Multiple Affordances of Language Corpora for Data-driven Learning, in Agnieszka Leńko-Szymańska, A. and A. Boulton (Eds.) John Benjamins Publishing Company: Amsterdam, pp. 109-128.
(First draft)

A corpus and grammatical browsing system for remedial EFL learners

Kiyomi Chujo
College of Industrial Technology, Nihon University

Kathryn Oghigian
Faculty of Science and Engineering, Waseda University

Shiro Akasegawa
Lago Institute of Language

Abstract

To address the need for corpora and corpus tools accessible to low-proficiency level EFL language students, we have created a free, grammatically-categorized browsing system based on a collection of copyright-free level-appropriate sentences called the Sentence Corpus of Remedial English (SCoRE). Teachers and students can search the database of sentences by grammatical category or target word to see complete example sentences which follow structural and lexical parameters identified as particularly relevant for Japanese EFL students. This database is based on a 30-million-word corpus from English secondary school textbooks used in Asian countries, American reading textbooks, English graded readers, and web-based children's news articles. This paper describes the creation of the Grammatical Pattern Profiling System (GPPS) browsing program and SCoRE, and discusses pedagogical applications.

Keywords

beginner level; SCoRE; EFL example sentences; grammatically-categorized corpus; GPPS; low proficiency; sentence-concordances

1. Appropriate level, needs-driven corpora for the EFL classroom

Second language proficiency is generally measured in Japan using TOEFL and/or TOEIC tests. Ranked average test score data for the TOEFL iBT for 2013 shows Japan near the bottom of all Asian countries (see Table 1), and

well below the majority of European countries (Educational Testing Service 2014). Similarly, Japan ranks 39th out of 45 countries in mean performance on the TOEIC (Educational Testing Service 2012). In his discussion on how low test results relate to educational policy in Japan, Yoshida (2008: 3) presented results from a 2004 National Institute for Education Policy Research study indicating that 53% of third-year junior high school students reported that they understood half or less of what was being taught in their English lessons. Not surprisingly, then, a study by Ono et al. (2005) found that first-year university students lacked knowledge of basic grammar that they were supposed to have learned in junior and senior high school. There have been numerous reforms in Japanese education over the last few decades, including the implementation of the JET Programme in 1987 in which thousands of native English-speaking university graduates have been hired to assist in classroom lessons in junior and senior high schools throughout Japan, to counter what was perceived as a rote-memorization and grammar-oriented approach with a more communicative approach (see JET Programme 2010). Based on the TOEFL scores seen in Table 1, neither this nor other reforms seem to have been particularly successful.

Table 1: Average TOEFL iBT test scores for the three highest and lowest ranked Asian and European countries, 2013

Asian Countries	TOEFL iBT
Singapore	98
India	91
Pakistan	90
…	
Japan & Mongolia	**70**
Cambodia	69
Tajikistan & Lao People's Dem. Rep.	68
Timor-Leste	62

European Countries	TOEFL iBT
Netherlands & Austria	100
Denmark	98
Belgium & Luxemburg & Switzerland	97
…	
Montenegro	79
Armenia	77
Kosovo & Turkey	76

With advances in technology and multimedia opportunities in education, another approach might be with data-driven learning (DDL). This kind of corpus linguistics methodology has been shown to have benefits (Gavioli & Aston 2001; Braun 2005; Huang 2008; Chujo et al. 2013; see also Flowerdew, this volume). However, it has a long way to go before being widely accepted in the mainstream second language (L2) classroom, in part because currently available corpora are not necessarily appropriate for low-proficiency learners, and because creating these resources is difficult and time-consuming. General corpora such as the British National Corpus (BNC 2007), and specific corpora such as the Michigan Corpus of Academic Spoken English (MICASE 2007) or the journal-based Springer Exemplar (<http://www.springerexemplar.com>) are often cited in studies successfully using DDL with intermediate and advanced learners. Very few

successes have been reported with beginner-level learners – not surprisingly, since appropriate corpora and corresponding material are difficult to obtain. In an investigation of 64 copyright-free e-texts, Chujo et al. (2007: 67) found that there was "an unfortunate shortage of copyright available e-texts at the beginner level." Only one title contained vocabulary understood by the average Japanese high school graduate at a 95% word coverage level (i.e. where known vocabulary would cover 95% of words encountered), this being postulated as the threshold for minimal reading comprehension of a text (Laufer 1992).

Gavioli and Aston (2001) have remarked on the need for teacher-selected or pre-edited graded or 'easy' concordances; this is underscored by Breyer (2009) who reported that 61% of teachers in her study were unable to find a corpus that was appropriate in topic or difficulty level for their students to use. She also reported that more teachers would use DDL to teach grammar if these materials were more readily available. Although many teachers have relied on the BNC, Allan (2009) points out that this corpus presents unfamiliar topics that are cut off from everyday life and that the truncated concordance lines visible in the usual KWIC (key word in context) format are difficult for students to manage. Similarly, lower-proficiency students may have difficulty with grammatically-complex concordance lines and colloquial usage found in general corpora such as the Corpus of Contemporary American English (COCA; see Davies 2008-).

Clearly, if DDL is to be considered for low-proficiency learners, there is a need to rethink available corpora, and perhaps the standard use of concordance lines. In an investigation of the "methodical challenges" of integrating corpora in secondary education, Braun (2007: 316) concluded that "it is time for a move from data driven learning (DDL) to needs-driven corpora, activities and methodologies." Toward that end, in an effort to increase efficiency and lessen the learning load required for grammatical items, Minn et al. (2005: 101) suggested including more usage data; however, they also noted that:

> Because of the labor-intensive nature of creating teaching material, large amounts of varied material cannot be made in a short time; …the quality of material largely depends on the creators' ability; [and]… many of the creators of such material are not native English speakers, so the expressions included tend to be lacking in variety, and the quality of example sentences cannot be immediately guaranteed.

