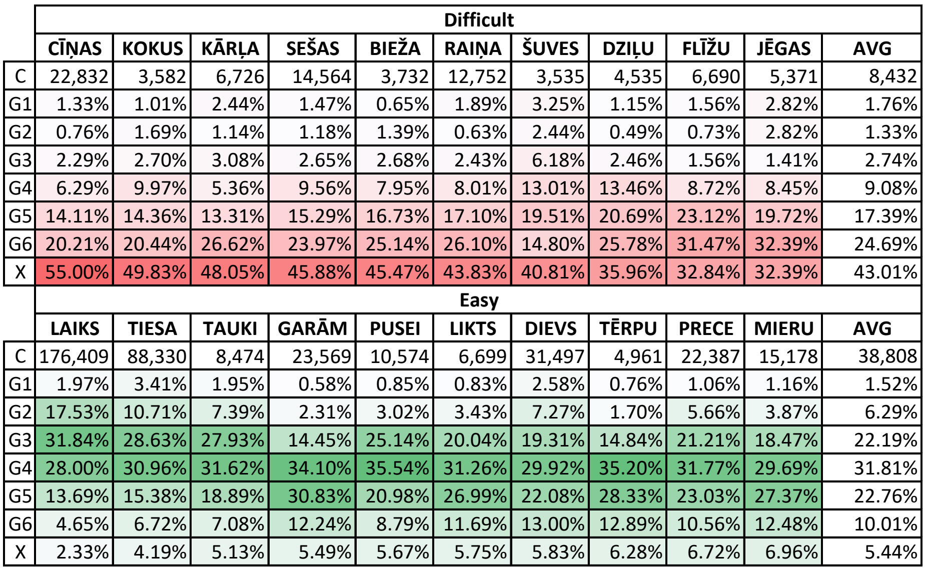
Fig. 2. Paths of previous guesses that lead players to the correct answer for the most difficult word of the day so far – “CĪŅAS”.

4. Conclusion

In this paper, we provided insight in a brief linguistic exercise that has become a fun pastime for a few minutes each day for many players around the world. The creation of a near complete Latvian version of the game is described with further hints on how to make it more or less challenging and the possibility of enriching the vocabulary by linking the game to an online thesaurus is examined. While providing a glimpse into the public perception of the Latvian version of the game, we also dive deep in analysing how the Latvian word game has been played over the first two and a half months, looking at players’ strategies, easier and more difficult words to guess.

In future work, we plan to automatically analyse each daily word morphologically and attempt to predict the difficulty level or even guess the distribution based on a machine learning model.

Table 2. Easiest and most difficult words to guess. Row C indicates the number of occurrences of the word in the specific form in the corpus, rows G1-G6 represent guesses, and row X represents failed games after 6 guesses. English translations of the words can be found in Appendix A.



	Difficult										
	CĪŅAS	KOKUS	KĀRĻA	SEŠAS	BIEŽA	RAIŅA	ŠUVES	DZIĻU	FLĪŽU	JĒGAS	AVG
C	22,832	3,582	6,726	14,564	3,732	12,752	3,535	4,535	6,690	5,371	8,432
G1	1.33%	1.01%	2.44%	1.47%	0.65%	1.89%	3.25%	1.15%	1.56%	2.82%	1.76%
G2	0.76%	1.69%	1.14%	1.18%	1.39%	0.63%	2.44%	0.49%	0.73%	2.82%	1.33%
G3	2.29%	2.70%	3.08%	2.65%	2.68%	2.43%	6.18%	2.46%	1.56%	1.41%	2.74%
G4	6.29%	9.97%	5.36%	9.56%	7.95%	8.01%	13.01%	13.46%	8.72%	8.45%	9.08%
G5	14.11%	14.36%	13.31%	15.29%	16.73%	17.10%	19.51%	20.69%	23.12%	19.72%	17.39%
G6	20.21%	20.44%	26.62%	23.97%	25.14%	26.10%	14.80%	25.78%	31.47%	32.39%	24.69%
X	55.00%	49.83%	48.05%	45.88%	45.47%	43.83%	40.81%	35.96%	32.84%	32.39%	43.01%
	Easy										
	LAIKS	TIESA	TAUKI	GARĀM	PUSEI	LIKTS	DIEVS	TĒRPU	PRECE	MIERU	AVG
C	176,409	88,330	8,474	23,569	10,574	6,699	31,497	4,961	22,387	15,178	38,808
G1	1.97%	3.41%	1.95%	0.58%	0.85%	0.83%	2.58%	0.76%	1.06%	1.16%	1.52%
G2	17.53%	10.71%	7.39%	2.31%	3.02%	3.43%	7.27%	1.70%	5.66%	3.87%	6.29%
G3	31.84%	28.63%	27.93%	14.45%	25.14%	20.04%	19.31%	14.84%	21.21%	18.47%	22.19%
G4	28.00%	30.96%	31.62%	34.10%	35.54%	31.26%	29.92%	35.20%	31.77%	29.69%	31.81%
G5	13.69%	15.38%	18.89%	30.83%	20.98%	26.99%	22.08%	28.33%	23.03%	27.37%	22.76%
G6	4.65%	6.72%	7.08%	12.24%	8.79%	11.69%	13.00%	12.89%	10.56%	12.48%	10.01%
X	2.33%	4.19%	5.13%	5.49%	5.67%	5.75%	5.83%	6.28%	6.72%	6.96%	5.44%

