Automatic Generation of English Lesson Materials for Native Speakers of Japanese

Jason Ginsburg
University of Aizu
Tsuruga, Ikki-machi, Aizu-Wakamatsu City
Fukushima, Japan, 965-8580
jginsbur@gmail.com

ABSTRACT
This paper describes a computer program, a Lesson Material Builder, that is designed to automatically generate editable lesson materials for teaching English to speakers of Japanese. This program has two functions, a) an Exercise Creator, and b) a Vocabulary Quiz Creator. The Exercise Creator automatically creates an editable reading exercise document, from an input text, in which target vocabulary words are highlighted and definitions in Japanese are provided. Taking a reading exercise, of the sort produced by the Exercise Creator, the Vocabulary Quiz Creator can create vocabulary quizzes that require students to match vocabulary words with their definitions.

Categories and Subject Descriptors
K.3.1 [**Computer Uses in Education**]: Computer-assisted instruction (CAI)

General Terms
Human Factors, Design, Theory

Keywords
English language teaching, reading, vocabulary

1. INTRODUCTION
Creating lesson materials for English language classes, or for any subject, can be time consuming. An instructor, if developing his/her own materials, must determine what types of materials are appropriate for the target students, what types of materials to create, how to create these materials, how to incorporate these materials into a lesson, etc. When developing lesson materials, there are different skills that an instructor can choose to focus on. Crucially, vocabulary skills are necessary - it is necessary for a student to understand most of the vocabulary words that he/she encounters in a foreign language in order to understand what he/she is exposed to. Reading English text requires the ability to understand the vocabulary words in the text. To this end, one could create a reading exercise in which target vocabulary words are highlighted and definitions are provided. It may also be desirable to give students vocabulary quizzes that test students' understanding of target vocabulary words, as well as motivate students to study. In this paper, we describe software, a Lesson Material Builder, that aids in the development of English language lesson materials of this sort.

The Lesson Material Builder is designed to help teachers of English create lesson materials for Japanese speaking students. This software does not replace the teacher - rather, it can be used to create lesson materials and quizzes that can be created from materials of the instructors' choosing and that can be edited as necessary. The user (e.g., an English instructor) chooses reading passages that are appropriate for the target students and creates lesson materials from these texts. Crucially, this software can help instructors create high quality lesson materials in a shorter amount of time than would normally be required, thus giving instructors more time to focus on teaching and evaluating students' work.

The core functions of this Lesson Material Builder are a) an Exercise Creator and b) a Vocabulary Quiz Creator. The Exercise Creator is fed a text document, from which it generates an editable output file in which target vocabulary items are highlighted and definitions are provided. This document can be used for helping English language learners read English. The Vocabulary Quiz Creator function is fed a document that lists vocabulary items and their definitions and then automatically creates quizzes that require students to match vocabulary words with their definitions. In this paper, we explain how this Lesson Material Builder works and we give some examples of its uses.

2. MATERIALS
This application was created in Python and uses the materials listed in Table 1. We created a unigram frequency list from a downloaded version of the entire Wikipedia corpus.[1] Then we created a list of all unigram types and the frequency (number of occurrences) of each type. This list is used to extract target vocabulary items from texts. We used a list

[1]This was downloaded from `http://en.wikipedia.org/wiki/Wikipedia:Database_download#English-language_Wikipedia` on September 6, 2010 and consists of all "articles, templates, image descriptions and primary meta-pages" of Wikipedia. After cleaning tags, via a Python script, this corpus has a size of 1,003,961,037 tokens.

HCCE '12, March 8–13, 2012, Aizu-Wakamatsu, Fukushima, Japan.