Minn et al. therefore created their own commercial website[1] to provide example sentences corresponding to English sentence patterns appearing in

[1] Bunpou Koumokubetsu BNC Youreishuu ('webpage for downloading BNC example sentences corresponding to chosen grammatical items'): <http://bnc.jkn21.com/search/login_ncube.cgi>.

secondary school textbooks, but these were extracted from the BNC and the sentences are therefore not ideal for the intended students. Although other English corpora do exist, some also exemplifying the structures in school textbook grammar (e.g. Tanaka et al. 2008), they are generally limited to high-level texts and are not ideal for low-proficiency learners.

The purpose of this paper is to describe two new resources created especially for low-proficiency students. The first is a new DDL tool, called the Grammatical Pattern Profiling System (GPPS). This is a free, web-based browsing program with a simple user-friendly interface in which the results appear as complete sentences rather than in KWIC format. The GPPS can be used to view and download example sentences of particular target grammar structures for use directly in the classroom or as a source of examples for prepared activities. The second is a database of example sentences called the Sentence Corpus of Remedial English (SCoRE) on which the GPPS is based. The sentences appear in three distinct levels (beginner, intermediate and advanced), and have been constructed according to criteria of reading grade, word familiarity and sentence length. Example sentences are being continually added to the grammatical categories, and the website for the GPPS and SCoRE will be made public as more data becomes available.

The creation of the GPPS is described in detail in the next section including its rationale, how grammatical categories were chosen, and how the search expressions were written. In Section 3, the creation of SCoRE is explained, along with how appropriate texts were sourced, the method for defining target level and three distinct proficiency levels, how sentence length was determined for each level, the rationale and procedure for creating the SCoRE sentences, and how the L1 (Japanese) translations were produced. Section 4 explores pedagogical applications, while Section 5 outlines the limits of this study.

2. Developing the Grammatical Pattern Profiling System (GPPS)

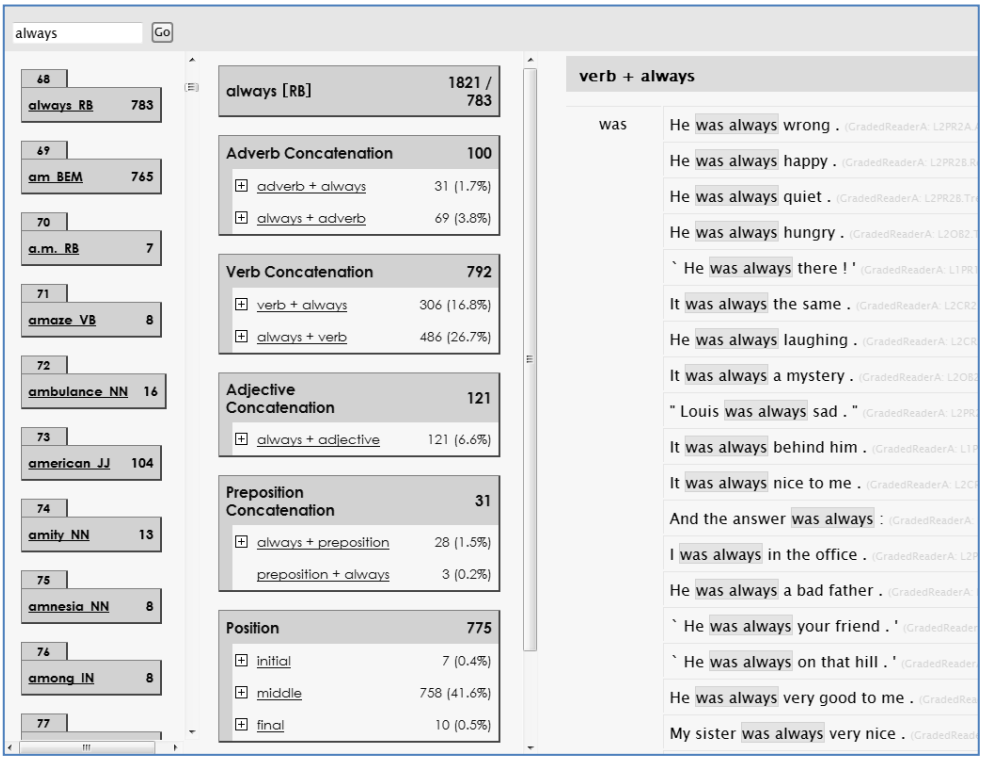
The goal for developing the GPPS was to create a tool that was very easy and intuitive for both students and teachers to use, and was web-based so it could be used in and out of the classroom without cost or registration. In order to be as practical as possible for low-proficiency learners, it was considered important that students be able to view complete and level-appropriate sentences. To be useful to teachers and materials writers, the browsing system was organized by grammatical category. Each step of this process is described in detail in the following section.

2.1. Using LWP-GRC as a model for the GPPS

Various aspects of the GPPS were modeled after a lexical profiling program called the LagoWordProfiler (LWP; for more technical detail, see Chujo,

Akasegawa et al. 2012). LWP for a Graded Reader Corpus (LWP for GRC) is a browsing system which provides example sentences by a grammatical pattern of a single search word; Figure 1 gives the example of *always*. It was developed based on a set of collocations and colligations in sentences extracted from a two-million-word graded reader corpus. As can be seen in Figure 1, such a graded corpus provides accessible examples for low-proficiency students, but the browser itself was designed for use by teachers and materials writers.

