

Article

Advanced Word Game Design Based on Statistics: A Cross-Linguistic Study with Extended Experiments

Jamolbek Mattiev ^{1,*} , Ulugbek Salaev ¹  and Branko Kavšek ^{2,3} ¹ Computer Science Department, Urgench State University, Khamid Alimdjan 14, Urgench 220100, Uzbekistan; ulugbek.salaev@urdu.uz² Department of Information Sciences and Technologies, University of Primorska, Glagoljaška 8, 6000 Koper, Slovenia; branko.kavsek@upr.si³ Department for Artificial Intelligence, Jožef Stefan Institute, Jamova Cesta 39, 1000 Ljubljana, Slovenia

* Correspondence: jamolbek.mattiev@famnit.upr.si or mattiev.jamolbek@urdu.uz; Tel.: +998-99-5854223

Abstract: Word games are of great importance in the acquisition of vocabulary and letter recognition among children, usually between the ages of 3 and 13, boosting their memory, word retention, spelling, and cognition. Despite the importance of these games, little attention has been paid to the development of word games for low-resource or highly morphologically constructed languages. This study develops an Advanced Cubic-oriented Game (ACG) model by using a character-level N-gram technique and statistics, commonly known as the matching letter game, wherein a player forms words using a given number of cubes with letters on each of its sides. The main objective of this study is to find out the optimal number of letter cubes while maintaining the overall coverage. Comprehensive experiments on 12 datasets (from low-resource and high-resource languages) incorporating morphological features were conducted to form 3–5-letter words using 7–8 cubes and a special case of forming 6–7-letter words using 8–9 cubes. Experimental evaluations show that the ACG model achieved reasonably high results in terms of average total coverage, with 89.5% for 3–5-letter words using eight cubes and 79.7% for 6–7-letter words using nine cubes over 12 datasets. The ACG model obtained over 90% coverage for Uzbek, Turkish, English, Slovenian, Spanish, French, and Malaysian when constructing 3–5-letter words using eight cubes.



Academic Editor: Wei-Chiang Hong

Received: 12 March 2025

Revised: 10 April 2025

Accepted: 15 April 2025

Published: 17 April 2025

Citation: Mattiev, J.; Salaev, U.; Kavšek, B. Advanced Word Game Design Based on Statistics: A Cross-Linguistic Study with Extended Experiments. *Big Data Cogn. Comput.* **2025**, *9*, 103. <https://doi.org/10.3390/bdcc9040103>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: word game modeling; letter frequency; character-level N-gram; model coverage; statistics

1. Introduction

Due to its importance and communicational use, vocabulary is regarded as one of the key elements that students must master when learning a language. Vocabulary plays an important role in the interpretation of spoken and written materials, as well as in learning the meanings and potential uses of new words, according to Viera [1]. Students' thinking and creativity during the language learning process are influenced by their vocabulary knowledge, which enhances the quality of their language acquisition. According to [2], second-language learners should possess sufficient vocabulary to ensure meaningful and successful conversations. Letter games, particularly those based on cubes, can help children from ages 3 to 13 enhance their vocabulary.

Research studies by Azar [3] and Rohani and Pourgharib [4] assert that games are useful for teaching vocabulary because they emphasize the keywords needed to complete the game's objectives. Alavi and Gilakjani [5] and Mageda and Amaal [6] show the

importance of word games in the teaching process for schoolchildren to improve their vocabulary retention skills. Bakhsh [7] defines the effectiveness of word games in teaching vocabulary and proves the usefulness of word games in explaining the meanings of words to young learners. According to Huyen and Nga [8] and Uberman [9], games establish a playful and enjoyable environment that helps young learners quickly gain new knowledge and remember it better. Moreover, word games help children retain knowledge more effectively by encouraging word recognition and letter organization skills. Players can identify letters and words in the first level of a word game. In the next phases, pupils learn how to construct words using the proper letter order while also strengthening their memory, which broadens their vocabulary.

The goal of the matching letter game (MLG) is to use letter cubes to create words by matching the letters on the cubes' sides. For children, Shchukina et al. [10] and Whitney [11] show that the MLG develops a variety of skills related to turning letters into words, and it is tailored for children aged from 3 to 13 years, offering varying levels of difficulty depending on the length of the words and the number of cubes required to build those words. The MLG is ideal for entertaining and educating young children and can enhance children's early language recognition abilities and word formation capabilities. It also encourages children to practice critical thinking and hands-on skills by sorting and analyzing the position of letters, as well as grouping letters to make words.

The game contains letter cubes; flashcards with predefined words, including a picture defining the word; and a game tray with a slot for cubes to match letters into a word. The child looks at the picture and word on a card and then tries to find each matching letter from the large letter cubes to form the word, placing the letters on the tray in the correct order. The MLG includes letter cubes, flashcards with predefined words explained by pictures, and a game tray with slots for placing the cubes while forming words. The child needs to analyze the picture and word on a flashcard and then try to match the letters by using larger letter cubes to construct the word. Once they find the letters, they are placed on the tray in the correct order to construct the word. The examples of matching 3–7 letter Uzbek words using eight cubes are presented in Figure 1. The cases in (a)–(e) describe the matching of 3-, 4-, 5-, 6-, and 7-letter words constructed using eight cubes for the Uzbek language.

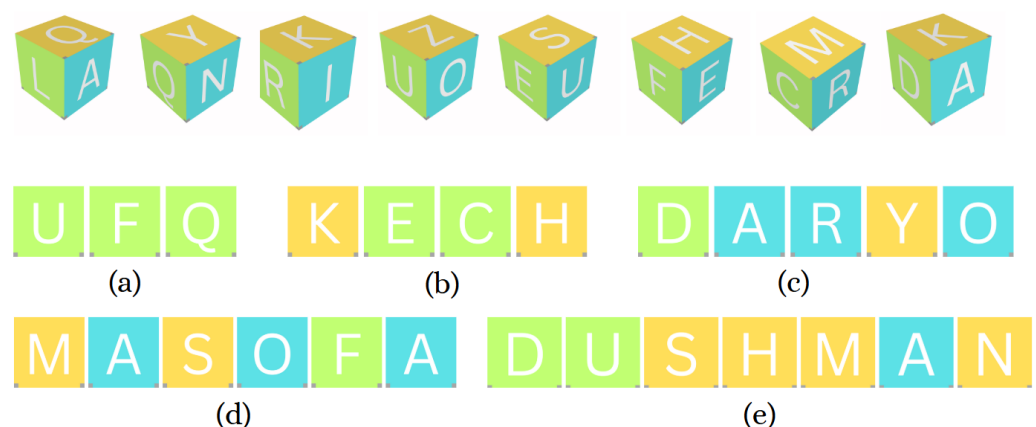


Figure 1. Examples of an English MLG: (a) 3-letter word; (b) 4-letter word; (c) 5-letter word; (d) 6-letter word; and (e) 7-letter word.

The existing cubic-oriented MLGs by Aristides et al. [12] were presented mainly in English and targeted children 3–8 years of age. Most of them have 8–16-letter cubes with options to create a maximum of 64 words from the word flashcards presented and 3–4-letter

words. That is, a lot of cubes are required to create a limited number of words. The MLGs have not been developed for low-resource (Uzbek, Kazakh, and Tatar) and other (Sloven and Poland) languages.

In this paper, we develop a novel methodology to design a cubic-oriented MLG based on the character-level N-gram technique and statistical analysis. Our main goal is to identify the optimal number of letter cubes while maintaining the overall word coverage. We further analyzed all possible combinations of interchanging cubic letters to improve the overall coverage, addressing the above-mentioned challenges. The model produces a set of letters that can be placed on cube faces to facilitate word formation.

The proposed model was evaluated on 12 datasets from various languages, including morphological features for the cases of 3–5-letter words with 7–8 cubes and 6–7-letter words with 8–9 cubes. The experimental results prove that ACG achieved relatively high coverage on all datasets for all cases. Our model obtained over 90% maximum coverage in constructing 3–5-letter words from eight cubes and 80% of maximum coverage in constructing 6–7-letter words using nine cubes on 8 datasets out of 12.

The rest of the paper is organized as follows: An overview of the related work is given in Section 2. Section 3 provides a detailed description of our proposed models. The attained results are comprehensively discussed in Section 4. The study is concluded and our future directions are outlined in Section 6.

