education sciences

IMPACT FACTOR 2.6

CITESCORE 5.5

Article

Closing the Gap: Automated Distractor Generation in Japanese Language Testing

Tim Andersson and Pablo Picazo-Sanchez

Special Issue

Language Education in the Digital Age: An International Perspective

Edited by

Prof. Dr. Cristina A. Huertas-Abril and Dr. Francisco J. Palacios-Hidalgo

MDPI

https://doi.org/10.3390/educsci13121203

education sciences

MDPI

Article

Closing the Gap: Automated Distractor Generation in Japanese Language Testing

Tim Andersson † and Pablo Picazo-Sanchez *,†

School of Information Technology, Halmstad University, 301 18 Halmstad, Sweden; timand18@student.hh.se
* Correspondence: ppicazo@hh.se
† These authors contributed equally to this work.

Abstract: Recent advances in natural language processing have increased interest in automatic question generation, particularly in education (e.g., math, biology, law, medicine, and languages) due to its efficiency in assessing comprehension. Specifically, multiple-choice questions have become popular, especially in standardized language proficiency tests. However, manually creating high-quality tests is time-consuming and challenging. Distractor generation, a critical aspect of multiple-choice question creation, is often overlooked, yet it plays a crucial role in test quality. Generating appropriate distractors requires ensuring they are incorrect but related to the correct answer (semantically or contextually), are grammatically correct, and of similar length to the target word. While various languages have seen research in automatic distractor generation, Japanese has received limited attention. This paper addresses this gap by automatically generating cloze tests, including distractors, for Japanese language proficiency tests, evaluating the generated questions' quality, difficulty, and preferred distractor types, and comparing them to human-made questions through automatic and manual evaluations.

Keywords: NLP; Japanese; cloze tests; education

check for updates

Citation: Andersson, T.; Picazo-Sanchez, P. Closing the Gap: Automated Distractor Generation in Japanese Language Testing. *Educ. Sci.* **2023**, *13*, 1203. https://doi.org/10.3390/educsci13121203

Academic Editors: Cristina A. Huertas-Abril and Francisco J. Palacios-Hidalgo

Received: 5 October 2023
Revised: 9 November 2023
Accepted: 24 November 2023
Published: 30 November 2023

Copyright: © 2023 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/).

1. Introduction

Thanks to the recent advances within the field of Natural Language Processing (NLP)—a sub-branch of artificial intelligence that concerns a computer's ability to interpret, manipulate, and comprehend human language—another field has also seen an emergence of studies and research: education. Specifically, the automatic generation of questions [1–3] has garnered a lot of interest in the last decade across a wide variety of educational fields. An efficient method for assessing comprehension, questions are an integral tool for any education setting that benefit both learners and educators [4,5]. We can find automatic question generation in maths [6,7], biology [8], law [9], medicine [10], and languages [11,12].

The generation of Multiple-Choice Question (MCQ) is a topic which has become popular within automatic question generation and language learning. These MCQs are widely used in standardized language proficiency tests, such as the *Test of English for International Communication (TOEIC)* (https://www.ets.org/toeic.html (accessed on 1 November 2023)), *Diplomas de Español como Lengua Extranjera (DELE)* (https://www.dele.org/ (accessed on 1 November 2023)), and the *Japanese Language Proficiency Test (JLPT)* (https://www.jlpt.jp/e/ (accessed on 1 November 2023)). However, manually generating MCQs is an arduous and time-consuming task, and creating "good" questions requires experience and resources [13–15].

Within MCQs, the most widespread type of test is the *multiple choice cloze test* [16], more commonly known as the "fill-in-the-blank" test, in which one word in a sentence is replaced with a blank space that students must fill in [17]. Writing cloze tests is challenging since generating easy or wrong *candidate distractors* may cause the tests to be low

quality [16]. For that reason, distractor generation is among the most challenging parts of MCQ generation [18] and, as far as we know, only a few studies have focused on this topic [19,20]. Strategies like using random words from the same context in which the original question sentence was chosen [21], selecting synonyms to the target word from a thesaurus regarding frequency [22], and dictionary-based collocation [23] are just a few examples of how researchers have proposed generating distractors.

Nevertheless, these methods alone do not necessarily produce adequate distractors since there are specific requirements distractors should satisfy [3,4]: (1) Be an incorrect answer to the question; (2) Be related to the correct answer, semantically or from the same category (i.e., nature, color, exercise); (3) Be grammatically correct and consistent with the difficulty of the correct answer, and; (4) Be of the same word class and a similar length to the target word.

To this end, researchers have proposed automatic distractor generation for many of the major languages of the world, e.g., Chinese Mandarin [11], Hindi [24,25], French [12], and English [3,19]. However, to our knowledge, Japanese lacks these types of studies. In this paper, by using Natural Language Processing (NLP) and adopting the official JLPT format, we automatically generate cloze tests and the distractors for the tests the official exam includes. Our goal is to investigate whether state-of-the-art NLP methods can be used to generate Japanese distractors and cloze tests, as well as to build the foundation from which future research can benefit. In more detail, we aim to address four research questions:

RQ1: Are our generated distractors indistinguishable from human-made distractors?

RQ2: Can we generate JLPT-level appropriate distractors?

RQ3: Can we use NLP methods to attach a valid difficulty rank to generated questions?

RQ4: Is there a preferred distractor type?

To address these questions, we use NLP to automatically generate cloze tests for three learning outcomes: kanji reading, kanji orthography, and vocabulary. Furthermore, we rank the questions based on the difficulty of the distractors. Finally, we evaluate our results in two ways: automatically and manually. We use NLP to get the difficulty of a cloze test for the automatic evaluation. For the manual evaluation, we asked professional Japanese teachers to evaluate the automatic cloze test questions through a questionnaire, similar to other works in this field [10,25,26].

2. Background

This section presents the concepts and terminology needed to understand the rest of the paper. In detail, we introduce NLP and some of the most important methods, the Japanese Language and some linguistic terminology, and explain how the official JLPT exam is structured, and the essential parts required to make cloze tests.

2.1. Natural Language Processing

NLP is a sub-branch of artificial intelligence that concerns a computer's ability to interpret, manipulate, and comprehend human language. The field of NLP has its roots in the Georgetown-IBM experiment from the 1950s, where researchers could automatically translate Russian to English [27]. NLP has since significantly improved and grown into areas such as text prediction for auto-correction, writing assistance, translation between languages, and chatbots. In this project, we use NLP methods and concepts that we will briefly explain in the following sections.

Word2vec is a method built to understand semantic relations between words similarly to how humans understand the relations between words, e.g., "king" and "queen". Since computers struggle with understanding human language, the method uses word embeddings, a vectorized representation of a word. Using "queen" as an example, the algebraic equation could look like $king - man + woman \approx queen$. Note that the result would not necessarily equate to "queen" but a vector close to that word.

Levenshtein distance measures the distance between two strings (*a* and *b*) by counting the number of operations required to transform string a into string b. The result is a numerical representation of the difference between the two strings, where a 0 represents that the two strings are the same.

N-gram is an uninterrupted sequence of words or tokens in a document. The "N" represents the number of splits the N-gram makes when fed a string. For instance, a sentence like "I enjoy playing games" would produce four tokens when N = 1, three when N = 2, two when N = 3, and one when N = 4 (see Table 1).

Table 1. Example of different N-gram tokens of the "I enjoy playing games" sentence.

N	N-Gram Type	N-Gram
1	Unigram	"I", "enjoy", "playing", "games"
2	Bigram	"I enjoy", "enjoy playing", "playing games"
3	Trigram	"I enjoy playing", "enjoy playing games"
4	Fourgram	"I enjoy playing games"

Bidirectional Encoder Representations from Transformers (BERT) is a language model for NLP that uses a bidirectional training of transformers to handle long-distance dependencies [28]. The model was released in 2018 by Google and has become ubiquitous within the NLP community. Only two years after its inception, over 150 studies had used BERT models in their work [29].

Masked Language Model (MLM) is a language model inspired by the *cloze task*, which predicts a missing word from a sentence. The missing word is often represented by the [MASK] token, giving the model its name. The [MASK] token focuses the model on a single word in a sentence and uses the surrounding words for context. The model then returns possible words and scores that fit into the given sentence. The MLM is a bi-product of creating the BERT model, since the creators needed a method to train the model. The training method revolved around the random removal of 15% [28] of words in a training document. This allowed the model to consider surrounding words and learn more about the context of words thanks to the bidirectional access the BERT model allows.

2.2. The Japanese Language

Japanese is a Subject-Object-Verb (SOV) language, whereas English is a Subject-Verb-Object (SVO) language. It means that in Japanese, the emphasis of a sentence is commonly in the final word. This can sometimes cause oddities when directly translating between English and Japanese (first row of Table 2). However, Japanese is flexible regarding word order and the omission of words entirely (second row of Table 2). For example, in Japanese, one can omit the subject of a sentence and the sentence will still make perfect sense in context. Also, Japanese has no spaces between words. Therefore, it can be problematic for non-native speakers to see where a word ends and one begins when reading Japanese text.

Throughout this paper, we provide rough translations of Japanese words or sentences used as examples. We use the Revised Hepburn Romanization system [30] to translate to Roman alphabetization.

Table 2. Examples of Japanese sentences, subject omission and a hiragana-only sentence.

Type	Japanese	Translation
Base	彼はボールを投げる	He is throwing a ball
Subject omitted	ボールを投げる	Throwing a ball
Hiragana	ははははながすきだ	My mother likes flowers

2.2.1. Writing System

The Japanese writing system consists of three alphabets, *Hiragana*, *Katakana*, and *Kanji*.

