School of Electronic Engineering and Computer Science

MSc Computing and Information Systems

Project Report 2018

Vocabulary extraction from text input with language-dependent difficulty scoring for ESL contexts

Rowena Rhiannon Jones

August 2018





# ACKNOWLEDGEMENTS

I would like to express my thanks to Dr Tony Stockman, who was the first to hear of my idea for this project and confirm its interest and validity.

My thanks also go to my supervisor Dr Matthew Purver, who took over the reins from Dr Stockman early on and provided useful advice, guidance and support.

Finally I express my gratitude to the many teachers of English as a Foreign Language in Paris and around the world who made this project possible by participating in my research. Thus my thanks go to: Helen Britt, Matthew Fletcher, Kimberley Gaugler, Sophie Lawson, Michaela McElhill, Margaret Stewart and Jane White.  
  
  
  
  
  
  
  
  
  
  
  
 i

# ABSTRACT

We create and evaluate a Python tool to extract and difficulty-score vocabulary items (both single word tokens and multi-word expressions) from input reading texts for use in English as a Second Language contexts. In addition to using the common measure of frequency within a reference corpus to assess a vocabulary item’s difficulty, we also consider the similarity between the English item and its translation to a given learner’s native language. Our tool extracts a number of such scoring features from vocabulary items, and uses linear regression combined with domain expertise to select appropriate features and coefficients for computing scores that best predict the difficulty scores obtained from teachers during primary research. Evaluation of the resulting tool shows a very favourable performance in terms of multi-word expression extraction against comparable tools. The utility of a frequency feature in predicting vocabulary difficulty is confirmed, with suggestion that our feature which incorporates native language into the overall difficulty level is also useful.

ii

# TABLE OF CONTENTS

ACKNOWLEDGEMENTS....................................................................................................... i

ABSTRACT ........................................................................................................................... ii

TABLE OF CONTENTS ....................................................................................................... iii

CHAPTERS

I: INTRODUCTION......................................................................................................1  
II: BACKGROUND: .................................................................................................... 2

A. Challenges in Teaching English as a Foreign Language ………...………....3

B. Determining Vocabulary Difficulty – Word Frequency, The Case for Native Language Consideration ………………………………………………………..... 5

C. Vocabulary Extraction From Text Input: The Challenge of Multiword Expressions ……………………………………………………………………..…. 7

D. Review of Existing Text & Vocabulary Tools for ESL Teachers and Learners ……………………………………………………..…….……..……..…. 9

III: METHODOLOGY……………………………………………..…….………………… 11

A. Establishing Vocabulary Extraction and Scoring Targets: Primary Research ……………………………………………………………..………..……....12

B. Python & Natural Language Toolkit (NLTK) …….……………………..……15

C. The British National Corpus .…………………….……………………..…… 17

D. English-to-French Wiktionary .…………………….…………………...……. 18

E. Evaluation Methods.…………………….……………….….………..………. 19

IV: IMPLEMENTATION .…………………….………………….…………..…...………. 20

A. Vocabulary Item Extraction ………………………………………………….. 21

B. Scoring Features: i. Vocabulary Item Frequency ……..……………...…… 24

B. Scoring Features: ii. POS-Tagged Frequency ……..…………………..….. 26

B. Scoring Features: iii. L1 Similarity Feature ……..……………………...….. 28

B. Scoring Features: iv. Part of Speech Tag Feature ……….……………..… 31

B. Scoring Features: v. Vocabulary Item Length .…………….………………. 32

C. Feature Contribution: Assigning an Overall Score ……………..…..…….. 33

V: EVALUATION ……………………………………………………………..………….. 35

A. Vocabulary Item Extraction ……………………………………………...…... 36

B. Difficulty Scores ……………………………..…………………………...…… 39

C. Vocabulary Item Rankings ……………………………………...…………… 41

VI: APPLICATIONS & EXTENSIONS .…………….………………………………..…………. 43

VII: CONCLUSION .…………………….………………………………………..…..…………. 45

REFERENCES .…………….……………………………………………………..…..…………. 46

APPENDIX A: SAMPLE TEXTS …………………………………...………………..…..………50

APPENDIX B: TEACHER RESEARCH DOCUMENT …………………………..………..….. 51

APPENDIX C: TEACHER RESEARCH RESULTS & RESULTING DIFFICULTY SCORES

…………………………..……………………………………………………….……..….. 52

APPENDIX D: POS TAG MAPPING …………………………………………….…..…….….. 54

APPENDIX E: GUIDE TO SUPPORTING MATERIAL ………………….………..…...…..… 55

# I: INTRODUCTION

Hundreds of thousands of English as a Second Language (ESL) teachers working around the world are increasingly required to personalise their lessons and incorporate authentic texts which have been selected with reference to students’ needs and interests (Hedge, 2000) (Kerr, 2016). When planning lessons based around these texts, teachers must predict the ‘difficult’ vocabulary items within that will be unknown to the learner and require teaching, exercises and practice (British Council, 2011).

A tool which can receive a teacher or learner’s chosen text as input, and output a difficulty-scored list of vocabulary items from within the text may be used by both new and experienced teachers to speed up their lesson planning, by learners who wish to study independently and as the basis for many other learning tools and resources. However, it is crucial to include multiword expressions in the output of vocabulary items, an NLP task which is significantly more challenging than the extraction of single word items.(Sag et al, 2003).