Table 1: Materials

Material	Size (Types)	Function
Unigram frequency list (from Wikipedia)	68,711,145	Determine target vocabulary terms
List of stop words	570	Eliminate stop words
Gene 95 English-Japanse dictionary	57,350	Provide definitions

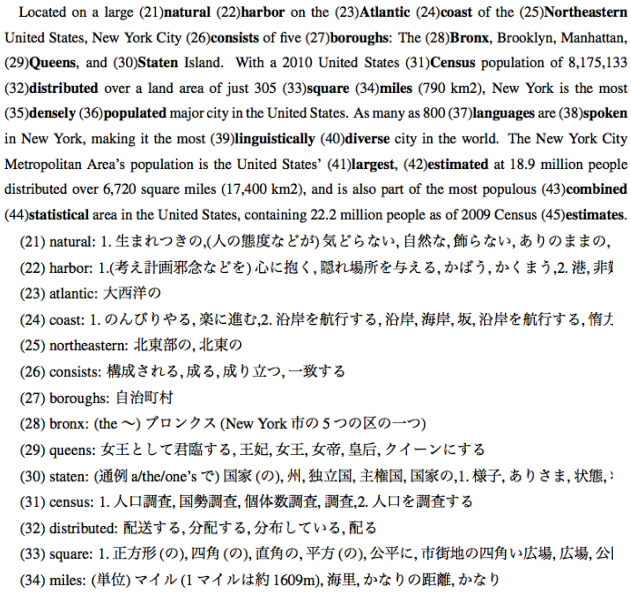
of stop words, which we obtained from the Natural Language Toolkit [1]. This list is used to exclude stop words when extracting target vocabulary terms. We also used the Gene 95 English-Japanese dictionary[2] to extract Japanese definitions for vocabulary terms. We next explain how our program works, and how it incorporates these materials into it.

3. EXERCISE CREATOR

The primary component of this software is the Exercise Creator function, which automatically creates editable lesson materials from text documents. The user loads in a document in LaTeX format. The program extracts target vocabulary words, finds their Japanese definitions, and then produces an output file in LaTeX format in which vocabulary words are highlighted and definitions are given. A user can then edit this output document.

Initially, the user creates a file in LaTeX. The user can simply cut and paste a document (e.g., an online article) into a .tex file and edit it as necessary. The file is loaded into the Lesson Material Builder. Then an output file in which target vocabulary words are selected and definitions are provided is produced. See Figure 1 for a screenshot of a portion of the output file.[3]

Once a text is input into the Lesson Material Builder, the user can implement a "Make Exercise" function. This element of the software works as follows. The input text is tokenized. For each word, the program then does the following. If the word is in the stop words list (see Table 1) or if it consists only of punctuation (e.g., symbols such as '.', ';','?', etc.), that word is not selected as a vocabulary item. If the word is still available (i.e., it is not a stop word or punctuation), then the program makes use of the unigram frequency list (see Table 1). The program extracts the top 500 most frequent non-stop-word unigrams in the unigram frequency list. If the target word is not in this list, then it is selected as a potential vocabulary word. Note that this cutoff of 500 is designed to exclude some of the most frequent English words, which students are likely to know. This cutoff can be altered; for example, the cutoff number can be increased, thereby resulting in extraction of fewer vocabulary terms. This may be useful for more advanced level students. If a word still is not eliminated, then there is a search for a definition of the extracted word in the Gene 95 English-Japanese dictionary (see Table 1). If a definition is not found, and if the word is capitalized, then it is eliminated (this serves the purpose of eliminating some names). Any remaining words are considered to be vocabulary items. Vocabulary items are stored and indexed. In the output file, each vocabulary item is printed out in bold text, with a preceding number. After the text, a list of all vocabulary items is given. Vocabulary items are preceded by the same numbers that precede the vocabulary items as they appear in the text, thus enabling a user to find the word in the text, and see how it is used in context. Each vocabulary item is followed by its definition, if a definition occurs in the dictionary.