Hiragana is a phonetic alphabet build on sound compounds which each represents one syllable. Hiragana is the first writing system both Japanese children and Japanese as a Second Language (JSL) beginners learn, and while one can construct complete Japanese sentence using only hiragana, doing so risks making sentences ambiguous (see the last row of Tables 2 and A3). The use-cases of hiragana are for any words that are not written in kanji, any particles binding together a sentence, and the grammatical endings for verbs, nouns, adjectives.

The *katakana* alphabet is built similarly to the hiragana alphabet but is characterized by its sharp edges. Katakana mainly appear in loan words, non-Japanese names and in Japanese comics, for example as *Yakuwari-go* [31] (character language) to represent a certain type of character archetype. For the sake of simplicity, in this paper, when we present a katakana word, it will either be a loan word or a non-Japanese name (see Table A1 in Appendix A for some examples).

Kanji was introduced into Japanese in the 5th century from the Chinese writing system. The characters are built on elements or radicals that produce meaningful words and can appear alone or in compounds to create different words (see Table A2 in Appendix A for examples of elements and radicals).

There is a predetermined way to write each kanji, usually starting from the top left and ending in the bottom right. The stroke order and count (order and amount of lines used to write a kanji) are also essential to understand since those are the two ways to assist when trying to find a kanji in a Japanese dictionary. The primary function of kanji is to combine the sounds represented by hiragana into words, adding meaning to the sounds and reducing the number of characters in a sentence (see Table A3 in Appendix A).

2.2.2. Japanese Grammar

In Japanese, different types of adjectives and verbs direct how a word is conjugated. The adjectives are *i-adjectives* and *na-adjectives*, while verbs are *ichidan*, *godan*, and *irregular* verbs. Only a handful of irregular verbs conjugate differently from the norm but are prevalent in everyday Japanese. Japanese also use a simple tense system that only covers past and present tenses. The person has no plural form or conjugation (as shown in Table 2). These peculiarities lead to the conjugation of words becoming a central pillar in grammar.

Politeness is also culturally significant in Japanese as there are grammatical conjugations and its own Part-of-Speech (POS) to adjust the politeness of a word. Sometimes, words with the same meaning differ due to this politeness. An example is the verb "to eat". The dictionary form is 食べる (taberu). However, when using polite speech, it turns into 食べます (tabemasu) followed by 召し上がる (meshiagaru), used when speaking to superiors or customers. Next is いただく (itadaku), referring to the action of oneself while speaking to a superior or customers. The meaning stays the same (to eat) for every word mentioned, but this change in politeness can cause issues with translations without context.

2.3. Linguistic Terminology

To clarify some concepts about linguistics, we include the definitions of synonyms, homonyms, phono-semantic compounds, and POS as follows.

Synonym is a word or phrase that means the same as another word or phrase. The Japanese language has many synonyms since it uses many loan words from other languages, mainly Chinese and English. We can find synonyms between languages like Native 車 (kuruma), Sino-Japanese 自動車 (jidōsha), and Western カー (kaa), which all translate to "car" [32] as well as full-native synonyms such as 話す (hanasu) and 喋る (shaberu) which both mean *to talk*.

Homonym is a word that does not share the same meaning as another word but is written or pronounced the same. When words are written the same, they are called *homographs*; when the pronunciation is the same, they are called *homophones*.
Some English examples of *homographs* are "bat" (the animal and the one used in baseball), "letter" (a letter you send to someone, and the one of the alphabet). Examples

of *homophones* is "one" and "won", "two", "to" and "too". In Japanese, *homographs* are less common thanks to kanji in the written language. Some kanji have different meanings (which have different pronunciations) depending on the context, such as 金, "gold" and "money", and 月, "month" and "moon". However, since Japanese has fewer sounds than English, *homophones* are far more prevalent [33]. Words with entirely different meanings are written similarly, but thanks to kanji, meaning can more easily be conveyed. Examples of Japanese *homophones* are 会う (au, to meet), 合う (au, to fit), 感じ (kanji, feeling), 漢字 (kanji, Chinese characters).

Phono-semantic compounds Kanji is made of compounds of elements, radicals, and sometimes a combination of other kanji. The different parts of a kanji can often be split into two parts: the *phono*, the sound of the character, and the *semantic*, the meaning of the character.

Part-of-speech (POS) is used to categorize and classify words according to their function in a sentence. Using the Universal Part-of-Speech Tagset as an example, there are 12 tags words that can be divided into [34]: (i) NOUN (nouns); (ii) VERB (verbs); (iii) ADJ (adjectives); (iv) ADV (adverbs); (v) PRON (pronouns); (vi) DET (determiners and articles); (vii) ADP (prepositions and postpositions); (viii) NUM (numerals); (ix) CONJ (conjunctions); (x) PRT (particles); (xi) '.' (punctuation marks), and; (xii) X (a catch-all for other categories such as abbreviations or foreign words).

2.4. The Japanese Language Proficiency Test

The JLPT (日本語能力試験 Nihongo Nōryoku Shiken) is a language criterion test created by the Japan Foundation (JF) and Japan Educational Exchanges and Services (JEES) in 1984. The test aims to measure the levels of language proficiency for non-native speakers through testing language knowledge (split into vocabulary and grammar), reading, and listening abilities. The test is a worldwide occurrence twice yearly and has seen a steady increase of examinees over the years (Reduced participation in 2020 due to the spread of COVID-19). In 2019, there were 1.36 million applicants and 1.16 million examinees worldwide [35]. Once a student passes the exam, they receive an official certificate that never expires. It can be used as proof of proficiency and in employment screening and evaluation for promotions and pay raises [36]. Thus, the JLPT has become integral for any foreigner who wishes to integrate into Japanese society.

There are currently five levels to the JLPT, starting from N5 (the lowest level) and going up to N1. The *N* before the number indicates the *New* JLPT [37] since the test changed in 2010. The JLPT is divided into "vocabulary", "grammar", "reading", and "listening" abilities. The N1 and N2 test levels are split into two sections where vocabulary, grammar, and reading abilities belong to the first, whereas "listening" belongs to the second. The remaining three test levels are instead divided into three sections, with the first containing the vocabulary ability. The second section is for grammar and reading, and the last is for listening ability (see Table 3).

Table 3. Structure of the different sections in the JLPT [38].

Level	Test Sections and Abilities (Time)		
N1	Vocabulary, Grammar, Reading (110 min)		Listening (55 min)
N2	Vocabulary, Grammar, Reading (105 min)		Listening (50 min)
N3	Vocabulary (30 min)	Grammar, Reading (70 min)	Listening (40 min)
N4	Vocabulary (25 min)	Grammar, Reading (55 min)	Listening (35 min)
N5	Vocabulary (20 min)	Grammar, Reading (40 min)	Listening (30 min)

Throughout the test, and in all blocks, students can expect cloze test-type questions where they must select the correct answer to a question from *four* choices. Only a single answer will be correct, making the answer unambiguous.

However, as it was pointed out [39], in JLPT: (1) it is hard to choose the appropriate difficulty level; (2) there is no way of evaluating communicative competence, and; (3) it

is a black-box process. Picking the appropriate difficulty level to focus on as a student is difficult and reflected in the low passing rates of certain JLPT levels [39]. Next, although the new JLPT claims to focus on communicative abilities [36], there are no speaking or writing tasks in the test, making it challenging to assess communicative competence. Lastly, the operators of the JLPT have reduced transparency regarding the test since they no longer release test content specifications and have reduced information about its procedures [40].

2.5. Cloze Tests

The standard cloze test has four parts: a stem, the keys, a target, and a focus area (see Figure 1). The *stem* is any sentence or paragraph previously containing the target but has been adjusted per the cloze test question type. The defining feature of a stem is that a *focus area* is present somewhere in the sentence or paragraph.

The *keys* are the choices given to complete a question. One of the choices is the correct answer, which completes the stem and makes the passage coherent. The wrong choices are known as *candidate distractors*, and these can fluctuate in number. Most commonly, the number of *candidate distractors* is between two to four [41,42].

The *target* is a word chosen to be used as the basis for *candidate distractor* creation. Depending on the type of cloze test question, the target acts as either the correct choice or as a hint to the correct choice among the keys.

The *focus area* is a highlighted area in a cloze test question, representing the area where the correct key will coherently complete the stem. The focus area may visually differ depending on the cloze test question type. The area may still have the target present in the stem, or it may be removed or changed in some manner. We include some examples of the types of cloze test questions in Japanese in Table 4.

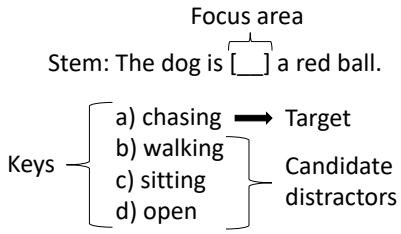

Figure 1. Example of a cloze test in English. The sentence containing the focus area is the stem. The keys contain three distractors (options b, c, and d) and the correct answer is called target a) which would make the stem coherent.

Table 4. Three examples of focus areas (hiragana, leave-in, and empty) together with the stem example and the types of possible distractors for every learning outcome.