Notably, many ESL teachers based in foreign countries work in contexts where all their students share a common native language. Yet little work exists to evaluate how etymological similarities between English and a given student’s native language may affect which vocabulary items are more ‘difficult’ for that learner. A French learner will not struggle to understand the English word *somnambulist* (a formal term for a sleepwalker) as it is highly similar to the French translation *somnambule,* a common French word. We thus explore similarity between native language and English as a measure of difficulty in our tool.

# II: BACKGROUND:

### **Chapter Summary**

**A. Challenges in Teaching English as a Foreign Language:** TEFL is a booming sector. Increasing call for ESL teachers to use authentic, bespoke texts when teaching reading and vocabulary suggests the potential value of a tool to aid lesson planning by processing such texts and predicting difficult vocabulary items within.

**B.** Determining Vocabulary Difficulty – Word Frequency, The Case for Native Language Consideration**:** NLP studies have primarily been concerned with predicting the difficulty level of an overall text, and not with detailed exploration of how individual words and vocabulary items may be assigned difficulty scores. The frequency of occurrence of a given word within a reference corpus is the typical means by word difficulty is assessed in the field. In the field of ESL, factors such as word length and phonology have also been considered as contributors to word difficulty. We make the case for a further contributing factor: the similarity of a given word to its translation in a learner’s native language.

C. Vocabulary Extraction From Text Input: The Challenge of Multiword Expressions: While identifying and extracting single-word vocabulary items from a given input text is made relatively simple with today’s NLP tools, doing the same for multiword expressions (MWEs) is significantly more challenging. We briefly consider what constitutes a MWE, why they are important in the fields of both ESL and NLP, and how NLP deals with these items.

**D. Review of Existing Text & Vocabulary Tools for ESL Teachers and Learners**: Having identified a need to incorporate personally chosen, authentic texts into ESL teaching, and the interest of extracting and difficulty-scoring vocabulary items from within these texts, we consider a number of relevant tools and resources already in existence and identify their strengths and limitations.

## A. Challenges in Teaching English as a Foreign Language

**The TEFL Sector**In an increasingly globalised world, Teaching English as a Foreign Language (TEFL) is a booming business sector. There are 1.75 billion learners of English worldwide, set to rise to 2 billion by 2020 (The British Council, 2013). While there are no firm figures on the number of ESL teachers, it is estimated that ‘250,000 native English speakers work as English teachers abroad in more than 40,000 schools and language institutes around the world” and that TEFL is a $63bn-a-year industry (internationalteflacademy.com, nd).

**The importance of texts to teach vocabulary and reading skills in TEFL**

The practice of bringing texts to the classroom to teach reading skills and vocabulary (both essential components of language learning) is commonplace in TEFL. Teachers working in mainstream education or preparing students for exams often follow a set curriculum with specific textbooks and oft re-used texts. Yet many ESL teachers work in less prescriptive contexts: from giving private one-on-one lessons, to teaching adults in-company for business purposes or simply in language schools without specific materials or curriculums to follow. These teachers must design their own lessons and source their own materials.

**Authentic materials and the communicative approach**

Since the 1970s, TEFL has been focused on a communicative approach, whereby the ‘PPP’ methodology – *Present* [the target language], *Practice* [using said language in controlled contexts] and *Produce* [the language in freer contexts] – is often employed (Ellis, 2003). In tandem with the communicative approach has come “pressure to use authentic materials…materials which have not been designed especially for language learners and which therefore do not have contrived or simplified language” (Hedge, 2000, p67). In fact, the presentation of vocabulary in ESL classes is therefore often achieved via the use of authentic materials, such as news articles, literary extracts or business texts.

**Personalised Learning**

Choosing these texts in relation to the learners’ interests and needs also plays into the trend for personalised learning, described as being ‘at the heart of educational programmes around the world’ (Kerr, 2016, p2) and ‘one of the great education challenges of the 21st century” (Trilling & Fadel, 2009, p33). Teachers are increasingly expected to tailor their lesson content – including reading texts and vocabulary sets - to individual learners. This can mean the student themselves select the texts they wish to study and the types of vocabulary they wish to learn, particularly in the context of English for Specific Purposes.

**English for Specific Purposes (ESP)**

Hand in hand with personalised learning comes English for Specific Purposes. This field emerged in the 1970s and is “designed to meet specific needs of the learner”, “centred on the language appropriate to [their] activities” (Dudley-Evans, 1998, p6). A common variety of ESP is Business English, itself having hundreds of sub-categories: ESL students from the finance industry will have different needs to those from the medical sector in terms of the texts they must be able to read and the specialist vocabulary needed for their jobs. ESL teachers may thus be expected to incorporate a diverse range of texts into their teaching.

**Lesson Planning**

In order to prepare a text-based lesson, particularly with an unfamiliar text chosen for a specific student, it is important that the teacher can predict which vocabulary items from the text will be already familiar to the learner and which will be new or difficult. This enables the preparation of exercises and practice activities focused on the difficult vocabulary. (The British Council, 2011).