Figure 1: LWP for GRC – Colligation/collocation information on the adverb always



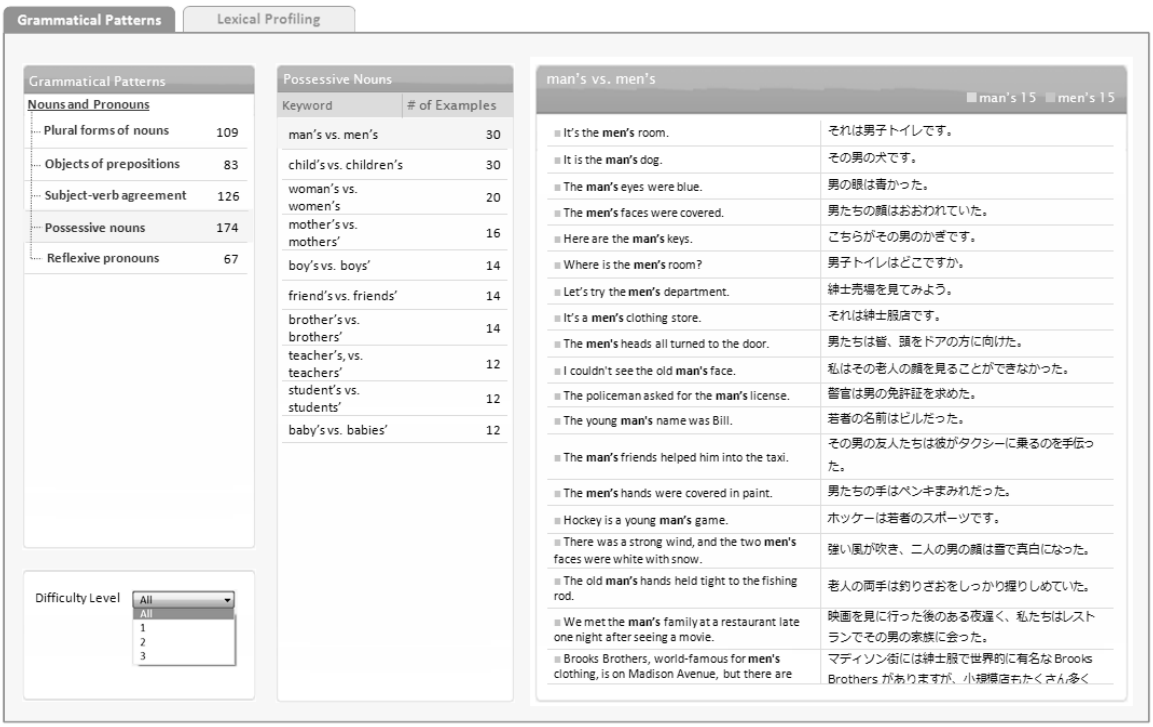
EFL teachers evaluating this program in a pilot workshop in Japan provided favorable feedback. They reported that the interface was user-friendly, and that the example sentences extracted from the graded readers were at an appropriate level for their teaching material. The limitations of the LWP for GRC were firstly that the example sentences were extracted by search word, not by grammatical category. The teachers preferred grammatical category-based searches both for creating teaching materials and for use by students for corpus-based language learning. Secondly, the LWP for GRC example sentences were extracted from commercial graded readers and were subject to copyright and therefore had limited application in materials development. Thirdly, the corpus was only two million words in size and needed to be expanded in order to collect additional 'easy' texts to provide enough example sentences to cover the variety of grammatical items. As a result of this feedback on the LWP, the GPPS was designed on a similar infrastructure, but to allow searches based on grammatical categories as well

as search words; and SCoRE was created with our own, copyright-free sentences based on a source corpus of 30 million words.

2.2. GPPS functionality

A screenshot of the GPPS is presented in Figure 2. The two uppermost left tabs show that the GPPS allows searches by grammatical pattern or word ("lexical profiling"). The screenshot shows the grammatical pattern for possessive nouns *man's* and *men's*. In the far left column, teachers and materials writers can view a hierarchy of related patterns and the number of example sentences that exist in the corpus for each pattern. In the second column, users can choose a particular lexical realization of the grammatical item. In the third and largest column, example sentences are given with L1 translations. The difficulty level can be chosen from a box on the bottom left.

Figure 2: GPPS screenshot showing grammatical categories for possessive nouns, and example sentences in three levels for man's *vs* men's



2.3. Selection of grammatical categories

Although it is possible to identify high-frequency grammar patterns from various corpora such as the BNC or COCA or from resources such as the *Longman Grammar of Spoken and Written English* (Biber et al. 1999), the focus here was on particular patterns identified as weak areas in our target population (low-proficiency Japanese senior high school students and first-

year university students). Grammatical categories identified in previous studies were mostly based on Japanese school textbook grammar items (e.g. Minn et al. 2005) or grammar items frequently targeted in certain standardized tests (e.g. Uchibori & Chujo 2005). The grammatical categories used for the GPPS were chosen by Chujo, Yokota et al. (2012, based on Murphy and Smalzer 2009, 2011) as being particularly relevant. In that 2012 study, a basic grammar proficiency test was created based on an investigation of English proficiency levels of junior high school students carried out by Shirahata (2008) and on an investigation of specific grammar weaknesses of high school students taking the TOEIC carried out by Uchibori et al. (2006). Test items that were incorrectly answered by more than 30% of the participants (first-year university students) in the 2012 study were selected for inclusion in the GPPS. In this way, it was possible to more accurately identify what was missing from the knowledge base of targeted students, and to use this as the basis for the GPPS rather than what was more frequent in general, native-speaker corpus data. Some examples of the chosen grammatical items are shown in Table 2. The percentages given after each item refer to the percentage of incorrect answers obtained from the 2012 study.

Table 2: Examples of targeted remedial grammatical items

Junior high school grammar items		Senior high school grammar items	
1	Possessive pronouns (47%)	1	Subjunctives (79%)
2	Plural forms of nouns (44%)	2	Relatives (61%)
3	Present perfect (43%)	3	Prepositions (60%)
4	Indirect questions (42%)	4	Negation (61%)
5	Passive (41%)	5	Conjunctions (50%)
6	Negation (37%)	6	Auxiliaries (45%)
7	Existential phrase (34%)	7	Gerunds (39%)
8	Tense (34%)	8	Adverbs (38%)

2.4. Creation of search expressions and patterns

Because a sentence typically contains multiple grammatical patterns, to enable the GPPS to search by any of the grammatical items assigned to a sentence, each sentence is tagged for its grammatical patterns in the form of XML attributes, as shown in Figure 3 where the sentence *people are living longer now* includes the progressive tense *are living* and the comparative *longer*. Multiple grammatical patterns can be added using the semicolon separator, as shown in the example sentence here as *grammatical pattern = 'progressive; comparative'*.