2. Scientific Background

Existing word game models are mainly focused on high-resource languages like English, French, or Spanish, and none of them are aimed at designing cubic-oriented word games. Therefore, our current study, developing cubic-based word game models for low-resource languages, can be considered a novel contribution to the field. We analyze some of the word game models in this section.

Study [13] developed a cubic-oriented model based on a character-level N-gram language model and showed the importance of word games in the vocabulary teaching process. That research was conducted exclusively for children to learn 3–5-letter words by using 5–8 cubes in Uzbek, Sloven, English, and Russian. The key differences between the current study and the study in [13] are that (1) we added a new optimization technique to improve the overall coverage and (2) comprehensive evaluations were performed on 12 datasets, including inflected words that were morphologically constructed, with extended cases of 6–7 letters using 8–9 cubes (for young learners and teenagers) to show the advantages of the novel models.

Uzbek, as an agglutinative language, forms many words by adding multiple affixes to a single root, leading to data sparsity in statistical models. The research in [14] addresses this challenge by developing a morphological analysis model that identifies key inflectional patterns. Our proposed model extends this by incorporating words with complex morphological structures, ensuring a more comprehensive analysis of the Uzbek language.

The study by Zaitun [15] analyzed the effectiveness of the Big Cube game in students' development of vocabulary mastery. For a "guessing with words and pictures" activity, 40 students took part in developing the game, conducted by the researchers. Through experimental testing, they confirmed that the game can effectively serve as an educational tool. Comparing the pre-test and post-test data, both consisting of 20 questions in parts of multiple choice and sentence completion, a positive impact of cubic games in the educational process was statistically confirmed.

Prior research has also explored cube-based word and puzzle games as a medium for assessing player interaction and cognitive behavior. For instance, Anadon et al. [16] proposed a two-stage analytical approach to characterize players of a cube puzzle game

by using performance clustering and decision trees to evaluate gameplay behavior and strategies. While their study focused on behavioral patterns and user categorization in a gamified puzzle setting, our research diverges by targeting linguistic and educational aspects—specifically the optimization of letter cube configurations for vocabulary development across multiple languages. However, both approaches underscore the flexibility and educational potential of cubic puzzle games, whether in cognitive or linguistic dimensions.

In the study in [17], the authors present a statistical approach for predicting word data, emphasizing the importance of accurate linguistic data forecasting in various computational applications. Their methodology involves analyzing word frequency distributions and applying statistical models to predict subsequent word occurrences. This approach is particularly relevant to our research on optimizing letter cube configurations for multilingual vocabulary games, as both studies leverage statistical analyses of language data to enhance linguistic applications. Integrating such predictive models can significantly improve the design and adaptability of educational tools across diverse languages.

Effectiveness in the utilization of word games in teaching English vocabulary in a school in Vietnam was established by Vu et al. [18] through a demonstration with two classes of students for eight weeks to assess the effectiveness of using games in teaching and improving learning and in enhancing the performance of students in post-tests, with a significant improvement in vocabulary retention. In the study, statistical support for using word games to improve language learning and vocabulary learning for students in schools has been demonstrated.

The study by Anugerah et al. [19] examined the effectiveness of instruction using the Build-A-Sentence cube game for instruction in simple past tense. With 16 students in the Eighth Level Q-Learning Course Pontianak, post-test achievement increased compared to the pre-test results, and it can be concluded that instruction with the use of the game helped students at a moderate level. This study also proves the usefulness of word games in teaching vocabulary to young learners.

The primary objective of our research work aligns with the objectives of the aforementioned studies, but our work introduces new approaches toward developing a cubic-oriented word game, and the present studies have focused predominantly on the integration of word games in the educational process. As a low-resource language, the Uzbek language lacks publicly accessible corpora, and algorithm development in such a scenario is even more challenging. Our main contributions and objectives are as follows:

- Within this study, we developed a new dataset for the Uzbek language by involving language experts. This can be considered a good contribution because the Uzbek language is a low-resource language and there are not enough publicly open corpora.
- We developed a novel methodology based on vowel–consonant patterns and statistical analysis to design a cubic-oriented MLG for low-resource and high-resource languages.
- The performance of the model was evaluated by comprehensive experiments on 12 datasets.

3. Methodology

In developing our approach, we established specific conditions and constraints. Since the game is intended to help children familiarize themselves with letters, we ensured that every letter of the alphabet appeared at least once on the cubes. To optimize the model, we implemented two key restrictions. First, each cube contains unique letters, meaning no letter is repeated on the same cube. This enhances gameplay by allowing players to form a wider variety of words, as only one face of a cube can be used at a time. Second, each cube is designed to include two or three vowel letters to maintain a balanced vowel distribution

across all cubes. It can be seen from [13] that vowels are the most used letters to construct words in different languages. By considering the above fact, it is important to use more vowel letters in cubes and place those vowels in different cubes. Our methodology mainly consists of the following 5 steps for constructing optimal cubes: data preparation, statistical analysis (generation of letter frequencies based on Uni-gram and Bi-gram techniques), identification of letter position, generation of cubes, and interchanging the letter cubes to form the final model. The proposed methodology is presented in Figure 2.

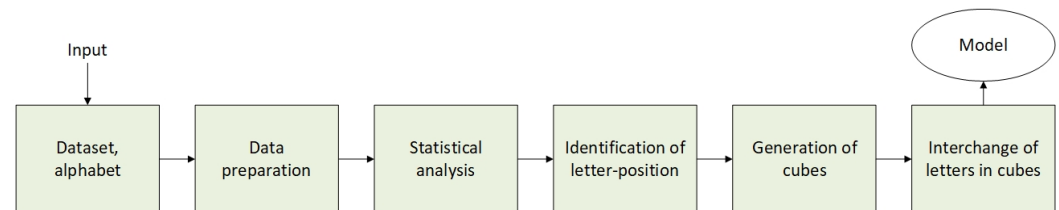


Figure 2. Overview of the developed model.

The following subsections describe the above-mentioned steps in detail.

3.1. Data Preparation

The dataset consists of the following 12 languages: German (de), English (en), Spanish (es), French (fr), Kazakh (kz), Malay (ms), Polish (pl), Russian (ru), Slovenian (sl), Turkish (tr), Tatar (tt), and Uzbek (uz). These languages belong to different linguistic families based on their structural characteristics. Uzbek, Kazakh, Tatar, Turkish, and Malay are examples of agglutinative languages in the collection, where words are created by adding affixes to a root. Complex inflections are used to express meaning in fusional languages like Spanish, German, French, Polish, Russian, and others. A thorough linguistic representation is ensured by the diversity that English and Slovenian provide as analytic and inflected languages, respectively, to the dataset. Since there is no dataset for young learners in the Uzbek language, we created a new dataset that includes 3–5-letter words (for young learners) and 6–7-letter words (for teenagers) with the help of language experts. The data preparation phase is defined as follows:

- (1) Collection of words. We extracted the 3–5-letter words that are appropriate for young learners from the largest dictionary book in the Uzbek language [20] and 6–7-letter words from [21], as well as a syntactic tagged corpus for the Uzbek language [22]. We extracted 3–5-letter words for the English language from ESL Forums <https://eslforums.com> (accessed on 20 September 2024), for the Kazakh language from Ref. [23], for the Russian language from [24], and for Slovenian from Ref. [25]. All words for German, Tatar, Spanish, Kazakh, Malay, Polish, Turkish, and French were extracted from Ref. [21].
- (2) Normalization. After the generation of the word list for our dataset, we performed normalization to simplify the coding.
 - (a) The Uzbek language involves digraphs with diacritic markings, and it is not an easy matter to calculate letter frequency in them. To bypass this, we replaced g' and o' with modified characters ($\bar{g}$ and $\bar{o}$), converting each digraph into a single character. Despite initially being regarded as two single characters in a word, with this substitution, the letter frequency could be determined with a high level of accuracy. For ch and sh , these two digraphs were analyzed and regarded as individual parts (s and h). As c is not part of the Uzbek alphabet, we retained it in its form of a digraph and regarded it as an individual character in the calculation of letter frequency.