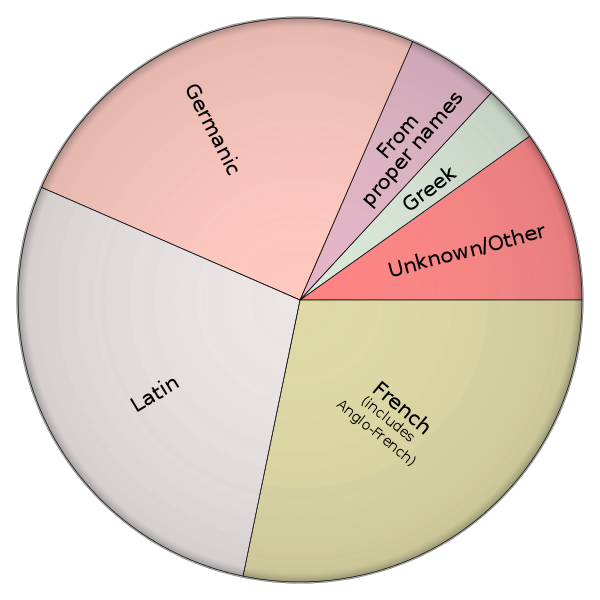
## B. **Determining Vocabulary Difficulty – Word Frequency, The Case for Native Language Consideration**

In the field of NLP, work has largely focused on predicting the difficulty level of an overall text rather than individual words / phrases within it. Word frequency in a reference corpus is generally used to measure individual word difficulty, while other features relate to grammar and syntax, sentence length etc. Curto, Mamede & Baptista (2015) extract other features, including parts-of-speech tags, words, averages and frequencies, from reading texts and use an annotated corpus of texts to train a classifier to assign discrete readability levels. Chen and Meurers (2016) use a rich representation of word frequencies to characterise the difficulty level of a text, considering frequency scores within clusters of words and in terms of standard deviation, rather than simply averaging the frequency scores of all a text’s words. Heilman et al (2008) and Aluisio et al (2010) use both lexical and grammatical features to measure the overall readability of a text, the latter in support of simplifying a complex text for ‘poor literacy readers’. In all of these studies, word frequency within a reference corpus has been a key measure of individual word difficulty within a reading text.

In the fields of ESL and educational psychology, word frequency also dominates as a ‘gold standard’ measure for determining vocabulary difficulty (West 1953), (Yoon 2012), (Cambridge University Press English Profile, 2018) (Breland 1996). We note that other features such as word length, word phonology and pronounceability have been explored in these fields (Spencer 2002), (Rodgers 2009).

Various papers explore the impact of **L1** (a given ESL learner’s native language) on various aspects of learning: writing (Sanmuganathan, 2014), grammar (Kosterina, A 2007) and pronunciation and listening (Tremblay, A. 2016). However, we do not believe that L1 has yet been explored as a formalised measure for predicting vocabulary difficulty for a given learner. We briefly make the case for inclusion of this as a measure in our tool:

* **Teaching in Homogenous L1 Contexts:** While some of the world’s ESL teachers are based in the UK, US and other locations where learners come from many different countries with different L1s, thousands of others (perhaps a majority) are based in foreign countries where all their learners share a common L1.

  
**Figure 1: Origins of English**Murraytheb, [English Wikipedia](https://en.wikipedia.org/wiki/) (2007)

* **Foreign Language Influences On English:** English shares many of its roots with other languages (see Figure 1). Many French and Latin-rooted words in English will be highly familiar to learners of Latin-rooted L1s. Consider the similarity in the translations of these English words, all of which have origins in French (or Latin via French):

|  |  |  |  |
| --- | --- | --- | --- |
| **English** | **French** | **Italian** | **Spanish** |
| respond | répondre | rispondere | respondeo |
| necessary | nécessaire | necessario | necesario |
| government | gouvernement | governo | gobierno |
| ruminate | ruminer | ruminare | rumiar |
| resplendent | resplendissant | risplendente | resplandeciente |
| somnambulist | somnambule | sonnambulo | sonámbulo |
| ***Figure 2: English words with Latin / French etymology***  *Translation source: Google Translate, 2018* | | | |

Note that the words ‘ruminate’, ‘resplendent’ and ‘somnabulist’ would be expected to have increasingly low frequency in a reference corpus, not being common, day-to-day words. It would thus be suggested that these are ‘difficult’ words for all learners, whereas in fact learners with an L1 of any of the three languages above are not likely to find these words as difficult to learn or understand as the more frequently occurring but Germanic / Old English rooted ‘ponder’, or ‘mull’ (as synonyms of ruminate), ‘pretty’ or ‘handsome’ (as variations of resplendent) or ‘sleepwalker’ (common word for somnambulist).

Our hypothesis is thus that incorporating L1 into a measure of vocabulary item difficulty will be valid and useful where the learner’s L1 is a Latin, Germanic or French language, yet not for L1s which are outside this group, for example Mandarin Chinese, Turkish, Afrikaans.

## **C. Vocabulary Extraction From Text Input: The Challenge of Multiword Expressions**

What is a multiword expression (MWE)? Given that it is not always simple to define what constitutes a word (Baldwin and Kim, 2010), defining a MWE is more complicated still. Broadly, we can think of an MWE as 1) a group of two or more words *and* 2) having a meaning or usage as a unit that transcends their individual component words *and* 3) commonly co-occurring as a group.