Learning Outcome	Focus Area Type	Focus Area Example	Possible Distractor Types
Kanji Reading	Leave-in	**[勉強]**	Synonyms, Homonyms, Hiragana
Orthography	Hiragana	**[べんきょう]**	Synonyms, Homonyms, Phono-semantic
Vocabulary	Empty	**[___]**	Synonyms, Homonyms, Part-of-speech

3. Dataset

Very little official data are available for the current JLPT since the operators keep most things regarding the test a secret [40]. Some official vocabulary, kanji, and grammar-point lists are available from before the JLPT changed into its modern form in 2010. However, there is a multitude of community-compiled online information regarding the JLPT. We use this unofficial information alongside old official information to cross-reference our data.

More concretely, in this paper, we use the website "nihongokyoushi-net" [43] as a data source. We use the grammar section of the website to collect all example sentences for each grammar point. As for the base of our vocabulary lists, we use the information from another website, "tanos" [44] as it is the same source used by one of the most extensive online Japanese dictionaries, "jisho" [45], for their information about the JLPT.

We also include any official data we can access, available on the JLPT website [46]. The questions we have access to from the website were published before the modernization of the JLPT in 2010 as an introduction to the revision of the new test. We use these sample questions as a benchmark when testing our generated questions. In total, we have access to 42 cloze test questions throughout all five of the JLPT levels.

4. Automatic Cloze Test Generation

This section presents the pipeline we use to generate distractors and cloze test questions. Since there are no prior ways to generate either distractors nor test questions in Japanese automatically, we first explore different distractors generation methods and create cloze tests for three learning outcomes: kanji reading, kanji orthography, and vocabulary. In the following, we explain every learning outcome in detail, while in Table 4, we include a summary of the possible distractor types based on the learning outcome.

Kanji reading 漢字読み (kanji yomi). A kanji reading test is a specialized test for the Japanese language since it relies on the swapping between different alphabets to create the distractors. The target in this question type must be a level-appropriate kanji, and the test uses the "Leave-in" focus area when creating the stem since the focus lies on the examinee's ability to know the reading of the target kanji. The possible candidate distractors are synonyms, homonyms, and hiragana distractors.

Orthography 表記 (hyouki). The kanji orthography test is similar to the kanji reading test, as it also relies on the swapping between alphabets. However, the focus area uses the "Hiragana" question type where the examinees are expected to read the target hiragana word and, through the context of the remaining stem, select the correct kanji provided in the keys. The possible candidate distractors are synonyms, homonyms, and phono-semantic distractors.

Vocabulary 語彙 (goi). A standard vocabulary test which does not have any preferences or limitations on what words can be used as a target as long as they are a part of the level-appropriate JLPT vocabulary list. The focus area for this question type uses the "empty-style", and its possible distractors are synonyms, homonyms, and POS.

Overview. To generate cloze tests, we first set the learning outcome and the JLPT level. Next, we automatically generate a sentence for the appropriate level using our corpus of Japanese sentences which we put together from the datasets we presented in Section 3. We then scan the sentence for a potential *target* by cross-referencing each word with a level-appropriate vocabulary list and randomly selecting one of the possible choices. Note that we want every word to have an equal possibility of becoming a target word since every possible target is relevant to the given JLPT level. However, we do not allow higher-level words to appear in the lower-level questions.

Once we have a target, we create the *stem* by extracting the target, adding a focus area and generating the candidate distractors. For the stem, we transform the target word of the original sentence in one of three ways based on the learning outcome: (1) Leave the word in and add brackets around the target to indicate focus area (see the first row of Table 4); (2) Change the target into another alphabet and add brackets around the word to indicate focus area(see the second row of Table 4); or (3) Extract the target and replace it with an empty box ([___]) to indicate where the word used to be (see the third row of Table 4).

4.1. Distractor Generation

There are a multitude of ways to generate distractors, such as generating random words [21], synonyms [22,47], homonyms [47], and dictionary bases collocation [23].

We used five different distractor types in our work over all of our learning outcomes. We use synonyms and homonyms distractors in all learning outcomes because they make functional distractors for most languages [48] and because Japanese is a morphologically rich language [47]. Next, we have three specialized distractors, which only appear in a single learning outcome each: (i) Hiragana distractors for the kanji reading questions;

(ii) Phono-semantic kanji distractors for the kanji orthography questions; and (iii) POS distractors for vocabulary questions.

The reasoning for picking these specific distractor types stem from that the hiragana and phono-semantic distractors are the distractor types that are used in the official JLPT tests for the respective question types. The POS distractor appears because we want to test how words that are essentially random, but with minor limitations, functions as distractors since this generation method is used in another work [47].

In the following paragraphs, we describe the distractors we consider in our work in more detail.

Synonym Distractor. *Synonyms* are a convenient way of generating distractors in most languages [47,48] while also being some of the most functional distractors [49]. We train a monolingual Japanese Word2vec model on the Japanese Wikipedia to produce our synonyms. Using a target word with the Word2vec model, we generate the top N synonyms and score pairs, which we then filter by saving only the words that appear in the desired JLPT levels vocabulary list. This filtering assures that we do not use inappropriate words and that we still have many words to use as distractors. The scores the method produces represent the closeness between the target and the generated word.

Homonym Distractor. In Japanese, homonyms (homophones specifically) can potentially produce many applicable distractors since the average homophony rate of Japanese is around 15%, which is very high considering that most languages have a homophony rate of around 4% [33]. This high rate means many words and kanji compounds are spelled similarly (see Table 5).

Table 5. Words that are spelled the same (hiragana column) but have a completely different meaning.

Kanji	Hiragana	Translation
漢字	**かんじ**	Chinese character
感じ	**かんじ**	Feeling
幹事	**かんじ**	Secretary

We generate homonym distractors using a Levenshtein distance since it is a string metric that does not require the compared words to be the same length [50]. We compare the target to our level-appropriate vocabulary lists and generate potential homonym distractors. We empirically set the distance between distractors to three since we noticed a considerable risk of not producing any valid distractors if the limit was lower than that limit. On the other hand, if the limit was higher than three the generated distractors would substantially differ from the target word.

When we create homonym distractors, we convert a word independently from the alphabet (hiragana or kanji compound) to katakana. We then convert the katakana word again to *Roman alphabetization* before measuring the distance. The reason is that words written using any of these alphabets tend to have a lower number of characters compared to when written in the Roman alphabetization. For example, the number of characters in Roman alphabetization of the words "kaisha" and "nihongo" is six and seven. They have four characters in hiragana (かいしゃ and にほんご, respectively), while in kanji, the number of characters is two (会社) and three (日本語), respectively. When calculating the Levenshtein distance, we prefer more variation (i.e., number of characters) for a better distinction between words.

Hiragana Distractor. Hiragana distractors are used in the kanji reading tests since the focus area of the stem still contains the target word, and the examinee needs to select the word's correct spelling from the keys. The goal is to have similar looking distractors so that the examinee can not simply guess the correct key by looking at the hiragana of the keys. That means that all distractors must be converted into hiragana and have limitations regarding what type of word we select as the target. The two types of target words that can appear in this question type are kanji compounds (学校—gakkō—school) or kanji with hiragana (踊る—odoru—to dance). In short, the idea is to have distractors of the same

length as the target in the case of a kanji compound target or to have words with the same grammatical form, which also contains the same number of hiragana as the target in the case of kanji and hiragana types.

Phono-semantic Distractor. We base the phono-semantic distractors on the fact that we can split most kanji into groups depending on their radicals or elements. This distractor type is used when generating distractors for the kanji orthography learning outcome, as the focus area of the stem contains the hiragana version of the target word, which is required to be a kanji or a kanji compound for this question type. To complete the stem, the examinee must select the correct kanji or kanji compound from the keys by using the context provided by the stem. The keys must, therefore, be similar so that the correct answer is not directly obvious. Note that the distractors produced by this method should look like the target word but do not necessarily need to be actual words. For example, the word 遅れる(okureru—to be late) could generate 達れる (no meaning) and 送れる (no meaning), which are not actual words but look like the correct word to the untrained eye. The same applies to kanji compounds like 予定 (yotei—plan; arrangement) which can generate 了定 (ryoutei?—no meaning) and 予足 (yosoku?—no meaning) as distractors.

Part-of-Speech Distractor. The POS distractor type only appears in our vocabulary cloze test. We generate these distractors by randomizing words of the same POS. More concretely, we create our POS distractors by scanning the appropriate JLPT vocabulary list and extracting N words of the same word class as the target.

Regardless of distractor type and generation method, there is one more issue to account for when making sure that tests are unambiguous: a distractor must be an incorrect answer to the question at hand. We use a public Japanese n-gram corpus [51] to compare trigrams, including all of our generated distractors. There are three versions of the corpus, each with differing frequencies of n-grams starting from ten or more, 100 or more, and 1000 or more occurrences of a given n-gram. We use the 100 or more frequency list in this work, which means that each time one of our distractor trigrams appears in the frequency list, we remove the distractor from the potential pool of usable distractors. This helps to improve the unambiguity of our generated tests.

4.2. Measuring the Difficulty

We propose an automatic algorithm that measures the difficulty of the generated tests. To this end, we use a combination of two scores: Word Score (WS) and Context Score (CS). The WS is a score assigned to each distractor and represents how close the distractor is to the word used to generate it. Each distractor has a WS, and it differs between distractor types. Synonyms use the score given by the Word2vec model as it represents the cosine similarity between words. Homonyms and hiragana distractors use the Levenshtein distance and we randomly assign a low score to POS distractors to avoid a random-type candidate adding significant weight to the difficulty algorithm. Lastly, the phono-semantic distractors use the difference in the number of strokes that make up a kanji as a score because the closer the number of strokes a candidate is to the target, the more likely the distractor kanji is to be built up of similar radicals and elements.