- (b) The Uzbek alphabet contains a phonetic glottal stop (*Tutuq belgisi*) character, which is not a letter but is included in the alphabet. Since only 18 words in the corpus of Uzbek contained this character, in an attempt to make the analysis easier, in filtering, we eliminated these words.
- (3) Filtering. After the normalization, the Uzbek corpus consisted of 18,523 words. To make our corpus more appropriate for kids and teenagers, 4 volunteer experts from the Uzbek Linguistics Department of Urgench State University were involved, and they helped eliminate infrequent and unfamiliar words not suitable for young learners. As a result, the filtered corpus consisted of 4558 three- to five-letter words and 8456 six- to seven-letter words. All other datasets were generated by filtering words with 3- to 7-letters and removing less frequently occurring words. The datasets are publicly available at <https://github.com/UlugbekSalaev/MatchingLetterGame> (accessed on 14 February 2025).

3.2. Descriptive Statistics

We computed the statistics on the selected languages to analyze the occurrences of vowel and consonant letters in words. The alphabets of the languages and their occurrences in the datasets including 3–5-letter words are shown in Table 1 and 6–7-letter words in Table 2 (infrequent letters are marked in bold).

Table 1 proves the active participation of vowel letters in all languages to construct 3–5-letter words, and these statistics can be considered in the letter placement phase of the model.

Table 1. Letter frequencies for the dataset representing 3–5-letter words and their frequency percentage.

	de	en	es	fr	kz	ms	pl	ru	sl	tr	tt	uz
e	14.6	e 10.4	a 15.1	e 11.2	a 11.8	a 16.6	a 9.9	a 10.8	a 11.2	a 12.8	a 10.9	a 14.1
t	8.8	a 9.2	o 10.0	a 8.4	e 6.8	i 8.3	o 6.2	o 9.0	e 9.3	e 8.4	ы 6.6	i 8.3
a	7.1	s 8.5	e 9.3	s 7.7	ы 6.7	u 6.9	i 6.0	к 6.2	o 8.5	i 7.6	т 6.2	o 8.2
n	6.7	o 7.4	s 7.4	r 7.1	т 6.1	e 6.1	e 5.5	р 6.0	г 6.8	г 5.6	е 6.2	г 6.0
r	6.4	r 6.0	r 6.6	i 7.1	i 5.2	r 5.8	r 5.2	т 5.9	i 6.6	к 5.6	к 5.7	л 5.1
s	5.2	l 5.4	i 5.6	o 6.3	р 5.0	t 5.8	u 4.4	e 5.2	t 5.3	n 5.5	р 5.1	s 4.9
i	5.2	t 5.4	l 5.1	t 5.7	н 4.9	s 5.8	s 4.3	л 5.0	n 5.3	л 4.9	н 5.1	т 4.9
l	4.8	i 5.4	t 4.6	u 5.2	с 4.6	k 5.5	y 4.2	и 4.8	к 5.3	т 4.6	э 5.0	u 4.8
o	4.6	d 4.6	n 4.5	l 5.0	л 3.9	n 5.1	k 4.2	с 4.6	l 4.7	s 4.1	и 4.5	n 4.4
h	4.1	n 4.3	c 4.2	n 4.8	қ 3.6	l 5.0	n 4.0	н 4.5	s 4.0	m 3.9	л 4.5	m 3.7
u	3.5	c 4.0	u 4.0	é 4.5	y 3.5	o 4.4	t 4.0	y 3.6	p 3.5	ı 3.9	y 3.8	q 3.6
g	3.4	u 3.7	p 3.3	c 3.6	к 3.3	m 3.9	m 3.6	м 3.1	vs. 3.3	u 3.9	с 3.6	k 3.4
m	3.1	b 3.6	d 3.2	m 3.1	o 3.2	p 3.4	d 3.6	и 3.1	d 3.2	d 3.4	м 3.3	y 3.1
b	3.0	p 3.2	m 3.1	p 3.0	д 3.0	b 3.2	w 3.5	в 3.0	u 2.9	o 3.0	г 2.3	h 3.0
d	2.9	m 2.9	b 2.4	d 2.7	м 2.9	d 2.9	z 3.4	д 3.0	m 2.9	y 2.8	o 2.3	b 2.9
f	2.4	h 2.7	g 2.1	g 2.4	п 2.5	h 2.6	p 3.3	б 2.7	b 2.6	ü 2.3	я 2.2	e 2.6
k	2.3	g 2.5	vs. 1.5	b 2.2	ш 2.3	g 2.2	l 3.2	ь 2.4	j 2.4	z 2.3	п 2.2	d 2.5
w	2.0	f 2.2	h 1.5	vs. 2.1	б 2.3	c 1.5	b 2.8	г 2.0	c 2.3	b 2.1	б 2.2	z 2.5
c	1.8	y 2.2	f 1.4	f 1.7	з 2.0	j 1.4	ı 2.8	з 1.9	g 1.9	ş 1.9	з 2.1	vs. 2.1
p	1.7	k 2.0	j 1.1	h 1.3	и 1.9	f 1.1	c 2.5	я 1.7	z 1.9	p 1.8	ч 2.1	ö 1.9
z	1.3	w 1.7	y 1.0	y 0.6	ж 1.8	w 0.9	а 2.4	ч 1.4	š 1.7	ç 1.6	ш 2.1	p 1.4
ü	1.0	vs. 1.1	z 0.8	x 0.6	ү 1.8	y 0.7	g 1.9	ы 1.4	ž 1.4	h 1.5	д 2.0	f 1.4
ä	1.0	x 0.5	k 0.6	j 0.5	й 1.7	vs. 0.4	e 1.7	ш 1.3	h 1.2	g 1.3	ү 1.9	g 1.3
v	0.9	z 0.4	ñ 0.4	k 0.5	ү 1.5	z 0.3	j 1.7	й 1.3	č 1.1	ö 1.2	ө 1.5	j 1.2
ö	0.5	j 0.4	x 0.4	è 0.4	ө 1.4	x 0.1	ó 1.4	x 1.3	f 0.8	f 1.2	й 1.4	g 1.1
j	0.5	q 0.1	q 0.3	z 0.3	г 1.3	q 0.1	ž 1.2	ж 1.2		vs. 1.2	ж 0.8	x 1.0
ß	0.4		w 0.3	ä 0.3	г 1.0		f 0.7	ë 1.0		c 0.9	ң 0.8	c 0.6
y	0.4			è 0.3	ң 1.0		é 0.7	ф 0.9		ğ 0.6	x 0.7	
x	0.3			w 0.2	э 0.9		h 0.7	ц 0.6		j 0.1	ю 0.6	
q	0.1			ö 0.2	я 0.6		š 0.5	ю 0.5			ф 0.6	
				û 0.2	x 0.4		ń 0.2	ш 0.3			в 0.6	
				q 0.2	ф 0.3		ž 0.1	э 0.3			э 0.5	
				î 0.1	ю 0.3			ь 0.1			и 0.2	
				ï 0.1	в 0.2						ь 0.2	
				œ 0.1	э 0.1						ж 0.1	
				ç 0.1	ь 0.1						ц 0.1	
				ë 0.1	ц 0.1						ь 0.0	
				à 0.1	ч 0.1							
				æ 0.1	и 0.1							
					ш 0.1							
					ь 0.1							
					ë 0.1							

It can be seen from the table that every language has some infrequent letters used to construct the words. We performed this statistical analysis to improve the overall coverage by replacing the least frequent letters ($\leq 0.2\%$) with the most frequent letters.