Located on a large (21)**natural** (22)**harbor** on the (23)**Atlantic** (24)**coast** of the (25)**Northeastern** United States, New York City (26)**consists** of five (27)**boroughs**: The (28)**Bronx**, Brooklyn, Manhattan, (29)**Queens**, and (30)**Staten** Island. With a 2010 United States (31)**Census** population of 8,175,133 (32)**distributed** over a land area of just 305 (33)**square** (34)**miles** (790 km2), New York is the most (35)**densely** (36)**populated** major city in the United States. As many as 800 (37)**languages** are (38)**spoken** in New York, making it the most (39)**linguistically** (40)**diverse** city in the world. The New York City Metropolitan Area's population is the United States' (41)**largest**, (42)**estimated** at 18.9 million people distributed over 6,720 square miles (17,400 km2), and is also part of the most populous (43)**combined** (44)**statistical** area in the United States, containing 22.2 million people as of 2009 Census (45)**estimates**.

(21) natural: 1. 生まれつきの,(人の態度などが) 気どらない, 自然な, 飾らない, ありのままの,
(22) harbor: 1.(考え計画邪念などを) 心に抱く, 隠れ場所を与える, かばう, かくまう,2. 港, 非
(23) atlantic: 大西洋の
(24) coast: 1. のんびりやる, 楽に進む,2. 沿岸を航行する, 沿岸, 海岸, 坂, 沿岸を航行する, 惰
(25) northeastern: 北東部の, 北東の
(26) consists: 構成される, 成る, 成り立つ, 一致する
(27) boroughs: 自治町村
(28) bronx: (the ～) ブロンクス (New York 市の 5 つの区の一つ)
(29) queens: 女王として君臨する, 王妃, 女王, 女帝, 皇后, クイーンにする
(30) staten: (通例 a/the/one's で) 国家 (の), 州, 独立国, 主権国, 国家の,1. 様子, ありさま, 状態,
(31) census: 1. 人口調査, 国勢調査, 個体数調査, 調査,2. 人口を調査する
(32) distributed: 配送する, 分配する, 分布している, 配る
(33) square: 1. 正方形 (の), 四角 (の), 直角の, 平方 (の), 公平に, 市街地の四角い広場, 広場, 公
(34) miles: (単位) マイル (1 マイルは約 1609m), 海里, かなりの距離, かなり

Figure 1: Portion of output exercise

Definitions are extracted in the following manner. If a target vocabulary word corresponds exactly to an entry in the dictionary, then the definition is extracted. If no entry is found, then two methods are used to check for a definition. First, the target word is, via a simple algorithm, stemmed; for example, "walked" becomes 'walk'. Then there is a search for the root form of the word; if found, then the relevant definition is extracted. If an entry still is not found, then there is another search, this time for any dictionary entries that are close in form to the vocabulary word. If no definition is found, then no definition is provided in the output file.

Consider the following text, an excerpt from a Wikipedia file on New York City.[4]

> Located on a large natural harbor on the Atlantic coast of the Northeastern United States, New York City consists of five boroughs: The Bronx, Brooklyn, Manhattan, Queens, and Staten Island. With a 2010 United States Census population of 8,175,133 distributed over a land area of just 305 square miles (790 km2), New York

[2] http://www.namazu.org/~tsuchiya/sdic/data/gene.html
[3] The output file is in .tex format. The screenshot is of the .pdf version of the output. See the following discussion.
[4] `http://en.wikipedia.org/wiki/New_York_City`

> is the most densely populated major city in the United States. As many as 800 languages are spoken in New York, making it the most linguistically diverse city in the world. The New York City Metropolitan Area's population is the United States' largest, estimated at 18.9 million people distributed over 6,720 square miles (17,400 km2), and is also part of the most populous combined statistical area in the United States, containing 22.2 million people as of 2009 Census estimates.