Consider the difference between the verb *look* in its simple form, meaning to regard or see, and the way its meaning changes when it is considered as an MWE with the addition of different particle: ‘*look up to’* (somebody you admire), ‘*look into’* (a crime), ‘*look up’* (a word in the dictionary). Note that some of these constructions are decomposable / syntactically flexible (*‘****look*** *the word* ***up*** *in the dictionary’)* while others are not (we can **‘*look up to*** a person’but not ‘**look** a person **up to’**). Going beyond verb particle constructions (known as phrasal verbs in the field of ESL), MWEs also comprise fixed idiomatic expressions (‘in a nutshell’, ‘top dog’, ‘get the ball rolling’), proverbs (‘*don’t count your chickens before they’re hatched’, ‘you can lead a horse to water but you can’t make it drink’),* compounds (*‘dry run’, ‘nest egg’, ‘fossil fuel’*) and more (Constant et al, 2017).

MWEs have long been deemed ‘a pain in the neck for NLP’, causing problems of overgeneration (‘*coffee table*’ is a MWE, while ‘*tea table*’ is not), idiomaticity *(*the meaning of *‘don’t count your chickens’* has little to do with counting or chickens) and flexibility (in ‘*look into my eyes’ and ‘look into the murder’,* the MWE ‘*look* *into’* has two distinctly different meanings)(Sag et al, 2003). For ESL learners, many of the same problems apply: a student may perfectly understand the component words of an MWE in their own rights, yet fail to understand the MWE as a whole, or perhaps erroneously believe they have understood it while failing in fact to grasp its significance. The sheer number of MWEs can pose a daunting challenge to learners, and yet comprehension of MWEs is essential to the task of learning English as a second language (Wood, 2004), with idioms, phrasal verbs and fixed expressions included in virtually all ESL curriculums and resources.

It is thus clear that MWEs within a text must be identified and extracted along with single word vocabulary items, for a tool such as ours to be useful and valuable in ESL contexts. But how to achieve this? In a recent survey of MWE resources in the field of NLP, it was summarised that many different proposals for dealing with MWEs exist, from statistical analysis to comprehensive linguistic analysis of texts, however “most proposals still concentrate on the creation of MWE lexicons or the automatic recognition of MWEs” (Losnegaard, Sangatila et al, 2016). The use of an external reference lexicon (or dictionary) of MWEs enables one to cross reference a given group of words from a text with the lexicon to determine its validity or not as an MWE – however, given the fore mentioned problem of decomposability / syntactical flexibility of some MWEs, such a method is still far from simple.

## D. Review of Existing Text & Vocabulary Tools for ESL Teachers and Learners

**Authentic articles with accompanying exercises**

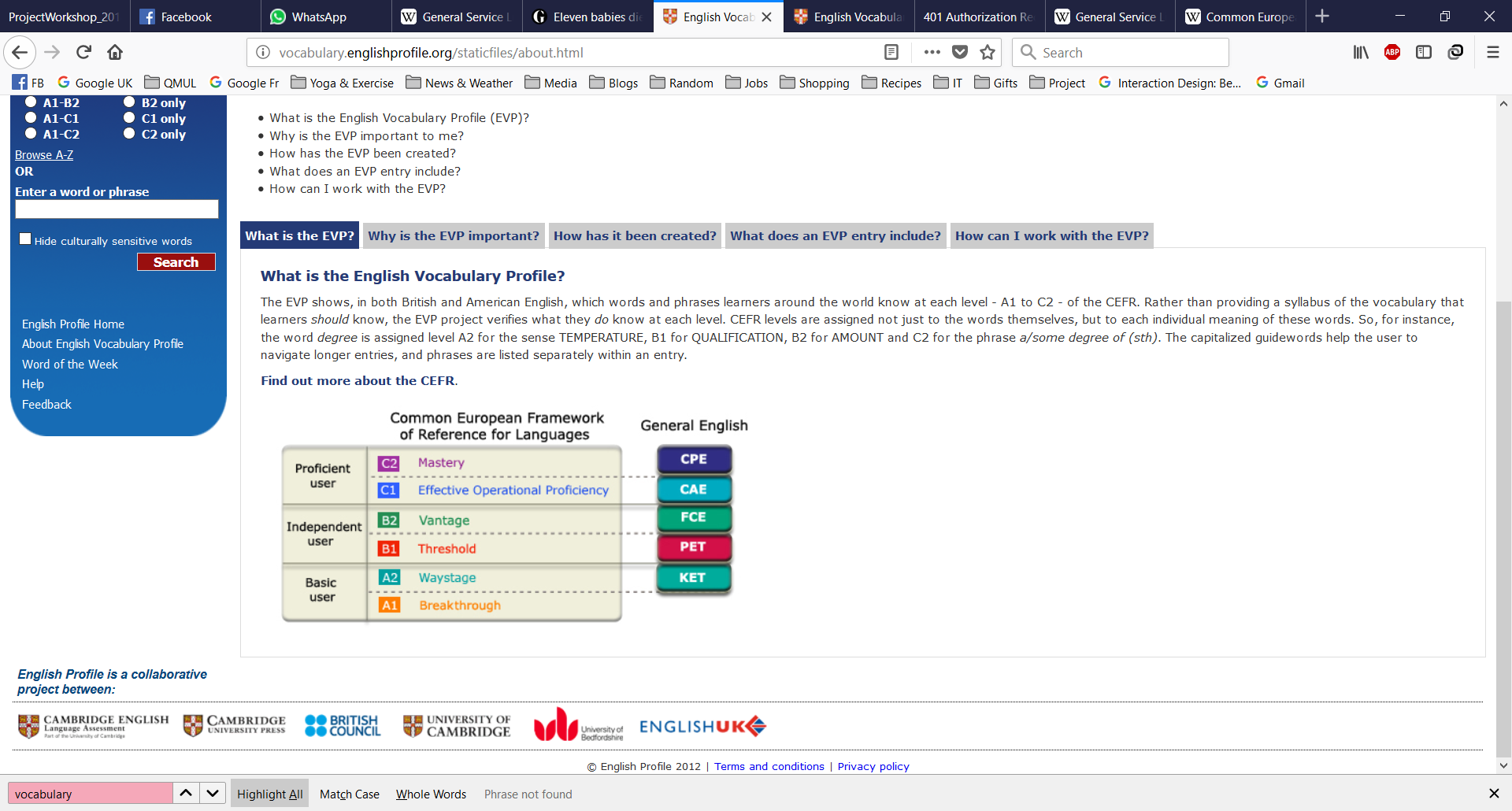
A handful of subscription-based websites (Market Leader Premier Lessons (nd), One Stop English 2000-2018) are the result of partnerships between ESL publishers and news outlets, and provide teachers with a frequently-updated bank of ‘real’ newspaper articles each accompanied with a series of vocabulary exercises. Teachers may search for articles relevant to particular business sectors or topics, and print them out with the accompanying exercises to use in lessons. However, the articles available from such websites represent only a fraction of the content from the partnered news outlet: finding a text on a specifically desired topic is not always possible. Teachers must also trust that the vocabulary selected for exercises is the most difficult / appropriate vocabulary for their learners.