Figure 3: An example of sentence data for people are living longer now

```
<example id="00001" keyword="live;people" grammatical_pattern ="progressive;
comparative" learning_level="A">
    <english>People are living longer now.</english>
    <japanese>人は以前よりも長生きするようになっている。</japanese>
</example>
```

3. Developing the Sentence Corpus of Remedial English (SCoRE)

We have called this database a 'sentence corpus' so users can differentiate it from a more traditional corpus of whole texts accessible via a KWIC presentation of truncated concordance lines. In order to produce a collection of example sentences to use in the GPPS, first the concept of text level was defined, then various texts were sourced and evaluated for appropriateness. Next, sentence length, three distinct proficiency levels and the optimal number of sentences were determined, and finally tailor-made sentences and translations were created. Each step is described in this section.

3.1. Defining target population proficiency levels

In order to build this specialized corpus, a common denominator was necessary to compare target population proficiency levels with potential corpus sources. Two indices were used to define texts at appropriate levels for the target population. These were US reading grade level and US word familiarity level, since these have been shown to be applicable and reliable for measuring the linguistic difficulty of English text (see Chujo et al. 2007, 2011). A text's reading grade level refers to the US school grade at which an average native-English-speaking child would be able to read and understand this particular text, measured here by the Flesch-Kincaid Formula (Micro Power and Light Co. 2003). Word familiarity grade level means at what US grade an average native-English-speaking child would understand the vocabulary of a text, as calculated using the data of Dale and O'Rourke (1981) and Harris and Jacobson (1972). Textbooks used in Japanese high schools were evaluated with these indices, and the results showed that the average Japanese junior high school English textbook corresponds to US school grades 2 and 3 (commonly ages 7 and 8), and that the average Japanese senior high school textbook corresponds to US school grades 4 and 5 (ages 9 and 10). Japanese remedial students (i.e. who failed to acquire the grammar and vocabulary taught in high school) generally do not advance beyond US school grades 4 and 5. Thus in order to create corpus data appropriate for lower-level students, the reading grade and word familiarity levels up to US school grade 5 were targeted. (For a more in-depth investigation of these indices, see Chujo et al. 2007, 2011.)

3.2. Sourcing potential corpus data

Once it was determined that the desired level of corpus data was from US school grades 1 to 5, potential corpus sources were located and evaluated with the same indices. Chujo et al. (2007, 2011) and Chujo, Nishigaki et al. (2012) examined four types of text which included: (1) American reading textbooks from grades 1 to 3; (2) English graded readers allotted the Yomiyasusa Level (YL) from 0.0 to 4.0 (a reading level of English books for Japanese students; Furukawa 2007); (3) English secondary school textbooks used in Asian countries; and (4) an 'authentic' English text collection (Utiyama & Takahashi 2003). ('Authentic' here refers to L1 texts produced for L1 readers as a whole.) These four types of text were evaluated for reading grade and word familiarity levels. Results showed that for reading grade level, a Japanese senior high school graduate would generally be expected to be able to follow American reading textbooks from grades 1 to 3; however, in looking at word familiarity, a high school graduate would not be able to understand approximately one fourth of the vocabulary of American reading textbooks from these grades. On the other hand, English graded readers seem to be appropriate for Japanese low-proficiency learners based on both reading and word familiarity. English textbooks used in Asia would also be appropriate although some vocabulary might be new to learners. However, in the authentic text collection, both reading grade and word familiarity indices show that Japanese students would have difficulty using authentic text, so these resources were not used.

A 30-million-word source corpus was therefore created from American reading textbooks for grades 1 and 2, English graded readers with a YL of 0 to 4, and English textbooks used in Asia. In addition, other resources found to be within the levels as outlined above were also included, such as website news stories for elementary-grade children (see Teaching Kids News at <http://teachingkidsnews.com/grades-2-8>).

3.3. Defining sentence length

Because the SCoRE database would comprise complete sentences, rather than whole texts accessible via KWIC presentation of truncated concordance lines, optimal sentence length was important. To calculate how a sentence is defined as beginner, intermediate or advanced level, several indices such as word length, sentence length, readability scores, and US word familiarity grade level have been shown to be effective (see Chujo et al. 2007, 2011). From these, two indices (sentence length and word familiarity) were chosen as most applicable to the present project as they evaluate the level of single sentences rather than a whole text (see Chujo, Nishigaki et al. 2012). Three-level distinctions were created so that students would be able to more easily understand the targeted grammar items and build confidence with beginner-level sentences and vocabulary, and then might be challenged with slightly longer and more difficult sentences. For

the purposes of SCoRE and the GPPS, sentence length for beginner/remedial level was established as eight words or less, intermediate level was from five to eleven words, and advanced level was longer than nine words. For word familiarity, beginner/remedial level included vocabulary from US school grades 1 and 2, intermediate level grades 1 through 3, and advanced level grade 4 and beyond. Example sentences are shown in Table 3.

Table 3: Example sentences for the passive voice using called

Beginner/Remedial level (8 words or less)	Intermediate level (5-11 words)	Advanced level (9 words or more)
What is it called?	What will their next CD be called?	The American School in Japan is usually called ASIJ.
My youngest son is called Bob.	My little brother was called Tommy by his friends.	Over the years he had been called many names.
What is this song called?	In the U.S., these are called shorts, not short pants.	I asked what this thing was called but no one knew.
What is your dog called?	A gardenia is called a *kuchinashi* in Japanese.	She has been called a genius by her contemporaries.
This game is called cricket.	Policemen are sometimes called cops.	A man whose wife has died is called a widower.
The teacher was called "Coach".	A lawyer is sometimes called an attorney.	What are the different phases of the moon called?