Table 2. Letter frequencies for the dataset representing 6–7-letter words and their frequency percentage.

de		en		es		fr		kz		ms		pl		ru		sl		tr		tt		uz	
e	17.7	e	13.1	a	16.2	e	12.3	a	12.8	a	16.5	a	10.1	o	9.7	o	10.0	a	12.8	a	11.5	a	14.6
t	9.4	s	8.5	e	10.5	r	8.7	ы	8.5	n	8.7	o	7.1	a	8.8	a	10.0	e	8.9	ы	7.7	i	12.8
n	8.5	r	7.8	r	8.8	s	8.2	e	7.0	i	8.6	i	6.9	e	7.5	i	9.4	i	7.7	e	6.9	o	6.2
r	7.4	a	7.4	o	8.7	a	8.0	н	6.6	e	8.5	e	6.4	и	6.7	e	9.3	n	6.7	н	6.5	r	6.2
a	5.5	i	6.9	s	7.5	i	7.0	т	6.3	r	6.2	r	5.2	р	5.7	n	6.6	r	6.1	р	6.2	n	5.9
s	5.3	t	6.4	i	6.1	t	6.3	i	5.8	u	5.4	z	4.7	т	5.6	r	5.8	l	5.6	л	5.8	l	5.8
i	5.3	n	6.1	n	5.6	n	5.8	р	5.3	t	5.4	n	4.5	н	5.6	l	5.2	ı	5.6	ə	5.8	s	5.2
l	5.1	o	5.3	t	4.9	é	5.5	л	4.5	s	5.1	w	4.3	л	4.7	t	4.3	k	4.7	т	5.6	t	4.7
g	4.4	l	5.2	d	4.4	o	5.3	c	4.3	k	4.7	y	4.1	c	4.7	s	4.1	t	4.3	к	5.4	d	3.9
h	3.8	d	5.2	l	4.4	u	4.8	д	4.0	m	4.3	k	4.1	к	4.0	vs.	3.9	m	4.2	c	3.6	u	3.3
u	3.0	c	3.8	c	4.3	l	4.5	қ	3.3	l	4.1	s	3.9	в	4.0	k	3.8	d	3.9	и	3.5	h	3.3
b	2.8	u	3.5	u	3.5	c	4.1	к	2.8	p	3.3	c	3.6	м	3.7	p	3.7	s	3.7	м	3.1	k	3.1
o	2.5	g	3.3	m	2.7	p	2.9	y	2.7	g	3.3	t	3.5	y	3.2	d	3.5	y	3.6	г	3.1	m	3.1
m	2.5	p	3.0	p	2.6	m	2.7	м	2.7	d	3.2	d	3.4	д	3.1	j	3.1	u	3.5	y	2.6	g	2.8
d	2.3	m	2.5	b	2.1	d	2.6	o	2.2	o	2.9	u	3.2	п	2.9	m	3.1	o	2.3	д	2.5	b	2.6
f	2.3	h	2.3	g	1.8	g	2.1	б	2.1	b	2.9	m	3.1	ы	2.8	z	2.5	ü	2.1	б	2.1	y	2.5
c	2.2	b	2.0	vs.	1.5	vs.	1.8	п	2.0	h	2.0	p	3.0	з	2.0	u	2.3	z	1.7	o	1.9	e	2.4
k	2.1	y	1.7	f	1.0	b	1.6	ғ	1.7	y	1.2	ı	2.9	б	1.9	b	1.9	b	1.7	ш	1.8	q	2.3
p	1.6	f	1.7	j	0.9	f	1.6	ш	1.7	j	1.0	l	2.5	я	1.8	č	1.6	ş	1.6	з	1.8	z	1.7
z	1.3	k	1.2	h	0.8	h	1.3	й	1.6	c	1.0	b	2.0	г	1.6	g	1.6	g	1.2	ч	1.7	ö	1.3
w	1.2	w	1.2	z	0.6	x	0.5	з	1.6	f	0.6	a	1.9	й	1.5	h	1.3	ç	1.2	п	1.5	vs.	1.3
ü	1.0	vs.	1.1	ñ	0.3	q	0.5	ж	1.5	w	0.5	j	1.7	x	1.4	š	1.1	p	1.0	ү	1.5	p	1.1
ä	1.0	x	0.3	q	0.3	y	0.4	и	1.4	vs.	0.3	g	1.6	ч	1.2	c	0.9	ğ	1.0	й	1.3	c	0.9
v	0.7	z	0.2	x	0.2	è	0.4	ң	1.4	z	0.2	ę	1.1	ь	1.1	ž	0.9	c	1.0	я	1.3	f	0.8
ö	0.5	j	0.2	y	0.2	j	0.3	г	1.3	q	0.1	ó	1.1	ж	1.0	f	0.3	h	1.0	ө	1.0	x	0.7
ß	0.2	q	0.2	k	0.1	k	0.2	ү	1.1	x	0.1	h	1.0	ю	0.7			ö	0.9	ң	0.9	ğ	0.7
j	0.2			w	0.1	ê	0.1	ө	1.0			č	0.9	ш	0.7			vs.	0.9	в	0.6	j	0.6
y	0.1					z	0.1	ү	1.0			ž	0.8	ц	0.6			f	0.7	x	0.6		
x	0.1					ô	0.1	ə	0.6			š	0.6	ë	0.6			j	0.1	ж	0.6		
q	0.1					ç	0.1	я	0.3			f	0.4	ф	0.4					ф	0.4		
						â	0.1	ф	0.2			ń	0.2	щ	0.3					э	0.3		
						û	0.1	x	0.2			ž	0.1	э	0.1					ю	0.3		
						î	0.1	в	0.2					ь	0.1					ь	0.2		
						ï	0.1	ц	0.1											h	0.1		
						w	0.1	ь	0.1											ж	0.1		
						œ	0.1	ю	0.1											ц	0.1		
						ë	0.1	э	0.1											ь	0.1		
								ч	0.1											щ	0.1		
								h	0.1														
								ь	0.1														
								щ	0.1														

Table 2 shows that the French and Kazakh languages have more infrequent letters. The main reason is that the French and Kazakh languages have more letters in their alphabets and some of them occur in words very rarely.

3.3. Proposed Method

We proposed a novel approach to designing a cube-oriented game by considering the statistical analysis of letters. In the first step (Algorithm 1), the frequencies of the letters for each dataset are generated based on Unigram and character-level Bigram techniques [26]. The use of Unigram and Bigram techniques was intentionally chosen to align with the primary goal of the study—placing letters in separate cubes to maximize coverage while forming words. Using higher-order N-grams, such as Trigrams, could result in multiple letters from the same N-gram being placed in the same cube, which would limit the flexibility of the game and reduce its intended effectiveness. In the second step, infrequent letters are eliminated from the alphabet based on the letter frequency tables and are later replaced by frequent letters. This step can help improve overall coverage. In the third step (Algorithm 2), potential cubes are generated from the letters produced in the previous step. The optimization technique is used by interchanging cubic letters in the final step (Algorithm 3).

Algorithm 1 Generation of Letter Frequencies (GLF) based on a character-level N-gram technique**Input:** An Alphabet A , dataset D , number of character n **Output:** Dictionary (list of key-value pair)**Initialization:** Empty dictionary *Frequency* to store the frequency percentage of character(s) N-gram, assign 0 to *total_ngram*

```

1: for each word  $\in D$  do
2:   for ( $i = 0; i \leq \text{word.length} - n; i++$ ) do
3:     if word[ $i : i + n$ ]  $\in A$  then
4:       if word[ $i : i + n$ ]  $\in \text{Frequency.keys}()$  then
5:         Increment Frequency[word[ $i : i + n$ ]] by 1
6:       else
7:         Frequency[word[ $i : i + n$ ]]  $\leftarrow 1$ 
8:       end if
9:     end if
10:    Increment total_ngram by 1
11:   end for
12: end for
13: for each key  $\in \text{Frequency.keys}()$  do
14:   Frequency[key]  $\leftarrow (\text{Frequency}[\text{key}] / \text{total\_ngram}) * 100$ 
15: end for
16: return Frequency

```

CLF takes the alphabet (list of letters), dataset, and n (1: Unigram; 2: Bigram) as input parameters. The CLF algorithm computes the word frequencies based on the character-level N-gram technique. If n equals 1, the method is called as a Unigram, producing a dictionary that contains a letter as key, with their frequency percentages as the value (e.g., a : 14.6; r : 6.2). When n equals 2, the method is called a Bigram, returning a dictionary of 2-letter word combinations with their corresponding frequency percentages (e.g., ar : 8.4, ut : 5.7).

Algorithm 2 Elimination of infrequent letters (EIL)**Input:** An Alphabet A , Dataset D , p threshold**Output:** A *Letter_list***Initialization:** Empty *Letter_list*

```

1: Alphabet  $\leftarrow \text{GLF}(A, D, 1)$ 
2: for each letter  $\in \text{Alphabet}$  do
3:   if letter.frequency  $> p$  then
4:     Letter_list  $\leftarrow \text{Letter\_list} \cup \{\text{letter}\}$ 
5:   end if
6: end for
7: return Letter_list

```

EIL takes the alphabet of the language, dataset, and threshold p as input parameters and returns the reduced list of letters. EIL removes the infrequent letters whose frequency percentage is lower than p . Algorithm 3 defines the procedure to generate all possible new sets by swapping a single letter between any two cubes within an initial set of cubes.