The CS is a score that represents how well a distractor fits into the context of a given sentence. We assign that score to each generated distractor and use it together with the WS to calculate the difficulty of a question. A customized state-of-the-art BERT fill-mask model generates this score. We restricted the softmax layers output to only include words we provide to the model, i.e., our distractors. The model attempts to fit the provided words into the stem and returns a number representing how well any distractor fits into the context of the stem. Note that a high context score does not equate to a suitable distractor since the higher the number, the higher the chance of that distractor being a valid answer to a cloze test, allowing for ambiguity in the test, which we must avoid.

Next, with the WSs, the CSs and the unwanted distractors filtered out, we normalize the values and attach each score to the respective distractor. Lastly, we assign a difficulty

rating to the finalized cloze test question by using the score of each distractor used to make up the complete cloze test question.

More formally, let $c_1, \ldots, c_n$ be a list of distractors. For every distractor, a tuple of WS and CS reflect a relation to the target word and how well the distractor fits into the stem (according to the BERT model), respectively. Given both the WS and CS, we compute the average ($\overline{wcs}_{c_n} = \frac{WS_{c_n} + CS_{c_n}}{2}$) for every distractor in the test. Finally, we get the difficulty_score as the average of the distractors average, i.e., difficulty_score $= \frac{\sum_{i=1}^{n} \overline{wcs}_{c_n}}{n}$.

Note that the resulting difficulty score is in a range between 0 and 1. Similar to other proposals in the field [14,52], we classified questions into five levels where *one* represents the easiest and *five* the most difficult (see Table 6).

Example 1. *Suppose we have three distractors c_1, c_2, and c_2. Whose WSs (WS_{c_1}, WS_{c_2}, and WS_{c_3}) are 0.6049, 0.5989, 0.5519 and their CS (CS_{c_1}, CS_{c_2}, and CS_{c_3}) are -0.5829, 1.6938, 0.8228, respectively. After normalizing the scores, computing their averages $\overline{wcs}_{c_1} = 0.3695$, $\overline{wcs}_{c_2} = 0.6202$, $\overline{wcs}_{c_3} = 0.5297$, and the difficulty score (difficulty_score = 0.5064), we conclude that the difficulty of the question is mid range.*

Table 6. Scoring difficulty range.

Range	Level	Difficulty
More than 0.80	5	Most Difficult
0.60–0.80	4	Difficult
0.40–0.59	3	Mid
0.20–0.39	2	Easy
Less than 0.20	1	Easiest

5. Evaluation

The cloze test evaluation is usually performed by experts in the field [53–55]. In this paper, we asked 14 experts (Japanese teachers) to evaluate the quality and the difficulty of the distractors we automatically generated through answering a questionnaire.

In that questionnaire, we used a mixture of our own automatically-generated questions and human-made questions from the "New Japanese-Language Proficiency Test Sample Questions". To unify the questions we performed the following changes: (i) Removed any spaces between words; (ii) Changed the focus area from "___" to "[___]"; (iii) Removed any furigana (smaller hiragana above a kanji which assist with reading); (iv) Reduced the target word (e.g., "[書いて]" to "[書い]て").

We made these changes because Japanese sentences do not usually include space between words and, in most cases, there is no furigana above a kanji. The reason for changing the focus area is a stylistic choice aimed at more clearly indicating the start and end of the extracted target word. Lastly, the reduction of the target words comes from a discrepancy between how Japanese is taught to natives versus JSL learners and which tokenizer one uses. Common English tokenizers separate words by white spaces, however since Japanese texts commonly do not contain white space between words, the English tokenizers do not handle Japanese text very well. Since our work exclusively handles Japanese text we thus decided to use the Japanese tokenizer "MeCab" [56]. The choice led us to have two different types of target words, one from the official JLPT questions and one from the MeCab tokenizer. To keep everything uniform we decided to unify our target words under the MeCab method since we have access to fewer official questions compared to unofficial questions. This change has no bearing on the results and is deemed the best solution to the problem at hand.

Similar to previous work [20], to validate our automatic distractor generation, the JLPT level accuracy, the difficulty, and the choice of distractor for each learning outcome, we asked 14 native Japanese teachers to answer our questionnaire. Most teachers (12) were still teachers of Japanese as a second language and had been teaching for ten or more years.

The JLPT is not the main focus for most teachers when creating lesson material. However, it remains a recurring part of the teaching process, as commented on by one participant, who said that the goal is to teach Japanese that is useful in everyday life. The JLPT structure becomes involved because of the grammar-building style of teaching.

5.1. Question Generation and JLPT Levels

In the questionnaire, we presented 30 cloze test questions, 10 for each learning outcome (kanji reading, kanji orthography, and vocabulary). Among the 30 questions, we included a mixture of 18 machine-generated and 12 human-made cloze tests questions. The split between question types are not balanced because we chose to randomly produce a mixture of questions for the questionnaire as a way to reduce potential bias from the teachers. Specifically, the question we asked the teachers was if they could differentiate between questions made by humans from the JLPT test and our automatically machine-generated questions. We also asked the teachers to judge how well each question fits into a given JLPT level.

For each questions about differentiating generation methods, the teachers were given three choices: (i) Machine-generated; (ii) Human-made; or (iii) Don't know. Table 7 represents the aggregated results of the answers given by the teachers (full table available in Table A4). The "Correct" column represents each time a teacher accurately judged a question as the correct generation method, while the "Wrong" column represents the inverse. From Table 7, we conclude that the vocabulary and kanji orthography questions are the most difficult ones for the teachers to detect. Approximately half of them (60%) could barely distinguish between human or machine-generated questions. However, they were better at differentiating the kanji reading questions (80% correct selection rate).

Table 7. Aggregated results of our Turing test. The answer columns represent how teachers judged the presented questions.

Learning Outcome	Correct	Wrong	Don't Know
Vocabulary	85 (60%)	43 (31%)	12 (9%)
Kanji Reading	112 (80%)	15 (11%)	13 (9%)
Kanji Orthography	83 (59%)	41 (29%)	16 (12%)

Regarding the JLPT level, as we can see in Table 8, the teachers deemed the questions to be appropriate for the assigned JLPT level 60% of the time. Again, the vocabulary and kanji orthography questions have similar results regarding the JLPT level assignment, while the kanji reading questions tend to be easier (34%) for the assigned level. In Table A5, Appendix A, we include the answers we received from the teachers split by questions and levels.

Table 8. Results of how well our automatically generated questions fit into the assigned JLPT level. The answer columns represent how teachers judged the presented questions.

Learning Outcome	Too Easy	Just Right	Too Hard	Don't Know
Vocabulary	15 (11%)	93 (66%)	17 (12%)	15 (11%)
Kanji Reading	47 (34%)	77 (55%)	9 (6%)	7 (5%)
Kanji Orthography	24 (17%)	84 (60%)	23 (17%)	9 (6%)

When we split the results and look at each generation method (see Table 9), we can see that the official human-made questions, on average, are seen as having the correct JLPT assignment 79% of the time, compared to the machine-generated questions, which are correctly assigned 49% of the time. A large part of the reduction in accuracy stems from the Kanji reading questions, which were assigned as "too easy" 57% of the time. With how often the teachers correctly judged the kanji reading questions as machine or human-made, there is an apparent issue with these distractor generation methods, which most likely

stems from the fact that the synonym and homonym distractors, in most cases, do not function as distractors for this learning outcome. A simple solution to this problem would be to only focus on the *similar-looking distractors* and improve that generation method.

Table 9. Split results of the assigned JLPT levels for questions.

Generation	Learning Outcome	Too Easy	Just Right	Too Hard	Don't Know
Human	Vocabulary	5 (12%)	35 (83%)	1 (2.5%)	1 (2.5%)
	Kanji Reading	8 (10%)	64 (80%)	7 (9%)	1 (1%)
	Kanji Orthography	4 (7%)	42 (75%)	8 (14%)	2 (4%)
NLP	Vocabulary	10 (10%)	58 (60%)	16 (16%)	14 (14%)
	Kanji Reading	39 (57%)	23 (34%)	0 (0%)	6 (9%)
	Kanji Orthography	20 (24%)	42 (50%)	15 (18%)	7 (8%)

Ultimately, we can generate level-appropriate cloze test questions most of the time using our unofficial dataset and distractor generation methods. It shows us that the available data online functions as a basis to work from. As the JLPT data collected by communities increase yearly, even without official data access, we expect to gain even better results with time.

5.2. Difficulty

To measure the difficulty, we randomly produced nine questions of random JLPT levels. For each question, we generated four sets of keys (A, B, C, D) and our difficulty-assigning algorithm assigned a difficulty level to each set. We then asked the teachers to rank each set of keys from one (easiest) to four (most difficult) without seeing how our algorithm ranked the sets.

Our automatic difficulty-assigning algorithm assigns difficulty with an error of ± 1 difficulty level 75% of the time, as shown in Table 10. In more detail, 25% belongs to perfect hits, meaning that our algorithm and the teachers agree on the difficulty level. The difficulty of the other 50% is, at most, one level up or down the difficulty automatically assigned. For example, if we analyze the second row of Table 10, our algorithm assigns a difficulty level of 2. In contrast, most teachers think that the difficulty level is 1 (7 out of 14), while four think that the difficulty level is 2. Finally, we marked with asterisks those questions that our algorithm differs from teachers' answers by more than one difficulty level (25% of the time). It is interesting to see that most of the incorrectly classified difficulties belong to the kanji reading learning outcomes, probably due to the fact that synonym and homonym distractors, in most cases, do not function as well as distractors for this learning outcome.