This text is fed through the Exercise Creator. The program searches through the text, finds target vocabulary items and their definitions, and a vocabulary list, consisting of vocabulary words and their definitions, is produced following the text. The resulting paragraph and a portion of the resulting vocabulary list are shown in Figure 1, which consists of screenshots of the .pdf file version of relevant portions of the output.[5]

In Figure 1, note that target vocabulary words are labeled and definitions are generally provided. Consider the word 'harbor', which is labeled as (22). This word is extracted because it is not a stop word and it is not in the list of the most frequent 500 non-stop-words in the unigram frequency list. This word is found as an entry in the dictionary. Thus, this word is tagged in the file, it is printed out in the vocabulary list below the text, and its extracted definition is provided. The definition is the entire extracted definition. Next, consider the word 'boroughs' (27). The singular version 'borough' appears in the dictionary, but the plural version does not. To extract the definition, the program first searches for 'boroughs'. When this is not found, this word is stemmed to 'borough', which matches an entry in the dictionary. The name 'Brooklyn', which appears in this paragraph does not occur in the vocabulary list. This is because it is identified as a name - it is not an entry in the dictionary, and it is capitalized. This contrasts with 'Bronx' (28), which just so happens to be in the dictionary, and thus, it is extracted.

Table 2 lists some vocabulary words from the above paragraph on New York City, together with their extracted definitions and the 'best' definition, where the best definition refers to the definition that corresponds to the appropriate word sense that corresponds to how the word is used in the text. Note that (21) 'natural' is provided with many definitions. Of these, the best one that corresponds to its use in the text is 自然な *shizenna* 'natural in nature'. Some of the definitions, such as 生まれつきの *umaretsukino* 'from birth' are not appropriate given the context. With respect to (27) 'boroughs', the program only finds one definition 自治町村 *jichichouson* 'borough' and this definition turns out to be appropriate. Interestingly, a definition is found for (28) 'Bronx', even though this is the name of a location. The definition provided, which describes that this is one of the 5 boroughs of New York City, is correct. A problem, however, arises for (29) 'queens', the name of a borough in New York. The program finds definitions that refer to a female monarch. None of these definitions are appropriate. The program extracts (39) 'linguistically', but no definition is found in the dictionary. Editing of the definitions in the output file is thus necessary.

Table 2: Definitions

Word	Definition	Best Definition
(21) natural	1. 生まれつきの,(人の態度などが), 気どらない, 自然な, 飾らない,2. ありのままの, 自然の, 加工しない, 野生の, うってつけの物 [人], ぴったりの物 [人], 本質, 原始状態,3. 自然, 天然, 全宇宙, 性質, 気質, 本位の, 本当の, 先天的な	自然な
(27) boroughs	自治町村	自治町村
(28) bronx	(the 〜) ブロンクス (New York 市の 5 つの区の一つ)	(the 〜) ブロンクス (New York 市の 5 つの区の一つ)
(29) queens	女王として君臨する, 王妃, 女王, 女帝, 皇后, クイーンにする	
(39) linguistically		

The Exercise Creator function is useful in that it aids in the process of developing target lesson materials for helping students with their reading skills. A human can take a text passage, find target vocabulary items, look up the definitions and then produce a document of the sort that is output by this program. However, this process is time consuming. This software makes this process much faster. This application automatically determines target vocabulary items. Definitions are automatically extracted and an editable LaTeX document is produced. A user who has basic skills in using LaTeX can then edit this output document. Certain vocabulary items that have been selected automatically by the program can be eliminated from the list of vocabulary items, if necessary - for example, the students may already know the item. Definitions, as noted above, must be edited. When appropriate definitions or no definitions are provided for target vocabulary items, the user can manually look up the word and find a suitable definition. However, in cases of input documents such as the one presented in this paper, most definitions are provided automatically. In the sample output document discussed here (see Figure 1 for a portion of this document), 114 vocabulary items were selected and definitions were found for all but 7 of these items. Of words for which definitions were found, in our opinion, there were 8 words for which certain problems arise (definitions are incorrect, etc.): 'referred', 'queens', 'Staten', 'Dutch', 'surrounds', 'wall', 'lower', and 'capitalization'.[6] While editing these definitions takes time,

[5]Due to lack of space, we do not show the entire output file.