Algorithm 3 starts by initializing an empty list to store unique sets. It iterates through each pair of cubes in the set, ensuring that each pair is considered only once. For each pair, the algorithm performs a nested iteration over all letters in the two selected cubes. A temporary copy of the initial set is created to avoid altering the original set. The algorithm then exchanges one letter between the two cubes, forming a new set. If the newly generated set is different from the initial set and is not already included in the list of unique sets, it is

added to the list. Finally, the algorithm returns all unique sets generated by the swapping process. The final model is outlined in Algorithm 4.

Algorithm 3 Generation of Potential Letter Cubes (GPLC)

Input: *Cubes* (2D array with size of $N \times 6$)

Output: A set U , optimized cube configurations (each configuration being a set of N cubes with 6 distinct letters).

Initialization: Assign an empty set U to store combinations of cubes, number of cubes N from *Cubes*.

```

1: for  $i, j \in \{1, \dots, 8\}, i < j$  do
2:   for  $k, l \in \{1, \dots, 6\}$  do
3:      $Cubes' \leftarrow Cubes$ 
4:     Swap  $Cubes'[i][k]$  and  $Cubes'[j][l]$ 
5:     if  $Cubes' \notin U$  then
6:        $U \leftarrow U \cup \{Cubes'\}$ 
7:     end if
8:   end for
9: end for
10: return  $U$ 

```

Algorithm 4 Generation of the optimized letter cube

Input: Alphabet A , a list of vowels V , dataset D , number of cubes N

Output: Set of *Cubes* (*Cubes* is a 2D array with size of $N \times 6$)

Initialization: Empty list L to store frequent letters, empty list DL to store the duplicate letters, Empty list SL to store the sequence of letters, empty 2D array *Cubes* with size of $N \times 6$ to store cubic letters

```

1:  $L \leftarrow EIL(A, D, 0.2)$ 
2:  $Unigram\_Frq \leftarrow GLF(L, D, 1)$  ▷ Returns dictionary which contains a letter (key) and its frequency (value), i.e., [ $a: 14.6, i: 12.8, \dots$ ]
3:  $Bigram\_Frq \leftarrow GLF(L, D, 2)$  ▷ Returns dictionary which contains 2 letters (key) and its frequency (value), i.e., [ $sa: 6.2, te: 2.3, \dots$ ]
4:  $Unigram\_Frq \leftarrow sort(Unigram\_Frq, value)$ 
5:  $Bigram\_Frq \leftarrow sort(Bigram\_Frq, value)$ 
6: for each  $key \in Bigram\_Frq$  do
7:   for each  $letter \in key$  do
8:     if  $letter \notin SL$  then
9:        $SL \leftarrow SL \cup \{letter\}$ 
10:    end if
11:  end for
12: if  $SL = L$  then
13:   break
14: end if
15: end for
16:  $SL \leftarrow SL \cup EIL(V, D, 5)$ 
17:  $SL \leftarrow SL \cup Unigram\_Frq.keys[0 : N * 6 - SL.length]$ 
18:  $SL \leftarrow sort(SL, descending\ by\ Bigram\_Frq.value)$ 
19: for  $(i = 0; i < N; i++)$  do
20:   for  $(j = 0; j < 6; j++)$  do
21:      $Cubes[i][j] \leftarrow SL[j * 6 + i]$ 
22:   end for
23: end for
24: return  $GPLC(Cubes)$ 

```

The input of Algorithm 4 is an alphabet of the language, vowel letters, dataset, and number of cubes. The first line extracts the frequent letters from the alphabet by removing the infrequent letters (frequency $\leq 0.2\%$). The second and third lines utilize the Unigram

and Bigram techniques to find the frequencies of one letter and two consecutive letters, and these frequencies are sorted by values in lines 4–5. Lines 6–15 form the sequence of unique letters from *Bigram_Frq*. In lines 16–17, vowel letters whose frequencies are higher than 5% are added to the letter list (*SL*) to fill the letter list (*SL*) up to $N \times 6$. The letter list *SL* is sorted based on the *Bigram_Frq* value to avoid placing the same letters into the same cube. Lines 19–23 place the letters in *SL* to the cube sides ($N \times 6$) to form the base cubes. The last line computes all the combinations of the potential set of cubes by using base cubes and returning the results.

4. Experimental Results

We performed experimental evaluation on 12 datasets. The proposed model was tested in four cases: 3–5-letter words on seven cubes; 3–5-letter words on eight cubes; 6–7-letter words on eight cubes; and 6–7-letter words on nine cubes. The detailed information about datasets is presented in Table 3.

Table 3. Description of the datasets.

Dataset	# of 3-Letter Words	# of 4-Letter Words	# of 5-Letter Words	Total	# of 6-Letter Words	# of 7-Letter Words	Total
de	439	924	1763	3126	2944	4100	7044
en	1026	2499	2499	6024	3295	3908	7203
es	483	1514	3541	5538	2500	3707	6207
fr	350	991	2293	3634	3430	4633	8063
kz	493	1210	2836	4539	4068	5371	9439
ms	438	1129	2262	3829	2275	2995	5270
pl	350	1274	3094	4718	4108	5448	9556
ru	516	1285	2507	4308	4017	5068	9085
sl	515	1598	4167	6280	4198	4991	9189
tr	417	1151	2824	4392	3563	5191	8754
tt	403	1089	2574	4066	4077	5401	9478
uz	518	1165	2875	4558	3774	4682	8456

Table 4 presents the experimental results of the *ACG* model for seven cubes with 3–5-letter words. The experimental results show the overall coverage (computed as the mean coverage across the entire set of cubes that returns from Algorithm 3), along with the standard deviations and the number of cubes generated obtained by interchanging cubic letters. The *Max* case indicates the highest coverage achieved by the *ACG* model for a dataset across all the sets of cubes.

It can be seen from Table 4 that the highest total coverage is observed in the *ms* dataset (92.4%), followed by *en* (90.6%) and *sl* (87.8%), while the lowest total coverage is obtained in the *tt* (55.3%) and *kz* (62.6%) datasets. For three-letter words, *ACG* performed the best result on *ms* (95.6%) and *sl* (95.4%) datasets, while the *tt* (75.1%) and *kz* (77.4%) datasets show lower coverage. The *Max* case results present that the *en* (91.8%), *ms* (93.2%), and *sl* (90.0%) datasets gained the highest peak coverage, proving that *ACG* performs best in these datasets when optimizing cube design for 3–5-letter words.

Table 4. Overall coverage (%) with a standard deviation for the ACG model in the case of 7 cubes with 3–5-letter words.

Dataset	# of Generated Cubes	3 Letters	4 Letters	5 Letters	Total	Max Case
de	596	91.8 ± 0.6	82.9 ± 0.9	70.5 ± 1.3	77.1 ± 0.9	78.9
en	512	92.4 ± 0.2	92.1 ± 0.4	88.4 ± 0.8	90.6 ± 0.5	91.8
es	525	84.0 ± 0.1	87.4 ± 0.3	86.3 ± 0.9	86.4 ± 0.6	87.5
fr	489	83.6 ± 0.7	78.5 ± 0.7	74.5 ± 0.9	76.5 ± 0.8	77.9
kz	598	77.4 ± 0.7	72.6 ± 0.9	55.8 ± 1.2	62.6 ± 0.9	65.7
ms	503	95.6 ± 0.2	95.2 ± 0.3	90.4 ± 0.8	92.4 ± 0.5	93.2
pl	591	86.3 ± 0.8	75.8 ± 1.1	58.0 ± 1.2	64.9 ± 1.0	67.8
ru	617	84.8 ± 0.7	76.7 ± 1.0	56.8 ± 1.8	66.1 ± 1.3	69.8
sl	544	95.4 ± 0.5	93.3 ± 0.5	84.8 ± 1.1	87.8 ± 0.8	90.0
tr	578	91.2 ± 0.7	85.3 ± 0.7	69.5 ± 1.1	75.7 ± 0.9	78.3
tt	600	75.1 ± 1.2	66.4 ± 1.1	47.5 ± 1.0	55.3 ± 0.9	58.7
uz	577	94.0 ± 0.5	87.6 ± 0.9	73.9 ± 2.0	79.7 ± 1.5	84.0

Although ACG gained 77.4% and 75.1% overall coverage in the *kz* and *tt* languages in constructing three-letter words from seven cubes, the coverage decreased around 30% in constructing five-letter words, which is an expected behavior.