Our difficulty-assigning algorithm can assign five difficulty levels, giving us more fine-grained control over the difficulties. However, we only make use of four levels for our questions. This is because, during the automatic generation stage, the algorithm struggled to find valid combinations of distractors to reach the highest level. This makes sense since we use a JLPT-curated vocabulary list to pull words from, which would not include overly difficult words for a given level. We did not see this as an issue since creating difficult questions for the sake of difficulty does not translate into good questions. Since the results show that leveling Diff4 is challenging, adding another level may only have caused further issues with the rankings.

Table 10. Question difficulty of randomly generated questions for random JLPT levels. The letters indicates the set of keys (three distractors and the target) and the NLP(Difficulty) column represents the difficulty assigned by our algorithm. Diff 1 is the easiest whereas Diff 4 is the hardest difficulty. Each number in these columns represents how often a difficulty was selected by a teacher. A bold letter represents a perfect hit between teacher and our algorithm while two asterisks represents a complete miss. The remaining letters are within ±1 of the assigned difficulty.

Learning Outcome	Question	JLPT Level	Set of Keys	NLP (Difficulty)	Diff 1	Diff 2	Diff 3	Diff 4
Vocabulary	1	N5	**A**	3	4	5	5	0
			B	2	7	4	2	1
			C	1	6	4	3	1
			D	4	4	1	2	7
	2	N2	**A**	1	7	4	3	0
			B	3	3	5	3	3
			C	4	3	3	6	2
			D	2	7	1	2	4
	3	N4	**A**	1	8	3	2	1
			B	2	6	2	3	3
			C	3	3	2	5	4
			D **	4	9	4	0	1
Kanji Reading	1	N4	A	2	5	2	2	5
			B **	4	5	5	3	1
			C **	3	6	4	1	3
			D	1	5	2	5	2
	2	N5	**A**	1	10	1	2	1
			B	3	3	3	2	6
			C	2	3	4	6	1
			D **	4	4	6	1	3
	3	N1	A	2	7	2	3	2
			B **	1	4	3	2	5
			C **	4	5	4	4	1
			D **	3	7	2	2	3
Kanji Orthography	1	N1	A **	4	5	2	4	3
			B	1	6	5	1	2
			C	2	5	5	4	0
			D	3	3	3	2	6
	2	N4	A	3	4	5	2	3
			B	1	5	4	4	1
			C	2	6	2	3	3
			D	4	5	0	3	6
	3	N2	**A**	1	7	3	1	3
			B	2	6	3	3	2
			C	3	3	4	2	5
			D **	4	5	3	5	1

5.3. Candidate Creation Types

For each of our three learning outcomes, we present the teacher with a stem and three sets of keys, one for each distractor generation method. The goal is to investigate which of the generated distractor types would be preferred for a given learning outcome. The evaluation is performed through having the teachers rank each of the distractor types from best fitting (one) to worst fitting (three) in relation to the presented stem. Table 11 shows the results for each question type.

The preferred vocabulary distractor types are synonyms distractors, whereas the POS distractors are the worst type for that specific question type. Kanji orthography questions also have a clear winner in the similar kanji distractors, with homonyms being the least

liked. Once again, the results become less clear for the kanji reading questions where the similar-looking word distractors are seen as both the best and worst distractor type, with synonyms and homonyms having an even spread.

Table 11. Candidate types for each learning outcome ranked by teachers from best to worst for a given stem. The numbers indicate how often the teachers selected a choice.

Learning Outcome	Distractor Type	Best	Avg.	Worst
Vocabulary	Synonym	22 (18%)	14 (11%)	6 (5%)
	Homonym	11 (9%)	23 (18%)	8 (6%)
	Part-of-Speech	9 (7%)	5 (4%)	28 (22%)
Kanji Reading	Synonym	12 (10%)	14 (11%)	16 (13%)
	Homonym	14 (11%)	15 (12%)	13 (10%)
	Similar looking words	16 (13%)	13 (10%)	13 (10%)
Kanji Orthography	Synonym	11 (9%)	27 (21%)	4 (3%)
	Homonym	2 (2%)	8 (6%)	32 (25%)
	Similar looking kanji	29 (23%)	7 (6%)	6 (5%)

In the follow-up questions to the preferred distractor type, we asked what the teachers deemed the most important when creating each question type. The answers were in line with what we can see in the previous tables, as the essential point for vocabulary questions is that the distractors should have a similar meaning to the target (79%); for kanji orthography, the essential point is that the distractors should look like the target kanji (71%). The answers for the kanji reading distractors are more even, where the distractors should have a similar ending hiragana (50%) and a similar spelling to the target word (58%).

Finally, we asked the teachers which question types they deem the easiest to hardest to create when making lesson material. Out of the three question types, the kanji reading type is the easiest to create, even though a majority put it at average difficulty. It is followed by the kanji orthography questions, which tend to lean towards the more difficult side, and finally, clearly the most difficult being vocabulary questions.

5.4. Cloze Test Examples

In Table 12, we include two example questions—vocabulary and kanji orthography—automatically generated by our model for the JLPT level 5, which we used in our questionnaire. The vocabulary question was wrongly judged by 10 teachers, who thought it had been generated by a human. A majority also judged the question to be "Just right" for the assigned level. This is an example of a generated question which was indistinguishable from a human-made question.

On the other hand, the kanji orthography question is an example of a generated question where almost every teacher was able to correctly judge it as generated by our model. They also judged the question to be "Too easy" for the assigned level, which is most likely the cause for the teachers to be able to correctly place the question.

Table 12. Two questions (vocabulary and kanji orthography respectively) used in our questionnaire. Teachers had to guess whether they were automatically generated or not, and measure the difficulty for JLPT N5.

Question	Generation Method			Difficulty			
	NLP	Human	Don't Know	Too Easy	Just Right	Too Hard	Don't Know
朝は何も食べません。牛乳だけ[__]ます。 (I don't eat anything in the morning. I only [__] milk) a. 飲み (drink) b. 着 (wear) c. 洗い (wash) d. 寝 (sleep)	4 (29%)	8 (57%)	2 (14%)	1 (7%)	9 (63%)	2 (14%)	2 (14%)

Table 12. *Cont.*

Question	Generation Method			Difficulty			
	NLP	Human	Don't Know	Too Easy	Just Right	Too Hard	Don't Know
アンジェラさんの走り[かた]はとてもかわいいです。 (Angela's running **[style]** is very cute.) a. 方 (style, manner of) b. 地図 (map) c. 出 (leave, exit) d. 下手 (poor, awkward)	13 (93%)	0 (0%)	1 (7%)	7 (50%)	1 (7%)	4 (29%)	2 (14%)

6. Discussion

The following section discusses the results and answers the four Research Questions (RQs) we initially proposed.

6.1. RQ1: Are the Generated Distractors Indistinguishable from Human Made Distractors?

The results show that teachers can detect whether a question was generated by a human or a computer with over 60% accuracy. Note that this probability is close to that of flipping a coin. However, teachers are incredibly adept at detecting kanji reading types, showing us that we must improve this subtype of questions.

Although the answer to RQ1 is that we cannot generate distractors totally indistinguishable from human-made ones, the results are promising, and there is potential for better results with an improved generation method.

6.2. RQ2: Can We Generate JLPT Level Appropriate Distractors?

We had concerns about how well-generated distractors would fit into the five JLPT levels since we trained our models (i.e., Word2vec, BERT) on native-level language from the Japanese Wikipedia. However, looking at every question, the teachers agreed in a unified majority that most questions were of the correct JLPT level. Splitting the results between the human-made and machine-generated questions (see Table 9), we can see how well the official questions are split compared to the machine-generated questions. According to the teachers, the kanji reading questions are too easy, and considering that they are also correctly judging between human and machine generation 80% of the time, there are some problems with this generation method. On the other hand, while there is some leeway between "too hard" and "too easy", the vocabulary and kanji orthography questions show adequate results, being correct over half of the time. In summary, there is still potential for improvements, as evidenced by the outcomes related to kanji reading distractors. Nevertheless, we maintain that our generation methods hold promise in consistently producing level-appropriate distractors for two of our three learning outcomes.

6.3. RQ3: Can We Use NLP Methods to Attach a Valid Difficulty Rank to Generated Questions?

As we demonstrated in Section 5.2, we can automatically assign the difficulty with an error of ± 1 difficulty level with high accuracy (75% of the time). However, for the remaining 25% we are far from reviewers' answers. Most of the incorrectly classified questions belong to kanji reading learning outcomes. The most likely reason is that synonym and homonym distractors do not function well together with this learning outcome compared to the *similar-looking distractor* type.

Even though the answer to this RQ is positive, we are currently working on two improvements to address the incorrectly classified questions. On the one hand, we are only focusing on the *similar-looking distractors* as a distractor type. On the other hand, we are improving the generation method to have more elaborated distractors for this particular learning outcome.

6.4. RQ4: Is There a Preferred Distractor Type?

The results of the vocabulary questions, where synonyms are the preferred distractor type, are in line with previous works for other languages [48], as well as teachers' priorities

regarding what the most crucial part of a vocabulary question creation is. While the POS is the worst distractor type among the presented types, the likely cause is our simple implementation of random POS words. We are currently working on a more optimal generation method for vocabulary questions to produce synonyms and add further limiters where only synonyms of the same POS are selected, combining the two generation methods.