[6]Definitions are found for 'referred', but they do not correspond to the use in the text. The definitions for 'queens' refer to a monarch, and not to the place in New York. Definitions are given for 'Staten' and 'Dutch' but they do not refer to the proper noun uses, as in 'Staten Island' and in

it takes much less time than manually looking up definitions for all vocabulary words. Also, the user does not have to spend much to any time (depending on the accuracy of the output, target students, etc.) determining which words to provide definitions for, since words are automatically selected by the program.

4. QUIZ CREATOR

Another function of this program is that it can be used to automatically create vocabulary quizzes from reading lesson materials. Specifically, the user loads in a file that corresponds to the type of file that is created by the Exercise Creator function (see Figure 1 above) - a document containing text in which vocabulary words are selected and definitions are provided.

The quizzes are created in the following manner. The user loads in the .tex source file. In this case, we loaded in an edited version of the file output by the Exercise Creator (a portion of the unedited output file is shown in Figure 1). Then the user selects a "Make Quiz" option. The program loads in the file and extracts all of the vocabulary words. Then the user chooses the number of vocabulary words to use for the quiz. The user can also choose not to use certain vocabulary words, if desired. Then the program automatically produces a quiz. A portion of the output quiz is shown in Figure 2. In this case, there are 100 possible vocabulary words and the user has chosen to make a quiz with 25 of these. The program chooses 25 of the vocabulary words. The original text is printed to the quiz, with the 25 vocabulary words in bold and numbered. The text is provided so that students can see the words in context. The program creates a table at the bottom of the text. The list contains the vocabulary words in one column and 25 randomized definitions in an adjacent column. Students have to match the vocabulary words with their definitions. In addition to creating a .pdf version of the quiz, the program also creates a version for use with the course management system Moodle (http://moodle.org/). This quiz is automatically created in Moodle XML format, which, when loaded into Moodle, produces an online quiz (see Figure 2). While looking at the paper quiz, students can enter their answers in the Moodle quiz, and the quizzes are automatically graded by Moodle.

This Quiz Creator has the following advantages. First, it automatically creates quizzes, thus saving the teacher time. Second, it creates quizzes by randomly selecting vocabulary items (the user determines how many), and thus it can be used to create many different quizzes. This can be useful when teaching different sections of the same class. The teacher can give each section of the same class the same lesson materials to study, but each section can be given a different quiz. Another advantage of this software is that it can be used to create a version of the quiz that can be given using Moodle. If students input their answers into Moodle, then their quizzes can be automatically graded by Moodle. Automatic grading for vocabulary quizzes of this sort is easy to implement on a computer since there is only 1 correct answer per quiz item (that is, if the quiz item does not have any flaws) and there is no actual need to read and comprehend English for grading. Automatic grading saves the teacher time. Instead of spending time with the tedious job of grading vocabulary quizzes, the teacher can focus on other tasks.

relating to the Netherlands. Definitions for 'surrounds' are not appropriate; the extracted definitions refer to the verbal usage of this word, but this word is used as a noun in the text. An appropriate definition for 'wall' is given, but 'wall' occurs in the name 'Wall Street', so it may be better to indicate 'Wall Street' as a vocabulary term than 'wall' alone. Definitions are extracted for 'lower' but they do not correspond to its use in the place name 'Lower Manhattan'. The definition that is extracted for 'capitalization' refers to its use with respect to capital letters, and not the financial use which appears in the text.