3.4. Defining the number of sentences

In order to decide the optimal number of sentences for each grammatical feature, student feedback from other DDL studies conducted over the last eight years was used. Students who studied basic grammar using a traditional KWIC format with ParaConc (Barlow 2004) were asked how many example lines they preferred to view. Based on data from Chujo, Oghigian et al. (in press), 68% of students preferred ten example sentences as a basis for observing patterns and making inferences, 23% preferred twenty sentences, 6% preferred five sentences, and 3% preferred fifty sentences. Thus we chose 10 for each level, with 30 example sentences for each grammatical pattern in order to provide a sufficient number for materials writers.

3.5. Using the source corpus as a model for SCoRE

Once the 30-million-word source corpus was complete, each targeted grammatical feature was analyzed to determine its most common patterns. Table 4 shows four grammatical patterns by way of example. One of the

most difficult patterns for lower-level Japanese students is the subjunctive. From the source corpus, it was determined that the most frequent verbs appearing in the subjunctive past included *be*, *go*, *want*, *get*, and *come*. This analysis indicated which verbs would need to be included in SCoRE, such as *I would buy it if it were cheaper* for the subjunctive past + *be.* For the subjunctive *wish*, the most frequent pattern was *wish * could,* followed by *be*, *have*, *would* or *know*. The highest frequency patterns for relative sentences were *a person who*, *a man who*, *someone who*, etc. The most frequent verb used in the passive voice was *got*, followed by *been*, *seen*, *gone*, *done*, *come* and others.

Table 4: Examples of high frequency parameters extracted from the source database for grammatical patterns

	Grammatical pattern	High frequency words
1	Subjunctive past	be, go, want, know, get, come, live, try, find, make, think
2	Subjunctive wish	could, be, have, would, know
3	Relative patterns	a person who, a man who, someone who, a thing that, a woman who
4	Passive voice verbs	got, been, seen, gone, done, come, made, given, lost, changed

Although the grammatical patterns were chosen based on the needs of the target population rather than by frequency in a native-speaker corpus, they were verified for structural authenticity with COCA. For example, the grammatical pattern * *wish * could tell* * was checked in COCA to confirm that it appears frequently in authentic texts (over 100 occurrences in this case).

Next the sentences in the source corpus were examined for suitability. Although they were taken from level-appropriate texts, many of the sentences were problematic. New sentences were therefore created, based on the data derived from the source corpus. The authenticity of a corpus is arguably its main attraction, but because corpora for the target population are not readily available, pedagogical criteria (appropriateness and usability) took priority. Firstly, it was essential to provide examples that could be used by teachers and materials writers, and 'fair use' copyright issues are somewhat unclear when applied to corpora. Although the new sentences were informed by the source corpus, i.e. in providing data on grammatical patterns, turning these into new sentences means the resource can receive wider distribution, and be used particularly by materials writers. Secondly, although many of the shorter sentences were appropriate and not unique (*We've won! She's married? I've decided.*), some sentences contained names or cultural allusions likely to be unknown by or irrelevant to the target population, or were difficult to understand without a context (see the examples in Table 5). Additionally, even though the database was level-appropriate, other sentences contained low-frequency words not necessarily

useful for this target population such as *The person who uses the heroin?* and *She sneers at people who are poor*. Thirdly, many sentences from the source corpus did not correspond to the sentence lengths defined for the three-level distinctions of beginner, intermediate and advanced. Fourthly, because some sentences were created for younger readers (e.g. from textbooks for US grades 1 and 2), age-related interests also needed to be considered. And finally, to modernize the corpus, sentences referring to current technology (mobile phones, websites, apps), contemporary companies (Sony, Nintendo, Apple), ideas (social media, video games, environmental issues) and popular culture recognized in Japan (*Harry Potter*, *Anne of Green Gables*) were also included. For all these reasons, a corpus of specially-written sentences was created.

Table 5: Examples of problematic sentences from the source corpus

	Examples of problematic sentences	Source
1	And there's a horse called Smoke.	gradedreader - Dead Man's Island
2	A broken neck, the doctor says.	gradedreader - Logan's Choice
3	But she never really forgot the speckled band.	gradedreader - Sherlock Homes
4	Peter came out from behind six broken TVs.	gradedreader - Jumanji
5	The wall opened, and Edwards saw a lot of coloured lights.	gradedreader - Men in Black
6	Hannah looked at Beth and called Dr . Bangs.	gradedreader - Little Women
7	Here in the United States ... in Washington?	gradedreader - Dante's Peak
8	The Rovers and United matches are always two-two or one-one.	gradedreader - Six Sketches

Three methods were used to produce SCoRE sentences. Some shorter sentences extracted from the source corpus were included because of their high frequency, which was verified in a general corpus. For example, in COCA, *I've seen worse* appears 21 times; * *game has started* appears five times. Often these were revised slightly, as in *Have you seen this website?* instead of *Have you seen Roz?*. Longer sentences (such as those shown in Table 5) were extracted from the source corpus and their patterns were used as a guide by a native English-speaking researcher for creating new sentences. For example, the first sentence in Table 5, *And there's a horse called Smoke*, might be used to frame the sentence *My horse is called Midnight* or *A young horse is called a pony* using the verb *called.* All sentences for each grammatical feature followed sentence length and word familiarity guidelines as outlined above. These were created by the native-English-speaking researcher, who has more than 25 years of experience as an L2 teacher, and were then verified by five other researchers. The resulting sentences excluded allusions to non-contemporary story lines or characters that may have appeared in the original sentences, such as the

reference to the Baudelaires or Count Olaf that occur in Table 6; similarly there are no low-frequency words and phrases that would be unfamiliar and thus perhaps not useful for low-proficiency learners (e.g. *I did find a man to mate*). Both Tables 6 and 7 show the basic pattern *I wish I could tell (someone)*; the sentences in Table 6 were extracted from the source corpus, and the sentences in Table 7 were created for SCoRE. These are not paired and there is no direct correlation; they are shown only for comparison.

Table 6: Examples of source corpus sentences extracted for the intermediate level

- I wish I could tell Lilly about Josh Richter talking to me.
- I wish I could tell them what I know, as they walked across the courtyard, raising small clouds of dust with every step.
- I wish you were nearby so I could tell you that I did find a man to mate.
- I wish I could tell you that the Baudelaires' first impressions of Count Olaf and his house were incorrect, as first impressions so often are.
- I wish I could tell you for sure, Jondalar, but I don't know.