The kanji orthography questions also have a clear winner with distractors that share similar visual characteristics as the target word. This result aligns with what we expected and is a significant reason for including this distractor type. When looking at the official kanji orthography questions, they follow a similar pattern for their distractors in most cases. Although it would be preferred if a distractor could share characteristics and be a synonym or homonym, this is harder to achieve with kanji since they are made up of elements and radicals, which does not necessarily mean similar-looking kanji share a spelling. The teachers also deemed having similar distractors the most important when creating kanji orthography questions. This result leads us to conclude that kanji orthography questions have an optimal distractor type.

The preferred distractor type must be clarified between homonyms or similar-looking words for the kanji reading questions. This is not surprising since the similar-looking word distractor is based around the homonym distractors but with further limitations during generation. According to the teachers, the kanji reading questions are among the easiest types of questions to make, which is in contrast to being the hardest to create by our automated generation methods (see Table 13). It has been the most troublesome distractor type to generate because of how limited the distractors may look. Since the target kanji is presented to the examinees in the stem, the distractors must stay within the target word. Otherwise, the distractor may make the question too easy to answer. This limitation is causing trouble throughout all questions related to kanji reading since all results for this question type are inconclusive. We expected homonyms to perform best since the distractors would all be in hiragana and must be spelled similarly. While somewhat true, this only works whenever the target word only contains a single kanji or is a kanji compound. Whenever the word is a mixture of kanji and hiragana, trouble arises since there are more factors that we need to take into account to make sure the distractors can not be removed directly as a possible choice.

Table 13. Teachers' responses regarding which type of cloze test questions are more difficult to create.

Learning Outcome	Easy	Avg.	Hard	Don't Know
Vocabulary	2 (14%)	4 (29%)	8 (57%)	0 (0%)
Kanji Reading	6 (43%)	7 (50%)	1 (7%)	0 (0%)
Kanji Orthography	3 (21%)	6 (43%)	4 (29%)	1 (7%)

In conclusion, the vocabulary and kanji orthography questions have a clear preferred distractor type, whereas the kanji reading results are less clear. We recommend simple single-type generation methods for all three question types. However, we also recommend that improvements are made to each distractor type for better results in the future. We are currently working on combining generation methods and a more comprehensive generating structure for the kanji reading questions to increase the generation of functional distractors.

6.5. Limitations

In this section, we discuss three of the main limitations our work has: lack of official data, low number of participants, and the unbalanced dataset.

The fact that the operators of the JLPT keep most parts of the test a secret causes issues since there is no up-to-date readily available data. To address this, we gathered data from online sources ("nihongokyoushi-net" [43] and "tanos" [44]). While the results we present show promise, we anticipate that official data would yield even better ones. Training models on each JLPT level instead of native-level text have the potential for much greater accuracy in terms of generating JLPT-appropriate distractors.

With regard to the participants in our questionnaire, our study included 14 teachers. While this number is significantly larger than in other studies in the field [20], in which authors assessed the quality of distractors by involving only four teachers, we still consider 14 teachers to constitute a relatively modest sample size. Although our findings are promising, it is advisable to validate them by incorporating a more extensive and diverse range of participants.

Lastly, it is important to note that we intentionally chose not to have a balanced dataset for the initial 30 questions in the questionnaire. We made this decision to prevent potential bias, ensuring that participating teachers would not have expectations about a specific number of questions being generated by humans or otherwise.

7. Related Research

The three areas we wish to address in this section is "AI in Education" and the main approaches to automatic distractor generation. The later can be split into two parts: distractor generation (where most works focus on) and distractor ranking.

AI in Education. Artificial intelligence gained prominence in the 2000s within the field of education, demonstrating its ability to have a positive impact [57,58]. However, it is important to note that many Artificial Intelligence (AI) techniques are originally designed for general applications and may not always address the specific needs of a particular domain [57]. While we are still a long way from replacing human educators with automated methods, AI can play a crucial role in educational assessments, such as automated essay scoring and computer adaptive tests [59]. In particular, the combination of NLP and education, has emerged as a promising research area [60]. NLP can help teachers and students in various ways, including automatic feedback analysis [61], the automated grading of open-ended questions [62], the inclusion of chatbots for student support [63,64], and the automatic generation of multiple-choice questions [20,65].

Distractor Generation. Early distractor generation methods used random words [21] or words with a similar frequency as the key [66]. These methods have since evolved into more modern versions, which tend to use semantically related words acquired through WordNet [66,67], a thesaurus [22], or n-grams and collocations [23,68]. Recently, word embeddings have been shown to be effective distractors, and the Word2vec method has gained popularity as a generation method [4,25,47], which achieves semantically similar words. Automatic distractors' generation in languages other than English has also gained more attention in the last decade with improvements to NLP methods, allowing for training models in multiple languages [48]. One work in a minor language, Lao, inspired our research since authors created distractors for a low-resource language using simple synonyms and homonym distractors [47]. Our paper aims to investigate five distractor generation methods over three types of questions using Word2vec, Hamming distance, and more specified generation methods for Japanese. To the best of our knowledge, there has been no prior work carried out on distractor generation in Japanese.

Distractor Ranking. Distractors are often generated by using the target word as a base. A method of then selecting appropriate distractors from the generated ones is by comparing scores assigned to each distractor at generation. Two such scoring methods are Path and WU-Palmer similarity scores [3]. Unfortunately, these methods miss out on any context a word has in a sentence since they simply score distractors using the target word. To solve this problem, some works used BERT models to generate distractors [11] and rank distractors generated through other means [26]. However, there are contradicting results regarding the usage of the BERT model for this purpose. Some authors used BERT as a method of generating distractors and stated that the distractors generated by BERT outperformed any other generation method [11]. However, three years later, other authors also focusing on distractors generation seemingly disproved the previous results concluding that BERT is ineffective for distractor generation [26]. These works used different languages and thus may have discrepancies in the results. However, we agree that BERT should not be used to generate distractors since it generates words that fit into a sentence, opening

up possibilities for unambiguous answers where multiple keys could give a complete sentence. Therefore, we limit ourselves to using BERT to score our distractors in tandem with similarity scores given by methods such as Word2vec.

8. Conclusions

We presented a novel approach to automatically generate cloze tests in Japanese for three learning outcomes: vocabulary, kanji reading, and kanji orthography. For each one, we generated three types of distractors for any given target word where two of the distractors are universal between all three question types (synonyms and homonyms) and one type is exclusive for each question (POS, similar-looking words, and similar-looking kanji respectively).

We evaluated our machine-generated distractors through a questionnaire where we asked 14 experts to evaluate each type of question. Our questions revolved around whether the machine-generated questions are indistinguishable from human-made questions, whether we could make JLPT-level relevant questions, whether it is possible to use NLP methods to attach a difficulty rank to a question, and whether there is a preferred distractor type for the given learning outcomes.

Limited by the amount of research in this area, our goal was to set the groundwork for future research and function as a benefit for teachers of the Japanese language or anyone who wishes to create Japanese cloze tests. The results show that we are close to automatically generating questions that are indistinguishable from human-made questions and we were able to empirically confirm that our questions are appropriate for the given JLPT levels. Although the difficulty-assigning algorithm must be further improved, it shows promising results already, and we demonstrated a clear preferred distractor type for two of our three learning outcomes.

Author Contributions: Conceptualization, T.A. and P.P.-S.; methodology, T.A. and P.P.-S.; software, T.A.; validation, T.A. and P.P.-S.; formal analysis, T.A. and P.P.-S.; investigation, T.A. and P.P.-S.; resources, T.A.; data curation, T.A.; writing—original draft preparation, T.A. and P.P.-S.; writing—review and editing, T.A. and P.P.-S.; visualization, T.A. and P.P.-S.; supervision, P.P.-S.; project administration, P.P.-S.; funding acquisition, P.P.-S. All authors have read and agreed to the published version of the manuscript

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: No new data were created or analyzed in this study. Data sharing is not applicable to this article.

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

Appendix A. The Japanese Language

Table A1. Some examples of katakana words.

Type	Japanese	Translation
Name	ティム	Tim
Fruit or color	オレンジ	Orange
Location	イギリス	England

Table A2. Examples of elements and radicals in kanji.

Type	Kanji	Parts
Elements	空	工, 八, 冖, 𠂉, 穴
Radicals	空	穴, 工

Table A3. An example of a sentence using only hiragana and the same sentence using a Kanji.

Type	Japanese	Translation
Hiragana	はは は はな が すき だ	My mother likes flowers
Kanji/Hiragana	母 は 花 が 好き だ	My mother likes flowers

Table A4. This table show the results of our Turing test. We show all ten questions for each learning outcome, along with the intended JLPT level of the question. The column for Generation method shows whether the questions come from a human (H) of from our generation method (NLP). The next three columns (Correct, Wrong, and Don't Know) represent the answer of the teachers.