**Text Input 🡪 Vocabulary Extraction Tools:**

In terms of tools which allow the user total control over text input, Vocabulary.com provides a ‘vocab grabber’ which allows students / teachers to paste a chosen input text and receive as output a list of vocabulary items extracted from said text (Vocabulary.com, 2018). These can then be studied via simple definition-matching exercises on the platform. The website Visual Thesaurus builds on this tool to provide a more graphical summary of the vocabulary items in the input text, displayed and grouped by size and colour with different filtering options - by frequency in the text, subject area etc (ThinkMap, 2018). Visual Thesaurus does provide a filter to sort the items by ‘how familiar the words are in written English overall’, though individual item scores are not assigned. We can also note that the VocabGrabber tool is limited in its ability to extract multiword expressions from input texts (see Chapter V: Evaluation, Part A).

**Vocabulary Scoring Tools:**

Cambridge University Press’s ‘English Vocabulary Profile’ web project assigns a discrete CEFR Level (see Figure 1), from A1 (‘easiest’) to C2 (‘hardest’) to a massive range of words, collocations and expressions (Cambridge University Press, 2015). This impressive tool from a highly authoritative source allows teachers and learners to browse vocabulary expected to be known at certain levels, search for a vocabulary item to find its level, and even input a raw text to receive individual vocabulary item levels. We can nonetheless identify two potential weaknesses with this project: firstly, it is based primarily on educational English contexts and uses word frequency within Cambridge exam papers and scripts as its primary difficulty categorisation tool. It’s utility for students studying business English is less clear. Secondly, much like with VocabGrabber, its Text Inspector tool’s ability to extract multiword expressions from texts is limited (see again Chapter V: Evaluation, Part A).



**Figure 3: Common European Framework of Reference for Languages Levels Chart** http://vocabulary.englishprofile.org/staticfiles/about.html

We also note that a commercial enterprise, Twinword, provides a paid-for API to evaluate the difficulty level of an individual word, or the overall difficulty level of a text. Much like English Vocabulary Profile, it uses word frequency and exam occurrence to assign difficulty scores to words, as well as ‘several other categories’ which are not specified. (Twinword, 2018). It is unclear to what extent it successfully deals with multiword expressions, since it assigns a difficulty level to an overall text without breaking down the difficulty of the vocabulary items within it.

# III: METHODOLOGY

### **Chapter Summary**

**A. Establishing Vocabulary Extraction and Scoring Targets: Primary Research:** A group of teachers working exclusively with French L1 students participated in primary research to identify the difficult vocabulary in six authentic sample texts. The aggregated research results are used to obtain normalised difficulty scores for vocabulary items within each sample text, thereby providing scoring targets for our study. Half of the scored texts will be used for the design and training of a tool which extracts vocabulary items and assigns difficulty scores, with the remaining half reserved for final testing and evaluation.

**B. Python & Natural Language Toolkit (NLTK):** We will use Python and the Natural Language Toolkit in the creation of this tool.

**C. The British National Corpus:** We use the BNC as our reference corpus for the computation of frequency-based difficulty scores for vocabulary items.

**D. English-to-French Wiktionary:** English-to-French Wiktionary is used for the computation of difficulty scores based on a vocabulary item’s similarity to the reader’s native language (for our purposes in this study, French).