Table 7: Examples of SCoRE database sentences created for the intermediate level

- I wish I could tell you how it happened.
- I wish I could tell you, but I just don't know.
- I wish I could tell you who was responsible.
- I wish I could tell you because then you would stop.
- I wish I could tell you how happy I am.

3.6. Translation

Each English example sentence is accompanied by a Japanese translation. To create these, machine translation software was used first, and then each translation was manually corrected separately by five Japanese native-speaker researchers. This translation step also served as a way to verify the English sentences because colloquial forms or obscure cultural references which were difficult to translate were identified and rejected. In these cases, the English sentence was revised or rewritten. This occurred in fewer than 2% of the example sentences. A small sampling of relative sentences using *whom* is shown in Table 8 with their translations. Although *whom* is used less and less frequently in American English (as a COCA search will show), it nevertheless remains on TOEIC tests and other proficiency assessments so was included as useful to the target population.

Table 8: Examples of relative sentences using whom *with Japanese translations*

English Sentences	Japanese Translations
Beginner/Remedial Level	
He is the man (whom) I love.	彼は私が愛する男性です。

She is the woman (whom) I married.	彼女は私が結婚した女性です。
He is the son (whom) I raised.	彼は私が育てた息子です。
She is the person (whom) I trust.	彼女は私が信頼している人です。
She is the person (whom) I respect.	彼女は私が尊敬する人です。
Intermediate Level	
These are the people (whom) I call my family.	こちらは私が家族と呼んでいる人たちです。
These are all the students (whom) I invited to my house.	こちらはすべて私の家に招待した生徒たちです。
These candidates were the ones (whom) I voted for.	これらの候補者は私が投票した人たちでした。
Here is a list of the friends (whom) I will travel with.	ここに私が一緒に旅行する友達のリストがあります。
Tom Cruise is an actor (whom) many fans enjoy watching.	トム・クルーズは多くのファンが楽しんでいる俳優です。
Advanced Level	
These are the candidates (whom) I supported in the last election.	これらの方々は前回の選挙で私が支持した候補者です。
Curie is one of many scientists (whom) the students will research this term.	キュリーは学生たちが今学期調査する科学者の一人です。
They are the engineers (whom) our company hired to repair the damage.	彼らはわが社が故障を直すために雇った技術者たちです。
The politicians (whom) I saw on television were arrested for taking bribes.	私がテレビで見た政治家たちは収賄で逮捕された。
Ben Howard is a wonderful new musician (whom) I had never heard of until recently.	ベン・ハワードは最近知った素晴らしい新人音楽家です。

Currently, the prototype GPPS has a small SCoRE database consisting of approximately 15,000 copyright-free sentences (25 grammatical categories x 10 search words x three levels x 10 sentences x the Japanese translation).

4. Pedagogical applications: Using SCoRE and the GPPS

One of the difficulties in teaching grammar using DDL for low-level EFL students in Japan has been a lack of level-appropriate example sentences. Using the GPPS and SCoRE, teachers and materials writers can find numerous, easily understood example sentences for students by simply selecting the targeted grammatical patterns. This would be a useful resource for language presentations in lessons, classroom or homework material, or quizzes. One application currently being investigated is to have students observe a KWIC presentation in a parallel concordancer such as AntPConc (Anthony 2013) to discover and form hypotheses about the language, and then use the GPPS to confirm and reinforce the grammatical rule in complete sentences. In addition, researchers may find the GPPS useful for comparing language patterns in English and Japanese. Once the GPPS is released, future studies will focus on developing classroom applications.

When creating DDL-based worksheets or materials for students using concordancers such as ParaConc or AntPConc, some grammatical patterns

lend themselves easily to concordance searches and a KWIC presentation. For example, a teacher could create a worksheet with instructions guiding students to search for * *books*, and students would easily be able to see various articles or determiners such as *the books*, *her books*, or *many books* in the resulting concordance lines. However, some grammatical features do not lend themselves to these kinds of simple KWIC searches. The relative clause (or 'contact clause', as it is known in Japanese English textbooks) is difficult for Japanese learners to understand because sometimes the relative word can be omitted (e.g. *the people (whom) we met last night were very nice*) and sometimes it cannot (e.g. *the woman* ***who*** *lives next door is a doctor*). It is difficult for teachers who are not specialist corpus users to find KWIC concordance patterns to show this kind of example. Because this specific grammatical feature has been identified and targeted as important to low-proficiency learners in Japan, sentences were specially created for it in SCoRE. Having these kinds of examples is one of the advantages of the GPPS.

A multilingual translation system is planned in the future so that the GPPS system will be available not only for Japanese EFL teachers and students, but also to English learners from other language backgrounds. This GPPS with SCoRE will be released as freeware on the DDL Open Platform with three additional corpus tools included (Chujo et al. 2013): WebParaNews, which is a web-based parallel concordancer that allows users to check word and phrase usage in an English and Japanese news corpus; AntPConc, which is a downloadable simple multilingual concordancer which works with corpora created by the users themselves; and LWP for ParaNews, which is a freeware lexical profiling program that allows users to check colligation/collocation usage in an English and Japanese news corpus. All four corpus tools (including the GPPS) are for bilingual or multilingual use. Teachers and students can investigate and observe the usage of words and phrases by search terms or by grammar patterns in English or Japanese, and can use more than one tool to observe a pattern.

5. Limitations of SCoRE and the GPPS

One of the most challenging aspects of this project has been the creation of example sentences. The aim was to create sentences that are interesting and easily understood while close to authentic sources and reflecting authentic patterns. Often language cannot be separated from culture, and this became evident when the translators were unable to understand some of the native speaker's sentences, for example *I wish I had a nickel for every time [something happened]*, or *it was no place for tourists after dark.* As educators, we are reminded that culture is very much a part of language learning. The method of creating sentences relies not only on empirical measures such as sentence length and word familiarity, but also on an intuitive understanding of sentences likely to be understood by low-

proficiency L2 learners. The three team leaders involved in creating, verifying and translating the sentences each have more than 25 years of experience as classroom teachers, and this type of semi-authentic text is meant as a balance between the more difficult real-world concordance data found in existing corpora and pedagogically-structured textbook grammar presentations.