Learning Outcome	Question	JLPT Level	Generation Method	Correct	Wrong	Don't Know
Vocabulary	1	N5	NLP	4	8	2
	2	N3	NLP	12	2	0
	3	N1	NLP	6	6	2
	4	N3	NLP	11	2	1
	5	N2	H	12	2	0
	6	N4	NLP	7	6	1
	7	N2	NLP	2	9	3
	8	N5	NLP	8	5	1
	9	N5	H	10	3	1
	10	N1	H	13	0	1
Kanji Reading	1	N4	H	7	5	2
	2	N3	NLP	10	3	1
	3	N2	H	12	1	1
	4	N3	NLP	11	1	2
	5	N4	NLP	13	0	1
	6	N3	H	13	1	0
	7	N2	H	12	1	1
	8	N5	NLP	10	2	2
	9	N1	NLP	13	0	1
	10	N3	H	11	1	2
Kanji Orthography	1	N3	H	8	4	2
	2	N2	H	9	3	2
	3	N5	H	9	4	1
	4	N5	NLP	13	0	1
	5	N2	NLP	7	7	0
	6	N2	NLP	5	7	2
	7	N1	NLP	4	8	2
	8	N3	NLP	9	4	1
	9	N4	NLP	9	2	3
	10	N3	H	10	2	2

Table A5. Results of how well our automatically generated questions fit into the assigned JLPT level. We show all ten questions for each learning outcome along with the intended JLPT level of the question. The following four columns (Too easy, Just right, Too hard, and Don't Know) represent the answer of the teachers.

Learning Outcome	Question	JLPT Level	Too Easy	Just Right	Too Hard	Don't Know
Vocabulary	1	N5	1	9	2	2
	2	N3	1	6	4	3
	3	N1	1	11	2	0
	4	N3	2	9	0	3
	5	N2	2	12	0	0
	6	N4	2	8	3	1
	7	N2	2	8	2	2
	8	N5	1	7	3	3
	9	N5	3	9	1	1
	10	N1	0	14	0	0

Table A5. *Cont.*

Learning Outcome	Question	JLPT Level	Too Easy	Just Right	Too Hard	Don't Know
Kanji Reading	1	N4	1	8	4	1
	2	N3	7	6	0	1
	3	N2	2	12	0	0
	4	N3	12	1	0	1
	5	N4	9	4	0	1
	6	N3	0	13	1	0
	7	N2	3	11	0	0
	8	N5	3	7	2	2
	9	N1	8	5	0	1
	10	N3	2	10	2	0
Kanji Orthography	1	N3	0	12	1	1
	2	N2	2	11	1	0
	3	N5	2	11	1	0
	4	N5	7	1	4	2
	5	N2	0	7	5	2
	6	N2	9	5	0	0
	7	N1	0	11	3	0
	8	N3	1	10	1	2
	9	N4	3	8	2	1
	10	N3	0	8	5	1

Table A6. Candidate types ranked by teachers from best to worst for a given stem. The numbers indicate how often the evaluators selected a choice.

Learning Outcome	Question	JLPT Level	Distractor Type	Best	Avg.	Worst
Vocabulary	1	N4	Synonym	7	5	2
			Homonym	4	8	2
			Part-of-Speech	3	1	10
	2	N1	Synonym	5	6	3
			Homonym	5	6	3
			Part-of-Speech	4	2	8
	3	N3	Synonym	10	3	1
			Homonym	2	9	3
			Part-of-Speech	2	2	10
Kanji Reading	1	N3	Synonym	5	3	6
			Homonym	3	6	5
			Similar looking words	6	5	3
	2	N3	Synonym	5	8	1
			Homonym	7	4	3
			Similar looking words	2	2	10
	3	N4	Synonym	2	3	9
			Homonym	4	5	5
			Similar looking words	8	6	0
Kanji Orthography	1	N2	Synonym	3	8	3
			Homonym	2	4	8
			Similar kanji	9	2	3
	2	N4	Synonym	4	10	0
			Homonym	0	2	12
			Similar kanji	10	2	2
	3	N3	Synonym	4	9	1
			Homonym	0	2	12
			Similar kanji	10	3	1

References

1. Araki, J.; Rajagopal, D.; Sankaranarayanan, S.; Holm, S.; Yamakawa, Y.; Mitamura, T. Generating Questions and Multiple-Choice Answers using Semantic Analysis of Texts. In Proceedings of the COLING, Osaka, Japan, 11 December 2016; pp. 1125–1136.
2. Satria, A.Y.; Tokunaga, T. Automatic generation of english reference question by utilising nonrestrictive relative clause. In Proceedings of the CSEDU, Porto, Portugal, 21–23 April 2017; Volume 2, pp. 379–386.
3. Susanti, Y.; Iida, R.; Tokunaga, T. Automatic Generation of English Vocabulary Tests. In Proceedings of the CSEDU, Lisbon, Portugal, 23–25 May 2015.

4. Zhang, C.; Sun, Y.; Chen, H.; Wang, J. Generating Adequate Distractors for Multiple-Choice Questions. *arXiv* **2020**, arXiv:2010.12658.
5. Thalheimer, W. Learning Benefits of Questions, V2.0. 2014. Available online: https://www.worklearning.com/wp-content/uploads/2017/10/Learning-Benefits-of-Questions-2014-v2.0.pdf (accessed on 1 November 2023).
6. Dave, N.; Bakes, R.; Pursel, B.; Giles, C.L. Math Multiple Choice Question Solving and Distractor Generation with Attentional GRU. In Proceedings of the EDM, Online, 29 June–2 July 2021.
7. Bakes, R. Capabilities for Multiple Choice Question Distractor Generation and Elementary Mathematical Problem Solving by Recurrent Neural Networks. Bachelor's Thesis, Pennsylvania State University, State College, PA, USA, 2020.
8. Agarwal, M.; Mannem, P. Automatic Gap-fill Question Generation from Text Books. In Proceedings of the BEA, Portland, OR, USA, 24 June 2011; pp. 56–64.
9. Maslak, H.; Mitkov, R. Paragraph Similarity Matches for Generating Multiple-choice Test Items. In Proceedings of the RANLP, Online, 6–8 September 2021; pp. 99–108.
10. Sajjad, M.; Iltaf, S.; Khan, R.A. Nonfunctional distractor analysis: An indicator for quality of Multiple choice questions. *Pak. J. Med. Sci.* **2020**, *36*, 982. [CrossRef]
11. Yeung, C.; Lee, J.; Tsou, B. Difficulty-aware Distractor Generation for Gap-Fill Items. In Proceedings of the ALTA, Perth, Australia, 23–25 May 2019; pp. 167–172.
12. Labrak, Y.; Bazoge, A.; Dufour, R.; Daille, B.; Gourraud, P.A.; Morin, E.; Rouvier, M. FrenchMedMCQA: A French Multiple-Choice Question Answering Dataset for Medical domain. In Proceedings of the LOUHI, Abu Dhabi, United Arab Emirates, 30 May 2022; pp. 41–46.
13. Nwafor, C.A.; Onyenwe, I.E. An automated multiple-choice question generation using natural language processing techniques. *arXiv* **2021**, arXiv:2103.14757.
14. CH, D.R.; Saha, S.K. Automatic Multiple Choice Question Generation From Text: A Survey. *IEEE Trans. Learn. Technol.* **2020**, *13*, 14–25. [CrossRef]
15. Kurdi, G.; Leo, J.; Parsia, B.; Sattler, U.; Al-Emari, S. A Systematic Review of Automatic Question Generation for Educational Purposes. *Int. J. Artif. Intell. Educ.* **2019**, *30*, 121–204. [CrossRef]
16. Zhang, Z.; Mita, M.; Komachi, M. Cloze Quality Estimation for Language Assessment. In Proceedings of the EACL, Dubrovnik, Croatia, 2–6 May 2023; pp. 540–550.
17. Huang, Y.T.; Mostow, J. Evaluating human and automated generation of distractors for diagnostic multiple-choice cloze questions to assess children's reading comprehension. In Proceedings of the AIED, Madrid, Spain, 21–25 June 2015; pp. 155–164.
18. Haladyna, T.M. *Developing and Validating Multiple-Choice Test Items*; Routledge: London, UK, 2004.
19. Susanti, Y.; Tokunaga, T.; Nishikawa, H.; Obari, H. Automatic distractor generation for multiple-choice English vocabulary questions. *Res. Pract. Technol. Enhanc. Learn.* **2018**, *13*, 15. [CrossRef]
20. Larrañaga, M.; Aldabe, I.; Arruarte, A.; Elorriaga, J.A.; Maritxalar, M. A Qualitative Case Study on the Validation of Automatically Generated Multiple-Choice Questions From Science Textbooks. *IEEE Trans. Learn. Technol.* **2022**, *15*, 338–349. [CrossRef]
21. Hoshino, A.; Nakagawa, H. A Real-Time Multiple-Choice Question Generation For Language Testing: A Preliminary Study. In Proceedings of the BEA, Ann Arbor, MI, USA, 29 June 2005; pp. 17–20.
22. Sumita, E.; Sugaya, F.; Yamamoto, S. Measuring Non-native Speakers' Proficiency of English by Using a Test with Automatically-Generated Fill-in-the-Blank Questions. In Proceedings of the BEA, Ann Arbor, MI, USA, 29 June 2005; pp. 61–68.
23. Liu, C.L.; Wang, C.H.; Gao, Z.M.; Huang, S.M. Applications of Lexical Information for Algorithmically Composing Multiple-Choice Cloze Items. In Proceedings of the BEA, Ann Arbor, MI, USA, 29 June 2005; pp. 1–8.
24. Das, B.; Majumder, M.; Phadikar, S.; Sekh, A.A. Automatic generation of fill-in-the-blank question with corpus-based distractors for e-assessment to enhance learning. *Comput. Appl. Eng. Educ.* **2019**, *27*, 1485–1495. [CrossRef]
25. Das, B.; Majumder, M.; Phadikar, S.; Sekh, A.A. Multiple-choice question generation with auto-generated distractors for computer-assisted educational assessment. *Multimed. Tools Appl.* **2021**, *80*, 31907–31925. [CrossRef]
26. Panda, S.; Palma Gomez, F.; Flor, M.; Rozovskaya, A. Automatic Generation of Distractors for Fill-in-the-Blank Exercises with Round-Trip Neural Machine Translation. In Proceedings of the ACL, Dublin, Ireland, 22–27 May 2022; pp. 391–401.
27. Hutchins, W.J. The Georgetown-IBM Experiment Demonstrated in January 1954. In Proceedings of the AMTA, Washington, DC, USA, 28September–2 October 2004; Frederking, R.E., Taylor, K.B., Eds.; Springer: Berlin/Heidelberg, Germany, 2004; pp. 102–114.
28. Devlin, J.; Chang, M.W.; Lee, K.; Toutanova, K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In Proceedings of the NAACL-HLT, Minneapolis, MN, USA, 2 June–7 June 2019; pp. 4171–4186.
29. Rogers, A.; Kovaleva, O.; Rumshisky, A. A Primer in BERTology: What We Know About How BERT Works. *Trans. Assoc. Comput. Linguist.* **2020**, *8*, 842–866. [CrossRef]
30. Hepburn, J.C. *A Japanese-English and English-Japanese Dictionary*; Trübner: London, UK; Maruya & Company: Tokyo, Japan, 1886.
31. Kinsui, S. *Virtual Japanese: Enigmas of Role Language*; Osaka University Press: Suita, Japan, 2017.
32. Hasegawa, Y. *The Routledge Course in JAPANESE Translation*; Routledge: London, UK, 2012.
33. Trott, S.; Bergen, B. Why do human languages have homophones? *Cognition* **2020**, *205*, 104449. [CrossRef] [PubMed]
34. Petrov, S.; Das, D.; McDonald, R. A Universal Part-of-Speech Tagset. In Proceedings of the LREC, Istanbul, Turkey, 21–27 May 2012; pp. 2089–2096.