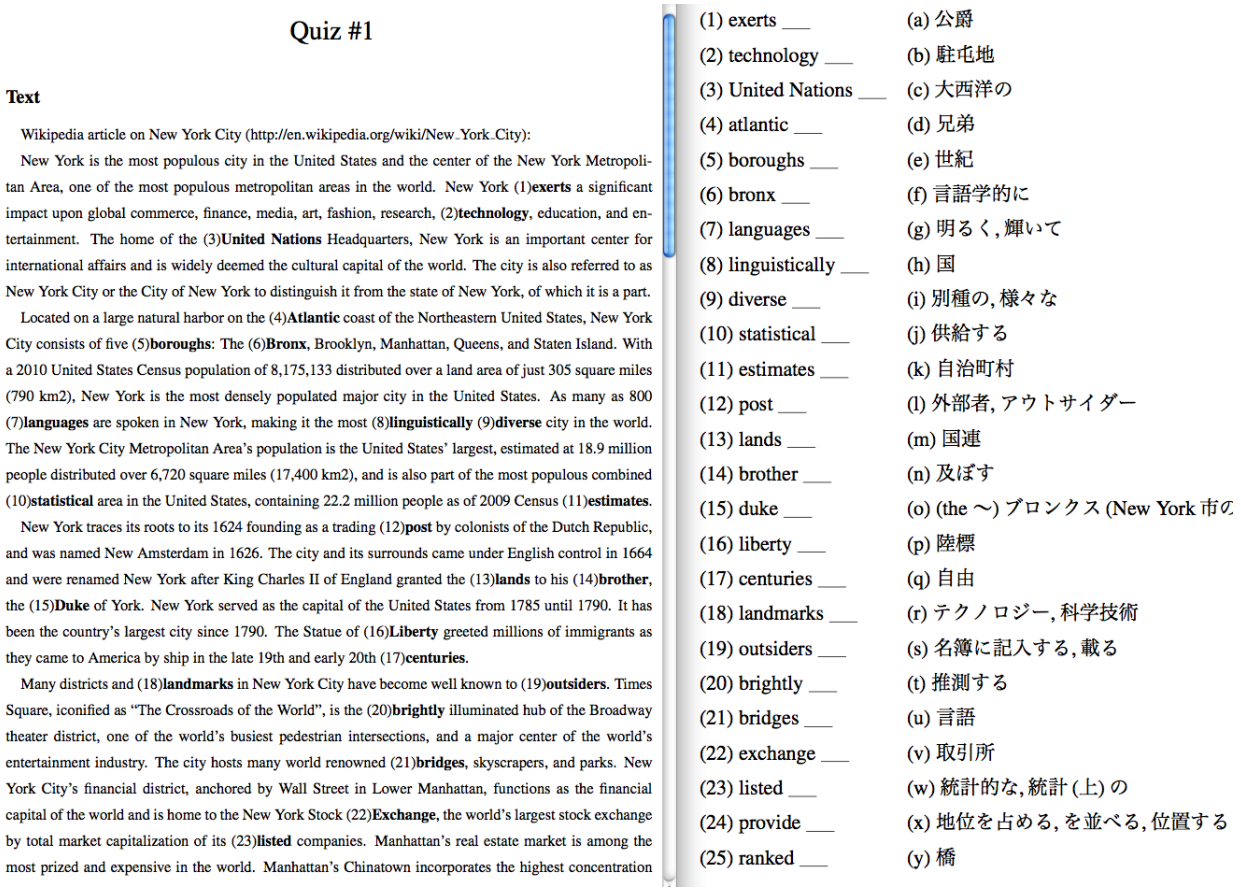
5. CONCLUSION

In this paper we have discussed a Lesson Material Builder that can be used to create editable lesson materials and to create vocabulary quizzes. This is a work in progress and there are numerous ways in which this application could be improved. The Exercise Creator function extracts vocabulary words and provides definitions. Currently, the program simply provides definitions, from a text dictionary, for each word, if a definition is available. This function, while useful, would be more useful if it were able to automatically provide appropriate definitions for words (e.g., instead of giving all definitions for a vocabulary item, it would provide the definition that corresponds to the appropriate word sense). Furthermore, the application extracts vocabulary terms that are unigrams; the ability to identify and define multi-word expressions would be useful. Development towards these goals could be achieved via incorporation of word-sense-disambituation, named entity recognition, and part-of-speech tagging functionalities. The Vocabulary Quiz function currently only creates quizzes that require students to match vocabulary words with their definitions. This function could be improved to create other types of vocabulary quizzes.