Another limitation of this project lies in the use of US reading grade and word familiarity levels, which are based on data from the 1970s and 1980s. No other comparable data has been found for more recent periods; in fact, the shift in demographics have radically changed as ESL speakers have immigrated to the US, so contemporary, reliable data for reading norms may be difficult to assess. In addition, the choice of grammatical categories may be criticized on the grounds that they do not always correspond to high-frequency items in a native-speaker corpus (cf. the example of *whom*, discussed above); however, they do reflect patterns most needed by remedial students in Japan or a general audience of beginner-level EFL learners.

Finally, the creation of a corpus, as noted by Minn et al. (2005), is both time- and labor-intensive, and because of this, the GPPS is currently limited in the number of sentences available, but it will be continually updated. Once the GPPS is opened to the public on the DDL Open Platform, grammar items and sentences can be added by expert users – EFL teachers around the world will be able to contribute based on their own needs and demands.

6. Conclusion

The Japan Times recently reported that the prime minister plans to invest in improving English language skills in Japan, and that from 2015, applicants for government jobs will have to submit their TOEFL test results (Hongo 2013). In a similar vein, the *Jiji Press* (17 March, 2013) reported that the TOEFL may be used in National Public Service exams. If the use of DDL is to be successful in L2 university classes as a means to improve language proficiency, there must be appropriate needs-driven corpora and corpus-based classroom-ready material for low-proficiency students. The project outlined here aims to address this with the creation of the GPPS and SCoRE. The grammatical structures included in the material are available for beginner, intermediate and advanced learners. Because the example sentences are based on graded texts approximately equivalent to US elementary school grades, and are written for different levels of proficiency, the basic vocabulary and sentence structures represented will allow students to focus on the particular grammatical patterns in question rather than high-level or obscure vocabulary, or complex or unrelated patterns of less normal usage.

Future tasks for the project will be to add more grammatical patterns, continue to create copyright-free sentences, add a read-aloud feature and a quiz-type question-creating function, and investigate and report classroom applications. The website will be made public as more data becomes available. It is hoped that this browsing system will bridge the gap between 'textbook language' and real communication in a way that also promotes the use of corpora in the remedial or lower-level language classroom as it provides multiple affordances to learners, teachers and materials writers.

Acknowledgements

Part of this research was funded by a Grant-in-aid for Scientific Research (21320107; 25284108) from the Japan Society for the Promotion of Science and the Ministry of Education, Science, Sports and Culture.

References

Allan, R. 2009. Can a graded reader corpus provide 'authentic' input? *ELT Journal* 63(1): 23–32.

Anthony, L. 2013. *AntPConc*, version 1.0.2. Tokyo: Waseda University. <http://www.antlab.sci.waseda.ac.jp> (1 June, 2014).

Barlow, M. 2004. *ParaConc*. Houston TX: Athelstan. <http://www.athel.com/mono.html> (1 June, 2014).

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. 1999. *Longman Grammar of Spoken and Written English*. Harlow: Pearson Education.

Braun, S. 2005. From pedagogically relevant corpora to authentic language learning contents. *ReCALL* 17(1): 47–64.

Braun, S. 2007. Integrating corpus work into secondary education: From data-driven learning to needs-driven corpora. *ReCALL* 19(3): 307–328.

Breyer, Y. 2009. Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning* 22(2): 153–172.

British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Oxford University Computing Services on behalf of the BNC Consortium. <http://www.natcorp.ox.ac.uk/> (1 June, 2014).

Chujo, K., Akasegawa, S., Nishigaki, C., Yokota, K. & Hasegawa, S. 2012. LagoWordProfiler ni yoru Eigo graded reader corpus no collocation/colligation hindo bunseki [LagoWordProfiler frequency analysis of collocations and colligations in an English graded reader corpus]. *Journal of the College of Industrial Technology, Nihon University* 45: 1–17.

Chujo, K., Anthony, L., Akasegawa, S., Nishigaki, C., Mizumoto, A. & Utiyama, M. 2013. Kyouiku riyou kanouna parallel corpus kensaku platform no kouchiku ni mukete [Toward building a parallel corpus

concordance platform for English language education]. Paper presented at the *39th Japan Association of English Corpus Studies (JAECS)*. Sendai: Tohoku University, 4 Oct. <http://english.chs.nihon-u.ac.jp/jaecs/Archive/CONF/CONF_39.pdf> (1 June, 2014)

Chujo, K., Anthony, L., Oghigian, K. & Yokota, K. 2013. Teaching remedial grammar through data-driven learning using AntPConc. *Taiwan International ESP Journal* 5(2): 65–90.

Chujo, K., Nishigaki, C., Yamaho, M. & Amano, K. 2011. Eigo shokyuusha muke corpus data tositeno kyoukasho text no tekisei ni kansuru kenkyuu [Identifying the suitability of textbook English for beginner-level corpus data]. *Journal of the College of Industrial Technology, Nihon University* 44: 33–46.

Chujo, K., Nishigaki, C., Yamaho, M. & Ochiai, T. 2012. Beikoku reading kyoukasho to Eigo graded readers no Eigo shokyuusha muke corpus data tositeno tekisei ni kansuru kenkyuu [Identifying the suitability of American reading textbooks and English graded readers for beginner-level corpus data]. *Journal of the College of Industrial Technology, Nihon University* 45: 29–42.

Chujo, K., Oghigian, K. & Akasegawa, S. In press. Developing a level-appropriate, grammatically-categorized browsing system of EFL example sentences for teachers and students. In *TaLC10: Proceedings of the 10th International Conference on Teaching and Language Corpora*, A. Leńko-Szymańska (ed.). Warsaw: Institute of Applied Linguistics, University of Warsaw.