**E**. Evaluation Methods**: Pearson’s Correlation Coefficient and the Mean Squared Error:** These calculations are used throughout the design and evaluation of our tool.

## A. Establishing Vocabulary Extraction and Scoring Targets: Primary Research

In order to design and test a new tool to extract vocabulary items from a text and assign them with difficulty scores, we must be able to check that the tool a) extracts all relevant vocabulary items (including multiword expressions) from the text, ready for scoring, and b) that the scores assigned to the vocabulary items by the program are relevant and useful in the real world of TEFL.

To this end, we incorporate domain expertise into our study. Six sample texts (emails, articles and report extracts each between 450 and 750 words in length, see Appendix A), with a survey document (see Appendix B) were given to a group of ESL teachers, working in France with exclusively French-L1 students. Each teacher was individually asked to identify the difficult vocabulary items in each sample text that they would expect to have to ‘teach’ to a typical intermediate or upwards level student.

**Limitations and Data Validity**

The research into word difficulty has been carried out with teachers and not students. The teachers participating in the study have an average of 10 and a median of 8 years of experiencing teaching ESL, and thus have a high level of expertise and experience in judging vocabulary likely to be difficult for their learners. By asking students themselves which vocabulary items in a text they think are ‘difficult’, we run the risk of students falsely believing that they understand certain items (particularly multiword expression and items with ambiguous meanings). For this reason, teachers, not students, have been preferred for establishing scoring benchmarks.

Similarly, teachers were not asked to score all the vocabulary within the text but rather just to identify that which they would judge difficult for intermediate learners. The justification for this is two-fold. Firstly, it is rare to use unaltered, ‘real world’ texts with learners below an intermediate level. These learners by definition still lack a critical mass of basic vocabulary and are generally unable to comprehend long real-world texts. Secondly, for teachers to score **all** vocabulary items in a text would be very time-consuming and be alien to the way that they work in the real world. By asking them to replicate the organic process of predicting the difficult words they will need to teach from a text, we hope to produce useful, meaningful results which are still effective in scoring words of pre-intermediate difficulty.

Finally, we acknowledge that our dataset of six sample texts is small. With no known publicly-available data for difficulty-scored vocabulary items using L1 as a measure, the collection of our own data was a necessary step. The data obtained from a group of expert teachers working with French L1 students, is high quality, but difficult to ‘scale up’ to more sample texts due to the time commitment required from research participants. We thus prefer a small amount of high quality data over potentially lower-quality data obtained by the potential crowdsourcing of learners or unknown, potentially inexperienced teachers, for the reasons described in the previous two paragraphs.

**Research Results**

Appendix C shows the aggregated research results from the six teachers. In using these results, we considered vocabulary items on a per-text basis, and assigned normalised difficulty scores as follows:

* Text items chosen by all teachers received a score of 1
* Text items chosen by none of the teachers received a score of 0
* Text Items chosen by some but not all teachers received a score between 0 and 1 based on how many teachers included the item in their ‘difficult’ list.

We can generally observe the following:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Text 1** | **Text 2** | **Text 3** |
| Total number of vocabulary items from each text scoring >0 | 46 | 49 | 40 |
| …which are multi-word expressions: | 20 (43%) | 17 (35%) | 5 (13%) |
| ***Figure 4: Teacher Research Results: Proportion of vocab items with score >0, which are MWEs.*** | | | |

* **MWEs:** The importance of MWEs inclusion as vocabulary items is evidenced by their significant inclusion as vocabulary items scoring >0 in our research (see Figure 3). With between 13 and 43% of the target vocabulary items being single word tokens for a given text, it is necessary to develop a method of identifying and extracting multiword expressions and putting them into the pool.

**⦁ L1 Contribution:** Teachers were invited to comment on their selection of vocabulary if they wished, in terms of their basis for selection and whether the native language of their students affected which items they thought would be difficult. The four teachers who engaged with these questions all had previous experience teaching in countries other than France. Three stated that they believed native language was an influencing factor in vocabulary difficulty and had taken this into account with their selection, with only one teacher saying they believed it was irrelevant. Two notable comments are as follows:

“*If choosing vocabulary for Chinese speaking students, rather than French speaking students, my choices would be broader. With no overlap/shared roots between Chinese and English, a teacher can’t assume that any vocabulary word will be familiar to a student.”*

*“I feel that different languages would find different vocab challenging. French speakers would probably know some of the words I [didn’t select] since they are cognates.”*

**Application of Research Results**The tool’s task will henceforth be defined as extracting all the vocabulary items identified as of interest (i.e. having a score >0) from our research participants. These ‘difficulty ‘scores will be used as the target scores for the program, which will incorporate L1 (French) Similarity as one of its scoring features. Half of the scored texts will be used during the overall creation and ongoing evaluation of the tool in terms of vocabulary item extraction, feature design and feature contribution.The remaining three texts will be reserved for the assessment and evaluation of the final tool.

## B. Python & Natural Language Toolkit (NLTK)

The tool will be created with Python and notably the Natural Language Toolkit (NLTK): a vast package of libraries and components designed for Python programmers working with human language data (NLTK, 2018). NLTK provides many useful tools for the purpose of text processing and feature extraction, which we briefly discuss below, in addition to other Python modules and packages used in our implementation.