Table 5 demonstrates that the overall coverage of the ACG model was noticeably improved on all datasets in the case where eight cubes were used for 3–5-letter words compared to seven cubes. ACG achieved a coverage below 80% for five-letter words on the *fr*, *kz*, and *tt* datasets. The key reason is that the Kazakh, Polish, and Tatar languages include more letters in their alphabets and more infrequent letters compared to other languages. In general, our model obtained over 90% of Max case on 8 datasets out of 12.

Table 5. Overall coverage (%) with the standard deviation of the ACG model in the case of 8 cubes with 3–5-letter words.

Dataset	# of Generated Cubes	3 Letters	4 Letters	5 Letters	Total	Max Case
de	790	96.8 ± 0.3	93.4 ± 0.4	86.3 ± 1.0	89.8 ± 0.6	91.4
en	720	94.5 ± 0.1	94.2 ± 0.3	93.5 ± 0.7	94.0 ± 0.4	94.8
es	740	90.5 ± 0.1	93.0 ± 0.2	91.5 ± 0.7	91.8 ± 0.5	93.0
fr	687	87.4 ± 0.1	84.9 ± 0.2	86.4 ± 0.3	86.1 ± 0.2	86.8
kz	792	86.5 ± 0.4	87.8 ± 0.4	78.9 ± 0.9	82.1 ± 0.7	83.4
ms	728	99.6 ± 0.1	98.8 ± 0.3	95.0 ± 0.9	96.7 ± 0.6	98.2
pl	791	92.4 ± 0.6	88.7 ± 0.7	78.9 ± 1.2	82.5 ± 1.0	85.0
ru	797	94.2 ± 0.3	92.0 ± 0.5	85.9 ± 0.8	88.7 ± 0.6	90.5
sl	750	99.2 ± 0.1	97.4 ± 0.5	93.3 ± 1.0	94.8 ± 0.8	96.6
tr	770	97.4 ± 0.3	95.1 ± 0.4	87.2 ± 1.0	90.3 ± 0.7	92.1
tt	773	90.6 ± 0.6	86.2 ± 0.6	78.0 ± 0.8	81.5 ± 0.6	83.2
uz	761	99.4 ± 0.2	97.7 ± 0.4	93.9 ± 0.8	95.5 ± 0.6	96.5

Interestingly, our model obtained worse coverage in the cases eight cubes were used for 3-letter words (90.5%) than 4–5-letter words (93% and 91.5%) on the *es* dataset. The key fact is that the *es* dataset has three times more four-letter words and seven times more five-letter words compared to three-letter words. Another reason is that the Spanish language has more morphological word forms.

Table 6 shows the overall coverage results of our model for the case of eight cubes designed to accommodate six- to seven-letter words. A general decline is seen as word

length increases, which is an expected behavior, with seven-letter words achieving lower coverage across all datasets, particularly *kz*, *tt* (43.3%), and *pl* (45.9%). The experimental results reveal significant differences across languages, with some datasets, such as *en*, *ms*, *fr*, *sl*, and *es*, exhibiting relatively high total coverage (above 75%), while other languages, such as *kz*, *pl*, and *tt*, show lower overall coverage (below 55%).

Table 6. Overall coverage (%) with the standard deviation for the ACG model in the case of 8 cubes with 6–7-letter words.

Dataset	# of Generated Cubes	6 Letters	7 Letters	Avg Total	Max Case
de	790	73.2 ± 2.1	52.9 ± 2.6	61.4 ± 2.3	67.7
en	720	84.9 ± 1.6	67.5 ± 2.4	75.4 ± 2.0	79.7
es	740	86.7 ± 1.6	70.1 ± 2.5	76.8 ± 2.1	82.2
fr	687	86.1 ± 0.9	75.8 ± 1.7	80.2 ± 1.3	82.8
kz	792	64.5 ± 1.4	43.3 ± 2.0	52.4 ± 1.7	56.9
ms	728	88.1 ± 1.7	72.1 ± 2.3	79.0 ± 2.0	85.8
pl	791	65.9 ± 1.6	45.9 ± 2.0	54.5 ± 1.8	58.3
ru	797	70.3 ± 1.1	48.8 ± 1.6	58.3 ± 1.3	61.4
sl	750	87.3 ± 1.7	71.5 ± 2.6	78.7 ± 2.2	83.6
tr	770	73.4 ± 1.3	49.7 ± 2.0	59.3 ± 1.7	62.9
tt	773	63.7 ± 1.1	43.3 ± 1.8	52.0 ± 1.4	54.7
uz	761	84.9 ± 1.5	66.4 ± 2.3	74.7 ± 1.9	77.8

It can be seen from Table 7 that the ACG model's performance was reasonably high across all datasets. ACG achieved over 75% of Max case on all datasets except *kz*, and these percentages were even higher on the *es*, *fr*, and *sl* datasets (over 90%). Our model generated a slightly lower number of cubes on the *fr* dataset (921 cubes), which means most of the generated interchanging letter cubes were the same on this dataset.

Table 7. Overall coverage (%) with the standard deviation for the ACG model in the case of 9 cubes with 6–7-letter words.

Dataset	# of Generated Cubes	6 Letters	7 Letters	Avg Total	Max Case
de	996	75.2 ± 3.1	58.5 ± 3.8	65.5 ± 3.5	76.0
en	977	87.3 ± 1.4	71.2 ± 2.8	78.6 ± 2.2	84.8
es	971	94.0 ± 1.2	83.7 ± 2.3	87.8 ± 1.8	92.4
fr	921	90.5 ± 1.0	84.7 ± 1.9	87.2 ± 1.5	90.3
kz	1014	80.7 ± 1.1	64.3 ± 1.9	71.5 ± 1.5	74.0
ms	986	89.9 ± 1.9	75.5 ± 3.2	81.7 ± 2.6	88.3
pl	1003	88.0 ± 0.8	76.6 ± 1.3	81.5 ± 1.1	84.3
ru	1023	86.5 ± 1.1	74.6 ± 1.8	79.8 ± 1.4	83.2
sl	974	94.7 ± 0.7	86.1 ± 1.4	90.0 ± 1.1	92.8
tr	987	83.9 ± 1.1	66.9 ± 1.6	73.8 ± 1.4	76.4
tt	1001	80.6 ± 1.0	68.5 ± 1.6	73.7 ± 1.3	75.4
uz	991	92.2 ± 1.2	78.9 ± 2.6	84.8 ± 1.9	88.7

In general, the ACG model acquired the intended coverage in all datasets in the case eight cubes were used for 3–5-letter words and nine cubes for 6–7-letter words. This was the expected behavior because the chance of constructing 3–5-letter words from eight cubes (containing 48 letters) and 6–7-letter words from nine cubes (containing 54 letters) became really high.

The total coverage of the ACG method for 3–5-letter words using seven and eight cubes are presented in Figure 3.

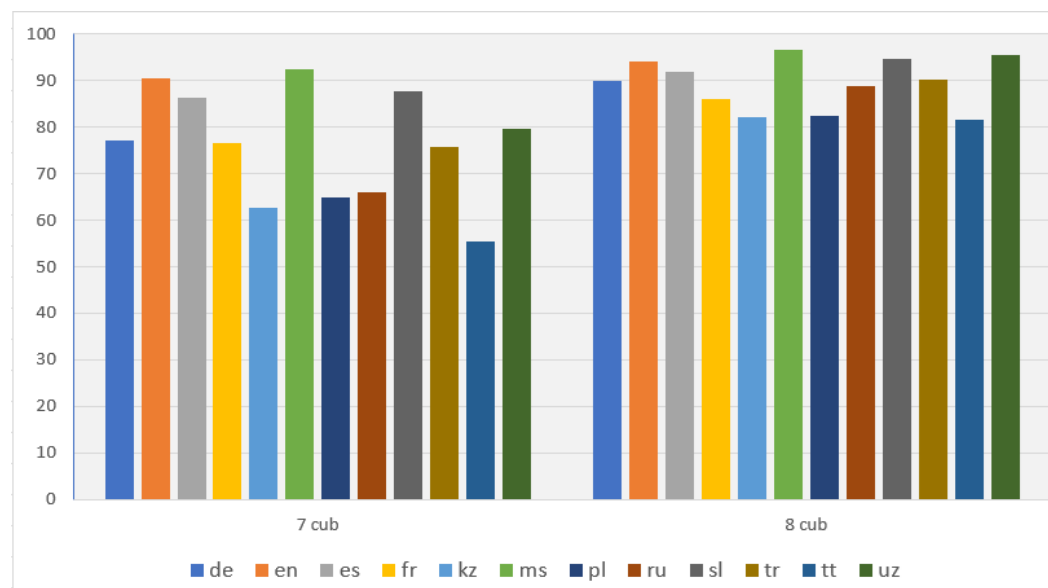


Figure 3. Comparison of the total coverage for 3–5-letter words using seven and eight cubes.