5. Acknowledgements

This work has received funding from the "European Social Fund via IT Academy programme" and the project "Research on Modern Latvian Language and Development of Language Technology" (No. VPP-LETONIKA-2021/1-0006).

References

Brown, K. A. (2022). MODEL, GUESS, CHECK: Wordle as a primer on active learning for materials research. In *npj Computational Materials 8, Article number 97*.

CBC-News (2022). New wordle clone contributes to revitalization of gitxsan nation's language. https://www.cbc.ca/news/canada/british-columbia/wordle-clone-gitxsan-language-revitalization-1.6354421. Accessed: 2022-04-10.

Čudare, A. (2021). Kā radās leģendārās spēles 'Lingo' un 'Karogs'. https://www.delfi.lv/campus/interaktivie-stasti/kurs-uztaisija-legendaras-speles-lingo-un-karogs. Accessed: 2022-05-19.

Heine, B. and Narrog, H. (2011). *Abbreviations*. Oxford University Press.

Ņikiforovs, P. (2011). Latviešu valodas vārdu locītājs. Qualification thesis, Latvijas Universitāte.

Schenck, L. M. (2022). Wordle adapted to indigenous languages. https://abcingles.net/2022/01/29/wordle-adapted-to-indigenous-languages/. Accessed: 2022-04-15.

Spektors, A., Auzina, I., Dargis, R., Gruzitis, N., Paikens, P., Pretkalnina, L., Rituma, L., and Saulite, B. (2016). Tēzaurs.lv: the largest open lexical database for Latvian. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2568–2571, Portorož, Slovenia. European Language Resources Association (ELRA).

Tiedemann, J. (2012). Parallel data, tools and interfaces in opus. In Calzolari, N., Choukri, K., Declerck, T., Dogan, M. U., Maegaard, B., Mariani, J., Odijk, J., and Piperidis, S., editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Victor, D. (2022). Wordle is a love story. `https://www.nytimes.com/2022/01/03/technology/wordle-word-game-creator.html`. Accessed: 2022-04-15.

Vincent, T. (2022). Wordle inspired games for the classroom. `https://learninginhand.com/blog/wordle-games-for-the-classroom`. Accessed: 2022-04-15.

Appendix A. Translations of Easy and Difficult Words

Table 3 shows the English translations and accompanying the part-of-speech tags (Heine and Narrog, 2011) of the easiest and most difficult words to guess (from Table 2).

Table 3. Translations of the easiest and most difficult words to guess with accompanying the part-of-speech tags.

Difficult	**Word**	Battle	Trees	Carl's	Six	Frequent
	Tag	N nom pl	N acc pl	N gen sg	Num nom pl	Adj nom sg
	Word	Rainis'	Stiches	Deep	Tiles	Sense
	Tag	N gen sg	N nom pl	Adj acc sg	N acc pl	N gen sg
Easy	**Word**	Time	Court	Fat	Past	Half
	Tag	N nom sg	N nom sg	N nom pl	Adv	N dat sg
	Word	Put	God	Outfit	Product	Peace
	Tag	V ptcp pst m	N nom sg	N acc sg	N nom sg	N acc sg

Appendix B. English Translations of Top 15 Guesses

English translations and accompanying the part-of-speech tags (Heine and Narrog, 2011) of the top 15 guesses at each turn (from Table 1) are shown in Table 4.

Table 4. Top 15 guesses at each turn translated into English with accompanying the part-of-speech tags. Words that were the actual answers within these days are marked in bold.