Chujo, K., Utiyama, M. & Nishigaki, C. 2007. Towards building a usable corpus collection for the ELT classroom. In *Corpora in the Foreign Language Classroom*, E. Hidalgo, L. Quereda, & J. Santana (eds), 47–69. Amsterdam: Rodopi.

Chujo, K., Yokota, K., Hasegawa, S. & Nishigaki, C. 2012. Remedial gakushuusha no Eigo shuujukudo to Eigo bunpou jyukutatsudo chousa [Identifying the general English proficiency and distinct grammar proficiency of remedial learners]. *Journal of the College of Industrial Technology, Nihon University* 45: 43–54.

Dale, E. & O'Rourke, J. 1981. *The Living Word Vocabulary*. Chicago: World Book-Childcraft International.

Davies, M. 2008-. *The Corpus of Contemporary American English: 450 million words, 1990-present*. <http://corpus.byu.edu/coca/> (1 June, 2014).

Educational Testing Service (ETS). 2012. *The TOEIC Test: Report of Test Takers Worldwide 2012*. <http://www.ets.org/s/toeic/pdf/2012_ww_data_report_unlweb.pdf> (4 June, 2014).

Educational Testing Service (ETS). 2014. *Test and Score Data Summary for TOEFL iBT® Tests*. <http://www.ets.org/s/toefl/pdf/94227_unlweb.pdf> (4 June, 2014).

Furukawa, A. 2007. *Yomiyasusa Level: A Reading Level for Japanese Students*. <http://www.seg.co.jp/sss/word_count/YL-20070621.html> (accessed 1 June, 2014).

Gavioli, L. & Aston, G. 2001. Enriching reality: Language corpora in language pedagogy. *ELT Journal* 55(3): 238–246.

Harris, A. J. & Jacobson, M. D. 1972. *Basic Elementary Reading Vocabularies.* New York: Macmillan.

Hongo, J. 2013. Abe wants TOEFL to be the key exam. *The Japan Times*, 25 March. <http://www.japantimes.co.jp/news/2013/03/25/national/abe-wants-toefl-to-be-key-exam> (1 June, 2014).

Huang, L.-S. 2008. Using guided, corpus-aided discovery to generate active learning. *English Teaching Forum* 46(4): 20–27.

JET Programme. 2010. *The Japan Exchange and Teaching Programme*. <http://www.jetprogramme.org> (1 June, 2014).

Jiji Press. 2013. Japan may introduce TOEFL as part of national public service exams. *NewsOnJapan.com*, 17 March. <http://newsonjapan.com/html/newsdesk/article/101506.php> (1 June, 2014).

Laufer, B. 1992. How much lexis is necessary for reading comprehension? In *Vocabulary and Applied Linguistics*, H. Béjoint & P. Arnaud (eds), 126–132. Basingstoke: Macmillan.

Michigan Corpus of Academic Spoken English. 2007. <http://quod.lib.umich.edu/m/micase/> (1 June 2014).

Micro Power and Light Co. 2003. *Readability Calculations*. Dallas TX. <http://www.micropowerandlight.com/rd.html> (1 June, 2014).

Minn, D., Sano, H., Ino, M. & Nakamura, T. 2005. Using the BNC to create and develop educational materials and a website for learners of English. *ICAME Journal* 29: 99–113.

Murphy, R. & Smalzer, R. 2009. *Grammar in Use: Intermediate*. Cambridge: Cambridge University Press.

Murphy, R. & Smalzer, R. 2011. *Basic Grammar in Use*. Cambridge: Cambridge University Press.

Ono, H., Muraki, E., Hayashi, N., Sugimori, N., Nozaki, H., Nishimori, T. Baba, M., Tanaka, K. Kuniyoshi, T. & Sakai, S. 2005. Nihon no daigakusei no kiso gakuryo ku kouzou to remedial kyouiku [A development of a placement test and e-learning system for Japanese university students: Research on support improving academic ability based on IT]. *NIME Research Report 6-2005*: 1–147.

Shirahata, T. 2008. Shougakusei to chuugakusei no eigo jyukutatsu do chousa [The investigation into the proficiency levels of English by both elementary school children and junior high school students]. In *Daini Gengo Shuutoku Kennkyuu wo Kiban to Suru Shou, Chuu, Kou,Dai no Renkei wo Hakaru Eigo Kyouiku no Sendou-teki Kennkyuu [Leading Research to Coordinate Primary, Secondary, and Tertiary Level English Education Based on Second Language Research]*, I. Koike (ed.), 166–190. Tokyo: Kaken Kenkyuu Seika Houkokusho.

Tanaka, S., Kobayashi, Y., Tokumi, M. & Asao, K. 2008. Gakkou Eibunnpou corpus no kouchiku no kokoromi [Development of a school grammar corpus of English]. Paper presented at *The 22nd Annual Conference of the Japanese Society for Artificial Intelligence*, Asahikawa, 12 June.

Uchibori, A. & Chujo, K. 2005. Daigaku shokyu level gakushuusha no eigo communiacation nouryoku koujou ni muketa CALL bunpouryoku youseiyou software no kaihatsu [The development of CALL material for grammar to improve communicative proficiency of beginner-level college students]. *Journal of the College of Industrial Technology, Nihon University* 38: 39–49.

Uchibori, A., Chujo, K. & Hasegawa, S. 2006. Towards better grammar instruction: Bridging the gap between high school textbooks and TOEIC. *Asian EFL Journal* 8(2): 228–253.

Utiyama, M. & Takahashi, M. 2003. English-Japanese translation alignment data. <http://www2.nict.go.jp/univ-com/multi_trans/member/mutiyama/align/index.html> (1 June, 2014).

Yoshida, K. 2008. TEFL in Japan: An overview. *Proceedings of the 15th World Congress of AILA*, 1–8. Essen: University Duisburg-Essen, 25 Aug. <http://pweb.cc.sophia.ac.jp/1974ky/TEFLinJapan.pdf> (1 June, 2014).