Figure 3 illustrates that significant improvement is observed in constructing 3–5-letter words using seven and eight cubes in terms of total coverage on the *kz*, *ru*, *pl*, *tt*, and *uz* datasets. Similar improvements are obtained for 6–7-letter words using eight and nine cubes, as shown in Figure 4.

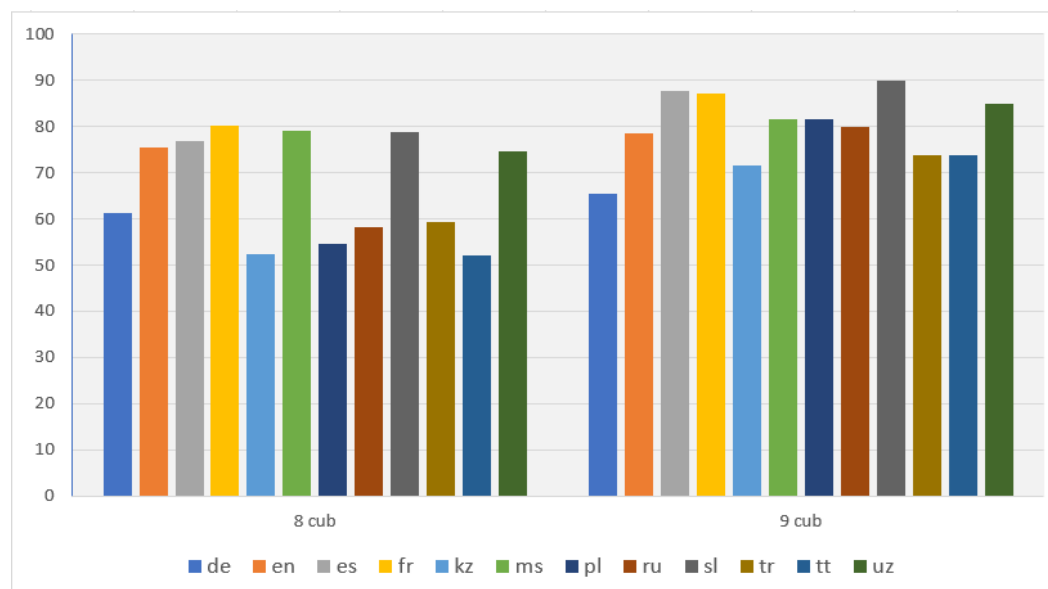


Figure 4. Comparison of total coverage for 6–7-letter words based on eight and nine cubes.

The training time was measured during the experiment to verify the temporal complexity of the suggested models, as indicated in Table 8. The model's training process was conducted on a computing system equipped with an Intel Core i5-1135G7 processor and 8GB DDR4 RAM.

Table 8. Overall training times in seconds for the proposed model.

Dataset	6–7 Letters		3–5 Letters	
	9 Cubes	8 Cubes	8 Cubes	7 Cubes
de	0.6962	0.5316	0.3197	0.2798
en	0.6420	0.4042	0.3260	0.2737
es	0.5455	0.4413	0.3214	0.2885
fr	0.5389	0.6160	0.3962	0.4126
kz	0.7133	0.6570	0.4005	0.5063
ms	0.7697	0.2471	0.1608	0.2819
pl	0.7588	0.5665	0.3090	0.1833
ru	0.6798	0.6512	0.4210	0.4813
sl	0.6780	0.5591	0.3924	0.3796
tr	0.5905	0.5379	0.3469	0.2871
tt	0.8481	0.8310	0.5322	0.3065
uz	0.7402	0.4367	0.3263	0.2874

5. Discussion of Results

The experimental results showed that the ACG model achieved reasonably high coverage across various languages. The coverage of the ACG model significantly improved with an increase in the number of cubes, which is an expected behavior. The proposed model obtained the intended coverage with eight cubes for 3–5-letter words and nine cubes for 6–7-letter words, which were selected as the optimal number of cubes in all datasets. The ACG model attained over 95% of *Max_coverage* on the Malay, Sloven, and Uzbek language datasets to construct 3–5-letter words with eight cubes. For the case of 6–7-letter words using nine cubes, our model achieved over 90% of coverage in Spanish, French, and Slovenian. The ACG model obtained worse coverage in constructing seven-letter words with eight cubes in Kazakh (43.3%), Tatar (43.3%), and Polish (45.9%). The main reason is that those languages have more letters in their alphabet compared to other languages. It is recommended to have more cubes in Kazakh, Tatar, and Polish to increase the coverage.

In [13], the MLG model was developed only for Uzbek, English (96.8%), Russian (84.1%), and Slovenian (94.2%). All languages are fusional except Uzbek, which is an agglutinative language, and these datasets included 3–5-letter words with non-morphologically formed words. We also used the same datasets for English (94.0%), Russian (88.7%), and Slovenian (94.8%) and achieved higher results in our present study.

We have elaborated on how different morphological structures—particularly in agglutinative versus fusional languages—affect word formation and model performance. Unlike fusional languages, agglutinative languages form words by attaching a series of affixes to a root, resulting in a high frequency of suffixes in morphologically complex words. Our analysis shows that suffixes appear less frequently in shorter words (3–5 letters), which leads to higher model accuracy for agglutinative languages such as Turkish, Kazakh, Tatar, and Uzbek (in Table 5). However, in longer words (6–7 letters), suffix usage becomes more prevalent, and agglutinative languages begin to exhibit complex morphological phenomena like allomorphy. These changes introduce variability, which affects the model’s ability to generalize, as seen in the performance differences shown in Table 7 for the same set of languages.

In exploring the statistical underpinnings of word game design across languages, our study leverages the associative networks to enhance gameplay and learning outcomes, aligning with broader research on semantic structures in language acquisition. Cox and Haebig (2023) [27] demonstrate that child-oriented word associations, elicited from adults imagining interactions with toddlers, significantly improve models of early lexical growth compared to unconstrained adult-oriented associations. Their findings reveal that child-oriented responses—characterized by simpler, higher frequency words with earlier ages

of acquisition are better at predicting vocabulary development, reflecting a semantic environment tailored to young learners. In contrast, our work extends beyond acquisition to application, utilizing cross-linguistic statistical patterns to design word games that engage diverse players; however, this similarly relies on associative networks to capture meaningful linguistic relationships. While Cox and Haebig focus on developmental modeling, our approach highlights the practical utility of such networks in interactive contexts, suggesting a complementary interplay between semantic structure and gamified learning.

6. Conclusions and Future Work

Since the matching letter game is an essential tool for teaching words to children and improving their vocabulary, a novel (ACG) model was developed to design an MLG based on the statistical analysis of letters. In this study, participants are given a set of cubes and tasked with forming the maximum number of words from the language. The testing process was conducted using sufficiently large datasets of words in different languages. The experimental results demonstrate the significance of our model by achieving relatively high coverage in all languages.

ACG obtained 94% average coverage over 12 datasets for 3-letter words, 94.4% average coverage for 4-letter words, and 87.4% average coverage for 5-letter words with eight cubes, which is selected as the optimal case for constructing 3–5-letter words. For 6–7-letter words, we selected nine cubes as an optimal case because the proposed model achieved relatively high coverage (86.6% average coverage over 12 datasets for 6-letter words and 74.1% for 7-letter words) in that case. The ACG model is easily interpretable and can be applied in any language. Our model obtained great results in terms of total coverage (over 95% for 3–5-letter words with eight cubes and around 85% for 6–7-letter words with nine cubes) in the newly created Uzbek dataset.

Our evaluation shows that our model performs well with languages that have complex morphological structures. In the future, we plan to improve and adapt ACG for other agglutinative languages, particularly those from the Turkic family.