Lastly, while the Lesson Material Builder is designed to serve as an aid to English language teachers who are teaching to students who speak Japanese as a native language, this type of application could be adapted for other teaching situations. It could be designed for speakers of a native language other than Japanese. This is fairly easy to implement, if an available dictionary for that language can be found. It could also be redesigned to provide definitions in English - this might be ideal in an English language class in a country where English is spoken, since oftentimes these classes have students who speak a diverse range of native languages. Lastly, the methods discussed here are not necessarily limited to teaching English. This sort of application could be designed for teaching languages other than English, as well as for designing lesson materials for other non-language academic subjects.

6. REFERENCES

[1] S. Bird, E. Klein, and E. Loper. *Natural language processing with Python*. O'Reilly, Sebastopol, CA, 2009.



Quiz #1

Text

Wikipedia article on New York City (http://en.wikipedia.org/wiki/New_York_City):

New York is the most populous city in the United States and the center of the New York Metropolitan Area, one of the most populous metropolitan areas in the world. New York (1)**exerts** a significant impact upon global commerce, finance, media, art, fashion, research, (2)**technology**, education, and entertainment. The home of the (3)**United Nations** Headquarters, New York is an important center for international affairs and is widely deemed the cultural capital of the world. The city is also referred to as New York City or the City of New York to distinguish it from the state of New York, of which it is a part.

Located on a large natural harbor on the (4)**Atlantic** coast of the Northeastern United States, New York City consists of five (5)**boroughs**: The (6)**Bronx**, Brooklyn, Manhattan, Queens, and Staten Island. With a 2010 United States Census population of 8,175,133 distributed over a land area of just 305 square miles (790 km2), New York is the most densely populated major city in the United States. As many as 800 (7)**languages** are spoken in New York, making it the most (8)**linguistically** (9)**diverse** city in the world. The New York City Metropolitan Area's population is the United States' largest, estimated at 18.9 million people distributed over 6,720 square miles (17,400 km2), and is also part of the most populous combined (10)**statistical** area in the United States, containing 22.2 million people as of 2009 Census (11)**estimates**.

New York traces its roots to its 1624 founding as a trading (12)**post** by colonists of the Dutch Republic, and was named New Amsterdam in 1626. The city and its surrounds came under English control in 1664 and were renamed New York after King Charles II of England granted the (13)**lands** to his (14)**brother**, the (15)**Duke** of York. New York served as the capital of the United States from 1785 until 1790. It has been the country's largest city since 1790. The Statue of (16)**Liberty** greeted millions of immigrants as they came to America by ship in the late 19th and early 20th (17)**centuries**.

Many districts and (18)**landmarks** in New York City have become well known to (19)**outsiders**. Times Square, iconified as "The Crossroads of the World", is the (20)**brightly** illuminated hub of the Broadway theater district, one of the world's busiest pedestrian intersections, and a major center of the world's entertainment industry. The city hosts many world renowned (21)**bridges**, skyscrapers, and parks. New York City's financial district, anchored by Wall Street in Lower Manhattan, functions as the financial capital of the world and is home to the New York Stock (22)**Exchange**, the world's largest stock exchange by total market capitalization of its (23)**listed** companies. Manhattan's real estate market is among the most prized and expensive in the world. Manhattan's Chinatown incorporates the highest concentration