G1	Tag	∑	G2	Tag	∑	G3	Tag	∑
Sun	N nom sg	5,341	**Time**	N nom sg	412	**Time**	N nom sg	433
Wall	N nom sg	3,179	Sun	N nom sg	355	**Court**	N nom sg	334
Court	N nom sg	1,579	Day	N nom sg	337	**Fat**	N nom pl	295
Day	N nom sg	1,476	Linden	N nom sg	290	**Thing**	N acc sg	273
Luck	N nom sg	1,449	Cat	N nom sg	284	**Half**	N dat sg	271
Milk	N nom sg	1,237	Milk	N nom sg	266	**Time**	N acc sg	247
Bread	N nom sg	1,217	Apple	N nom sg	262	**Product**	N nom sg	230
Linden	N nom sg	1,159	Sun	N loc sg	207	**God**	N nom sg	225
Link	N nom sg	958	Wall	N nom sg	205	**Put**	V ptcp pst m	224
Box	N nom sg	952	Thing	N nom sg	204	**Halves**	N nom pl	214
Apple	N nom sg	950	Bread	N nom sg	203	**Market**	N acc sg	214
Cat	N nom sg	869	Tire	N nom sg	192	**Outfit**	N acc sg	206
Streets	N nom pl	676	Luck	N nom sg	188	**Peace**	N acc sg	201
School	N nom sg	673	**Fat**	N nom pl	175	**At once**	Adv	200

G4	Tag	∑	G5	Tag	∑	G6	Tag	∑
Outfit	N acc sg	432	**Clean**	V inf	382	**Tiles**	N acc pl	345
Clean	V inf	382	**Do**	V prs 2 pl	364	**Rainis**	N nom sg	296
Half	N acc sg	382	**Longer**	Adv cmp	359	**Pleasure**	N acc sg	295
Later	Adv cmp	380	**Basket**	N loc sg	350	**Gently**	Adv	289
Away	Adv	364	**Own**	Pron dat pl f	340	**Frequent**	Adj nom sg	278
Time	N nom sg	354	**Outfit**	N acc sg	337	**Built**	V ptcp pst f	258
Course	N loc sg	353	**Worth**	Adj N sg m	334	**Own**	Pron dat pl f	250
Paint	N acc sg	340	**Land**	N dat pl	327	**Take**	V deb	245
Put	V ptcp pst m	339	**Pour**	V prs 2 sg	324	**Wide**	Adj nom f	243
Product	N nom sg	338	**Away**	Adv	322	**Good**	Adj nom pl f	238
Table	N acc sg	335	**Good**	Adj nom pl	321	**Land**	N dat pl	237
God	N nom sg	332	**Later**	Adv	321	**Knowing**	V ptcp	236
Block	N acc sg	331	**Likewise**	Adv	317	**Pour**	V prs 2 sg	234
Market	N acc sg	328	**Take**	V deb	317	**Neck**	N acc sg	226

Appendix C. Distributions of Words in the Corpora

Figures 3 and 4 show the distribution of unique n-character words and total word counts of each length in the corpora of ∼32M unique Latvian sentences from Opus. We can see that 5-character words rank only 8th within the corpora, having 54,953 unique forms. However, in terms of total appearances in the corpora, 5-character words are almost as frequent as 2-character words, ranking 5th overall with 47,468,991 total appearances.

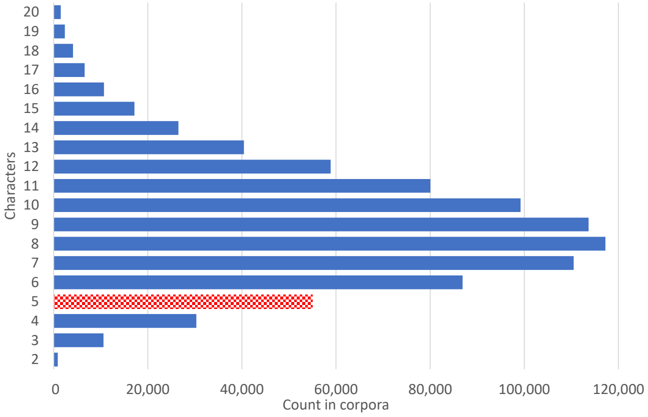


Fig. 3. Distribution of unique n-character words in the corpora.

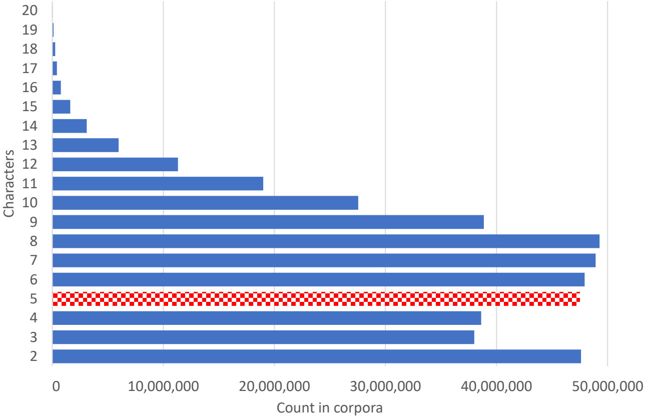


Fig. 4. Total distribution of n-character words in the corpora.

Received May 20, 2022 , accepted July 6, 2022