Author Contributions: Conceptualization: U.S., J.M. and B.K.; methodology: U.S. and J.M.; software: U.S.; validation: J.M. and U.S.; formal analysis, J.M., U.S. and B.K.; writing—original draft preparation: J.M. and U.S.; writing—review and editing: J.M. and B.K. All authors have read and agreed to the published version of the manuscript.

Funding: Jamolbek Mattiev acknowledges funding by the Agency for “Innovative Development” of the Republic of Uzbekistan; grant: UZ-N39. The first and third authors acknowledge the Slovenian Research Agency ARRS for funding under project J2-2504. They also gratefully acknowledge the European Commission for funding the InnoRenewCoE project (Grant Agreement #739574) under the Horizon2020 Widespread-Teaming program and the Republic of Slovenia (Investment funding of the Republic of Slovenia and the European Union of the European Regional Development Fund).

Data Availability Statement: The dataset used in this paper can be found on <https://github.com/UlugbekSalaev/MatchingLetterGame>, accessed on 14 February 2025.

Conflicts of Interest: The authors declare no conflict of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ACG	Advanced Cubic-oriented Game
MLG	Matching letter game
GLF	Generation of Letter Frequencies

EIL	Elimination of Infrequent letters
GPLC	Generation of Potential Letter Cubes
de	German
en	English
es	Spanish
fr	French
kz	Kazakh
ms	Malay
pl	Polish
ru	Russian
sl	Slovenian
tr	Turkish
tt	Tatar
uz	Uzbek

References

1. Viera, R.T. Vocabulary knowledge in the production of written texts: A case study on EFL language learners. *Rev. Tecnol. ESPOL (RTE)* **2017**, *30*, 89–105.
2. Alqahtani, M. The importance of vocabulary in language learning and how to be taught. *Int. J. Teach. Educ.* **2015**, *3*, 21–34. [\[CrossRef\]](#)
3. Azar, A.S. The Effect of Games on EFL Learners' Vocabulary Learning Strategies. *Int. J. Basic Appl. Sci.* **2012**, *1*, 252–256. [\[CrossRef\]](#)
4. Rohani, M.; Pourgharib, B. The Effect of Games on Learning Vocabulary. *Int. J. Basic Appl. Sci.* **2013**, *4*, 3540–3543.
5. Alavi, G.; Gilakjani, A.P. The Effectiveness of Games in Enhancing Vocabulary Learning among Iranian Third Grade High School Students. *Malays. J. ELT Res.* **2019**, *16*, 1.
6. Najjar, M.; Masri, A. The Effect of Using Word Games on Primary Stage Students' Achievement in English Language Vocabulary in Jordan. *Am. Int. J. Contemp. Res.* **2014**, *4*, 144–152.
7. Bakhsh, S. Using Games as a Tool in Teaching Vocabulary to Young Learners. *Engl. Lang. Teach.* **2016**, *9*, 120. [\[CrossRef\]](#)
8. Huyen, N.; Nga, K. Learning Vocabulary Through Games. *Asian EFL J.* **2003**, *5*, 4.
9. Uberman, A. The use of games for vocabulary presentation and revision. *Forum* **1998**, *36*, 20–27.
10. Shchukina, T.J.; Mardieva, L.A.; Alyokine, T.A. Teaching Russian Language: The Role of Word Formation. In *Teacher Education-IFTE 2016, Volume 12. European Proceedings of Social and Behavioural Sciences*; Valeeva, R., Ed.; Future Academy: Kazan Federal University: Kazan, Russia, 2016; pp. 190–196. [\[CrossRef\]](#)
11. Whitney, C. How the brain encodes the order of letters in a printed word: The SERIOL model and selective literature review. *Psychon. Bull. Rev.* **2001**, *8*, 221–243. [\[CrossRef\]](#) [\[PubMed\]](#)
12. Aristides, V.; Monica, G.; Maria, F. Christos, T. Utilizing NLP Tools for the Creation of School Educational Games. In *Educating Engineers for Future Industrial Revolutions. ICL 2020. Advances in Intelligent Systems and Computing*; Auer, M.E., Rüttmann, T., Eds.; Springer: Tallinn, Estonia, 2020. [\[CrossRef\]](#)
13. Mattiev, J.; Salaev, U.; Kavsek, B. Word Game Modeling Using Character-Level N-Gram and Statistics. *Mathematics* **2023**, *11*, 1380. [\[CrossRef\]](#)
14. Salaev, U. UzMorphAnalyser: A morphological analysis model for the Uzbek language using inflectional endings. *AIP Conf. Proc.* **2024**, *3244*, 030058. [\[CrossRef\]](#)
15. Zaitun, M.; Fitri, A.J.E. Big Cube Game: An Instructional Medium Used in Students' Vocabulary Mastery. *J. Engl. Lit. Educ.* **2020**, *7*, 101–106.
16. Anadón, X.; Sanahuja, P.; Traver, V.J.; Lopez, A.; Ribelles, J. Characterising Players of a Cube Puzzle Game with a Two-Level Bag of Words. In *Proceedings of the 29th ACM Conference on User Modeling, Adaptation and Personalization (UMAP '21)*, Utrecht, The Netherlands, 21–25 June 2021; pp. 47–53. [\[CrossRef\]](#)
17. Jiang, L. Word Data Prediction Based on Statistical Method. *Trans. Comput. Sci. Intell. Syst. Res.* **2024**, *5*, 1662–1670. [\[CrossRef\]](#)
18. Vu, N.N.; Linh, P.T.M.; Lien, N.T.H.; Van, N.T.T. Using Word Games to Improve Vocabulary Retention in Middle School EFL Classes. In *Proceedings of the 18th International Conference of the Asia Association of Computer-Assisted Language Learning (AsiaCALL-2-2021)*, *Advances in Social Science, Volume 621, Education and Humanities Research*, Ho Chi Minh City, Vietnam, 26–27 November 2021; pp. 97–108.
19. Anugerah, R.; Wijaya, B.; Bunau, E. The use of build-a-sentence cubes game in teaching simple past tense. *J. Pendidik. Dan Pembelajaran Khatulistiwa (JPPK)* **2016**, *5*. [\[CrossRef\]](#)
20. Madvaliyev, A.; Begmatov, E. *O'zbek Tilining Imlo Lug'ati*; Mahmudov, N., Ed.; Akadem-nashr: Tashkent, Uzbekistan, 2012.

21. Goldhahn, D.; Eckart, T.; Quasthoff, U. Building Large Monolingual Dictionaries at the Leipzig Corpora Collection: From 100 to 200 Languages. In Proceedings of the 8th International Language Resources and Evaluation (LREC'12), Istanbul, Turkey, 21–27 May 2012.
22. Sharipov, M.; Mattiev, J.; Sobirov, J.; Baltayev, R. Creating a Morphological and Syntactic Tagged Corpus for the Uzbek Language. In Proceedings of the International Conference and Workshop on Agglutinative Language Technologies as a Challenge of Natural Language Processing, ALTNLP 2022, Koper, Slovenia, 6–8 June 2022; pp. 93–98.
23. Allaberdiev, B.; Matlatipov, G.; Kuriyozov, E.; Rakhmonov, Z. Parallel texts dataset for Uzbek-Kazakh machine translation. *Data Brief* **2024**, *53*. [[CrossRef](#)] [[PubMed](#)]
24. OpenCorpora: An Open Source Initiative for Building a Free and Comprehensive Corpora for Russian and Other Slavic Languages. Available online: <http://opencorpora.org/> (accessed on 15 October 2024).
25. Kaja, D.; Simon, K.; Peter, H.; Tomaž, E.; Miro, R.; Špela, A.H.; Jaka, Č.; Luka, K.; Marko, R.-Š. Morphological Lexicon Sloleks 2.0, Slovenian Language Resource Repository CLARIN.SI, ISSN 2820-4042. Available online: <http://hdl.handle.net/11356/1230> (accessed on 10 October 2024).
26. Jurafsky, D.; Martin, J.H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition with Language Models*, 3rd ed.; Pearson/Prentice Hall: Upper Saddle River, NJ, USA, 2009. Available online: <https://web.stanford.edu/~jurafsky/slp3/> (accessed on 15 January 2025).
27. Cox, C.R.; Haebig, E. Child-Oriented Word Associations Improve Models of Early Word Learning. *Behav. Res.* **2023**, *55*, 16–37. [[CrossRef](#)] [[PubMed](#)]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.