**Natural Language Toolkit**

* **Tokenize:** NLTK’s tokenize module allows us to split an input string into a collection of tokens, or substrings. Tokenization can be at sentence level (one token = one sentence) or at word level (one token = one word). Tokenization will be essential to our task of extracting vocabulary items from raw text input.
* **Frequency Distribution:** When given a collection of tokens, the frequency distribution function tells us for each token (or ‘sample’), how many times it appears in the collection. This function will be highly useful in the design of our word frequency feature, when compiling word frequencies in reference corpora.
* **Corpus Reader:** The corpus reader package provides a reader interface for working with external corpora (collections of text). Notably, we use the BNCCorpusReader in our creation of a Frequency-scoring feature.
* **Part of Speech (POS) Taggers:** NLTK’s tag package provides automated tagging of tokens in a sentence with supplementary Part of Speech information. Tokens are thus identified as verbs, nouns, punctuation etc. POS tags enable us to improve our extraction of MWEs, more accurately translate vocabulary items, and can be considered in their own right as a scoring feature.
* **Stemmers & Lemmatizers:** Stemming a vocabulary item strips it of any prefixes or suffixes, reducing it to its stem, for example *misapprehending* becomes *apprehend.* NLTK’s Wordnet Lemmatizer is similarly used to find the base lemma of the word, but uses a sophisticated lexical database (rather than a more primitive string rule-based stripping method) to obtain a word’s lemma, for example *study* for *studies* (a stemmer may well return *studi).* NLTK’s Porter Stemmer and WordNet Lemmatizer are highly useful in many aspects of text translation and MWE extraction in our tool.
* **Bigrams & N-grams:** NLTK’s Bigrams module allows us to extract adjacent word pairs from tokenized strings, so that [‘His’, ‘name’, ‘is’, ‘Jack’] returns bigrams of (‘His’, ‘name’), (‘name’, ‘is’), (‘is’, ‘Jack’). The N-grams module allows us to specify the number of adjacent words to return in each tuple (so a bigram is in fact an n-gram where n=2). We will principally use the n-grams feature in MWE identification.

**Re:** The re (regex) module provides regular expression matching operations, permitting us to search for, match and substitute specified text patterns within an input text. This is particularly useful for dealing with punctuation, symbols, non-standard tokens (e.g. emoticons), and wrangling of raw text data.

**Levenshtein Module** (Pypi, 2014): This module provides built-in support for the Levenshtein Edit Distances calculation, which calculates the similarity between two strings. The edit distance is the minimum number of edit operations required to transform string one into string two, via character insertions, deletions or replacements. This measure will be the backbone of our feature which considers the similarity between an English word and its French translation, as it equates to the word’s difficulty.

**SciKitLearn Linear Regression Module** (Scikit-learn, 2017): We use this package to calculate appropriate co-efficients for our scoring features, by fitting a linear model to our feature score data, using our teacher-obtained difficulty scores as targets.

**Other Packages & Tools:** Numpy, Pandas and Matplotlib are also used within our tool for the processing, manipulating, storage and visualisation of data.

## C. The British National Corpus

(See also: Supporting Material: *Creating frequency distributions from the BNC corpus.ipynb)*

For the purpose of measuring vocabulary item frequency as it relates to difficulty, it will be necessary to use an external reference corpus to obtain frequency outcomes.

**Reference Corpus Selection**

NLTK includes a number of corpora as part of its standard installation, most notably the Brown corpus, the first ever million-word electronic corpus of English (Bird et al, 2014)**.** The Brown corpus includes text from 500 different sources, from news articles and government reports to romance novels. Producing a frequency distribution from this corpus gives based on single-word, all lower-case samples gives us 49,815 samples and 1,161,192 outcomes.

This corpus is problematic for our purposes in two respects: its age and its size. The corpus was created in 1961, and as such its texts are somewhat dated. Words relating to topics which have become commonplace only since this date, for example *renewable* (in the case of energy) or *internet* have deceptively low frequency counts. Secondly, while a million-word corpus may seem sizeable, it is in fact relatively sparse for NLP purposes.

Instead, we turn to the 2007 edition of the British National Corpus (BNC), a 100-million-word XML-based corpus which incorporates a wide range of both spoken and written language sources (The British National Corpus, 2007). The size, availability and modernity of this corpus make it highly suitable for our purposes, giving us 642,054 single word samples with 111,977,277 outcomes.

## D. English-to-French Wiktionary

(See also: Supporting Material: *Creating an Eng-to-French dictionary from Wiktionary data.ipynb)*

**Translation versus Etymology**While it may be simpler to look up the etymology of an English word (in terms of its roots being in French, Latin, Germanic etc) and use the result in assigning a difficulty score based on etymological commonality with learner’s native language, the approach would be hindered by the false cognates problem. The English word *actually* has Latin etymology, and seems similar to the French word *actuellement.* However, this is a false cognate: the correct translation of *actually* is *en fait*, with the French word *actuellement* meaning *currently.* The existence of many such false cognates precludes a simple etymology lookup. We prefer a more informed translation of the English vocabulary item, and a Levenshtein distance calculation. For this, we need a translation dictionary.