(1) exerts ___ (a) 公爵
(2) technology ___ (b) 駐屯地
(3) United Nations ___ (c) 大西洋の
(4) atlantic ___ (d) 兄弟
(5) boroughs ___ (e) 世紀
(6) bronx ___ (f) 言語学的に
(7) languages ___ (g) 明るく, 輝いて
(8) linguistically ___ (h) 国
(9) diverse ___ (i) 別種の, 様々な
(10) statistical ___ (j) 供給する
(11) estimates ___ (k) 自治町村
(12) post ___ (l) 外部者, アウトサイダー
(13) lands ___ (m) 国連
(14) brother ___ (n) 及ぼす
(15) duke ___ (o) (the ～) ブロンクス (New York 市の
(16) liberty ___ (p) 陸標
(17) centuries ___ (q) 自由
(18) landmarks ___ (r) テクノロジー, 科学技術
(19) outsiders ___ (s) 名簿に記入する, 載る
(20) brightly ___ (t) 推測する
(21) bridges ___ (u) 言語
(22) exchange ___ (v) 取引所
(23) listed ___ (w) 統計的な, 統計 (上) の
(24) provide ___ (x) 地位を占める, を並べる, 位置する
(25) ranked ___ (y) 橋

Figure 2: Portion of output quiz: .pdf file

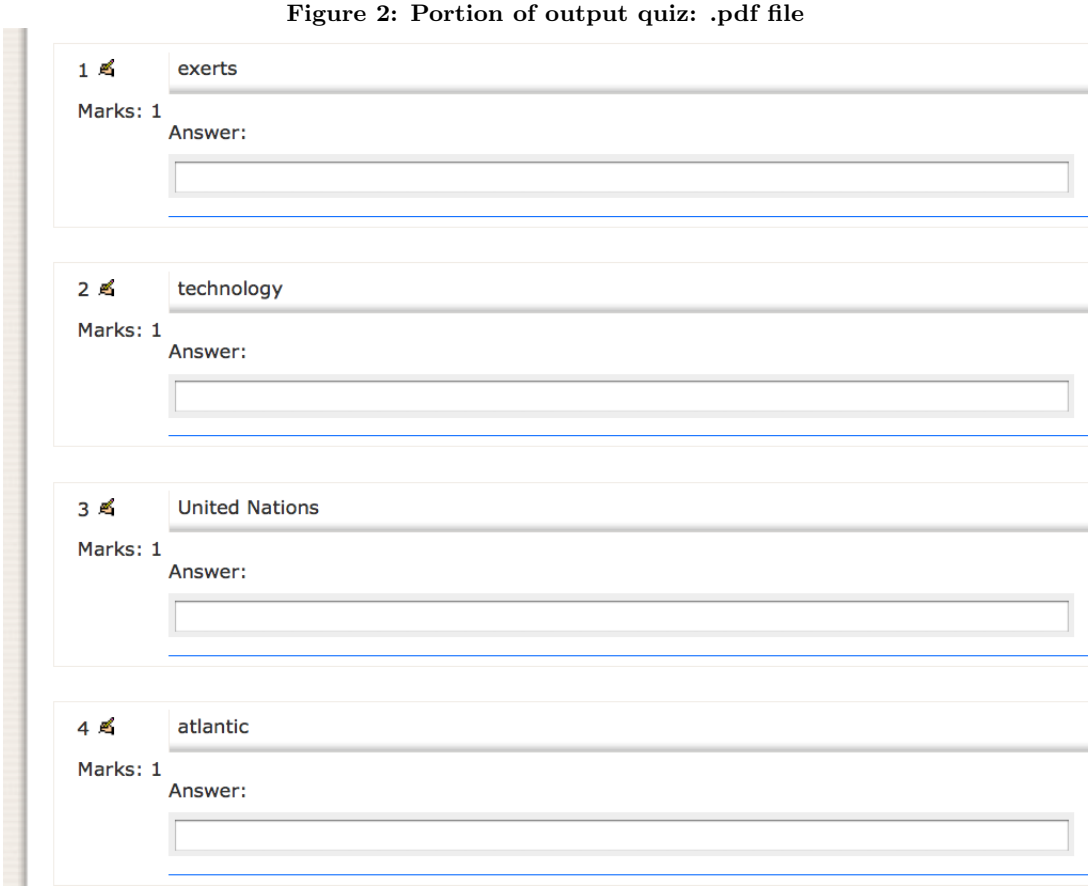


Figure 3: Portion of Moodle Quiz