35. Foundation, J.; Exchanges, J.E.; Services. Japanese-Language Proficiency Test—Statistics. 2023. Available online: https://www.jlpt.jp/e/statistics/archive/202301.html (accessed on 27 February 2023).
36. Objectives and History | JLPT Japanese-Language Proficiency Test. Available online: https://www.jlpt.jp/e/about/purpose.html (accessed on 25 August 2023).
37. Foundation, J.; Exchanges, J.E.; Services. The New Japanese Language Proficiency Test Guidebook, Summarized Version. Available online: https://www.jlpt.jp/e/reference/pdf/guidebook_s_e.pdf (accessed on 27 February 2023).
38. Composition of Test Sections and Items | JLPT Japanese-Language Proficiency Test. Available online: https://www.jlpt.jp/e/guideline/testsections.html (accessed on 25 August 2023).
39. Nishizawa, H.; Isbell, D.R.; Suzuki, Y. Review of the Japanese-Language Proficiency Test. *Lang. Test.* **2022**, *39*, 494–503. [CrossRef]
40. Iles, T.; Rojas-Lizana, S. 'Changes' to the new Japanese-Language Proficiency Test: Newly emerged language policies for non-Japanese and Japanese citizens. *Electron. J. Contemp. Jpn. Stud.* **2019**, *9*, 8.
41. Raymond, M.R.; Stevens, C.; Bucak, S.D. The optimal number of options for multiple-choice questions on high-stakes tests: Application of a revised index for detecting nonfunctional distractors. *Adv. Health Sci. Educ.* **2019**, *24*, 141–150. [CrossRef] [PubMed]
42. Guo, H.; Zu, J.; Kyllonen, P. A Simulation-Based Method for Finding the Optimal Number of Options for Multiple-Choice Items on a Test. *ETS Res. Rep. Ser.* **2018**, *2018*, 1–17. [CrossRef]
43. Japanese NET (日本語NET). Available online: https://nihongokyoshi-net.com (accessed on 30 August 2023).
44. Japanese Language Proficiency Test Resources. Available online: https://www.jlpt.jp/e/index.html (accessed on 30 August 2023).
45. Japanese-English Dictionary. Available online: https://jisho.org (accessed on 30 August 2023).
46. New Japanese-Language Proficiency Test Sample Questions | JLPT Japanese-Language Proficiency Test. Available online: https://www.jlpt.jp/e/samples/forlearners.html (accessed on 30 August 2023).
47. Qiu, X.; Xue, H.; Liang, L.; Xie, Z.; Liao, S.; Shi, G. Automatic Generation of Multiple-choice Cloze-test Questions for Lao Language Learning. In Proceedings of the IALP, Singapore, 11–13 December 2021; pp. 125–130.
48. Han, Z. Unsupervised Multilingual Distractor Generation for Fill-in-the-Blank Questions. Master's Thesis, Uppsala University, Department of Linguistics and Philology, Uppsala, Sweden, 2022.
49. Goodrich, H.C. Distractor efficiency in foreign language testing. In *Tesol Quarterly*; Teachers of English to Speakers of Other Languages, Inc.: Alexandria, VA, USA, 1977; pp. 69–78.
50. Yujian, L.; Bo, L. A Normalized Levenshtein Distance Metric. *IEEE Trans. Pattern Anal. Mach. Intell.* **2007**, *29*, 1091–1095. [CrossRef] [PubMed]
51. Japanese N-gram (コーパス—日本語ウェブコーパス) 2010. Available online: https://www.s-yata.jp/corpus/nwc2010/ngrams/ (accessed on 30 August 2023).
52. Ebel, R.L. Procedures for the Analysis of Classroom Tests. *Educ. Psychol. Meas.* **1954**, *14*, 352–364. [CrossRef]
53. Trace, J.; Brown, J.D.; Janssen, G.; Kozhevnikova, L. Determining cloze item difficulty from item and passage characteristics across different learner backgrounds. *Lang. Test.* **2017**, *34*, 151–174. [CrossRef]
54. Olney, A.M.; Pavlik, P.I., Jr.; Maass, J.K. Improving reading comprehension with automatically generated cloze item practice. In Proceedings of the AIED, Wuhan, China, 28 June–1 July 2017; pp. 262–273.
55. Brown, J.D. Cloze Item Difficulty. *Jpn. Assoc. Lang. Teach. J.* **1989**, *11*, 46–67.
56. Kudo, T. Mecab: Yet Another Part-of-Speech and Morphological Analyzer. 2005. Available online: http://mecab.sourceforge.net/ (accessed on 31 July 2023).
57. Zhai, X.; Chu, X.; Chai, C.S.; Jong, M.S.Y.; Istenic, A.; Spector, M.; Liu, J.B.; Yuan, J.; Li, Y. A Review of Artificial Intelligence (AI) in Education from 2010 to 2020. *Complexity* **2021**, *2021*, 1–18. [CrossRef]
58. Zawacki-Richter, O.; Marín, V.I.; Bond, M.; Gouverneur, F. Systematic review of research on artificial intelligence applications in higher education–where are the educators? *Int. J. Educ. Technol. High. Educ.* **2019**, *16*, 1–27. [CrossRef]
59. Gardner, J.; O'Leary, M.; Yuan, L. Artificial intelligence in educational assessment: 'Breakthrough? Or buncombe and ballyhoo?'. *J. Comput. Assist. Learn.* **2021**, *37*, 1207–1216. [CrossRef]
60. Kurni, M.; Mohammed, M.S.; Srinivasa, K. Natural Language Processing for Education. In *A Beginner's Guide to Introduce Artificial Intelligence in Teaching and Learning*; Springer: Berlin/Heidelberg, Germany, 2023; pp. 45–54.
61. Shaik, T.; Tao, X.; Li, Y.; Dann, C.; McDonald, J.; Redmond, P.; Galligan, L. A review of the trends and challenges in adopting natural language processing methods for education feedback analysis. *IEEE Access* **2022**, *10*, 56720–56739. [CrossRef]
62. Smith, G.G.; Haworth, R.; Žitnik, S. Computer science meets education: Natural language processing for automatic grading of open-ended questions in ebooks. *J. Educ. Comput. Res.* **2020**, *58*, 1227–1255. [CrossRef]
63. Molnár, G.; Szüts, Z. The role of chatbots in formal education. In Proceedings of the 2018 IEEE 16th International Symposium on Intelligent Systems and Informatics (SISY), Subotica, Serbia, 13–15 September 2018; pp. 197–202.
64. Pérez, J.Q.; Daradoumis, T.; Puig, J.M.M. Rediscovering the use of chatbots in education: A systematic literature review. *Comput. Appl. Eng. Educ.* **2020**, *28*, 1549–1565. [CrossRef]
65. Mitkov, R.; Mitkov, R.; Ha, L.A. Computer-aided generation of multiple-choice tests. In Proceedings of the HLT-NAACL 03 Workshop on Building Educational Applications Using Natural Language Processing, Edmonton, AB, Canada, 31 May 2003; pp. 17–22.

66. Brown, J.C.; Frishkoff, G.A.; Eskenazi, M. Automatic Question Generation for Vocabulary Assessment. In Proceedings of the HLT, Vancouver, BC, Canada, 6–8 October 2005; pp. 819–826.
67. Zesch, T.; Melamud, O. Automatic Generation of Challenging Distractors Using Context-Sensitive Inference Rules. In Proceedings of the BEA, Baltimore, MD, USA, 26 June 2014; pp. 143–148.
68. Hill, J.; Simha, R. Automatic Generation of Context-Based Fill-in-the-Blank Exercises Using Co-occurrence Likelihoods and Google n-grams. In Proceedings of the BEA, San Diego, CA, USA, 16 June 2016; pp. 23–30.

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.