**English to French Wiktionary**

The English Wiktionary is ‘a collaborative project to produce a free-content multilingual dictionary” which aims to “describe all words of all languages using definitions and descriptions in English”. Under the Creative Commons act, an electronic download of the English🡪French database entries is available (Matthias Buchmeier / Wiktionary, 2018).

Each of the 72,827 entries in the dictionary has the following format:

***English word/MWE – Part of speech – Further detail / definition – French Translation***

Many entries do not contain a direct translation but rather a reference to ‘see x entry’. Still others contain not only one single translation, but up to 26 suggested translations for a single English entry. Finally, multiple entries exist in some cases for the same English word + POS tag combination, where it has different meanings. To prepare the dictionary for our purpose, we use Pandas to wrangle the data into a usable form (see Chapter V Implementation: Part B iii for further details of this process.)

## E. Evaluation Methods

We use two key measures throughout the design and evaluation of our tool:

**Pearson Correlation Coefficient:** The linear relationship between scores derived from a particular feature, and the teacher-derived difficulty scores is measured via the Pearson Correlation Coefficient. The coefficient varies from -1 to 1, with -1 or +1 implying an exact linear relationship (negative or positive), while 0 implies no correlation. A p-value is also calculated, indicating the statistical significance of the found correlation. (StatSoft, Inc. (2013))

**Mean Squared Error:** We use the mean squared error to measure the average squared difference between teacher difficulty scores and feature scores. This allows us to obtain an overall error rate as well as examine more closely those vocabulary items with the most significant error throughout the design process, helping to identify potential problems in our feature scoring parameters.

# IV: IMPLEMENTATION

### **Chapter Summary**

The tool is comprised of a number of Python classes which each take a different role:

**A. Vocabulary Item Extraction:** Out TextItems class takes an argument of raw text, and returns a list of vocabulary items therein, including both single word items and MWEs.

**b: Scoring Features:** We extract a number of features from each vocabulary item and examine each feature’s validity as a contributing factor to an overall difficulty score. Each feature is controlled by a particular class which extracts feature information from the list of vocabulary items, then computes and returns a normalised score between 0 and 1, where 0 is posited to correspond to least difficulty, and 1 to most difficulty.

B. Scoring Features: i. Vocabulary Item Frequency:Computation of a score which measures the vocabulary item’s frequency of occurrence in an external reference corpus.

B. Scoring Features: ii. POS-Tagged Frequency:As above, but we now calculate frequency scores based on vocabulary items taken as a tuple pair with their POS tag. As such, (*run, NOUN)* will have a different POS-Tagged frequency score to *(run, VERB)*.

**B. Scoring Features: iii. L1 Similarity Feature:** Assignment of a score to the vocabulary item based on its similarity (measured formally via the Levenshtein Distance) to its translation into the specified L1, in this case French.

**B. Scoring Features: iv. Part of Speech Tag Feature :** Calculation of a score based on the vocabulary item’s POS tag in its own right.

**B. Scoring Features: v. Vocabulary Item Length:** We compute a score based on the length of the vocabulary item.

**C. Feature Contribution: Assigning an Overall Score**: We use linear regression to identify the most useful features and their coefficients for predicting teacher scores.

## A. Vocabulary Item Extraction

(See also: Supporting Material: *tool / TextItems.ipynb )*

**Text Input**Our TextItems class is designed to extract vocabulary items. We instantiate the class by passing it a plain-text, utf8-encoded file containing the input text. This is initially stored as a single string.

**Vocabulary Item Extraction: Single Word Tokens**

Using first sentence tokenization and then word tokenization, we can progressively break this string up into first sentences, and then words and punctuation, therefore extracting a list of word (or symbol) tokens within the text. This allows us to extract single word vocabulary items at the simplest level.

**Vocabulary Item Extraction: MWEs**

As discussed however, it is essential to deal with MWEs within the text, and this must be integrated with the tokenisation and extraction of single word items so as to avoid any words being extracted twice (once as a single word and then once again as part of an MWE). We use a lexicon approach to deal with MWEs. Of nine compiled and annotated MWE lexical resources licensed for free download and use under the Creative Commons Attribution, the *Multiword entries from English Wiktionary* lexicon is by far the most complete, containing 82,175 MWEs (Schneider et al, English Multiword Expression Lexicons EnWikt, 2014).

At a basic level, we can now loop through this list of MWEs, search for each within our text, and upon finding one, add it to the pool of vocabulary items to be scored. We refine this process somewhat as follows:

1. All MWEs in the lexicon and all text tokens in the input text are transformed to lower case for matching purposes (to avoid matches being missed at the start of sentences where a word is capitalised)
2. We create an index for the input text, whereby each single word token has a unique corresponding variable in a simple Boolean array with of equivalent length. All values are set to True. This permits us to track which tokens are part of MWEs, and therefore should not be considered as single word tokens.
3. An algorithm loops through each sentence of the input text. Firstly, it checks whether an MWE in the lexicon exist in the raw string of each sentence. When a match is found, the sentence is tokenized and transformed into a series of n-grams of corresponding length to the MWE. A simple helper method then returns the position of the MWE n-gram in the sentence. The corresponding Boolean values to these token positions in the index array are set to False. The discovered MWE is added to a list.
4. All single word tokens from the text which have not been used in an MWE at the end of the loop (i.e. whose Boolean flag in the index array is still True) are added to the vocabulary item collection, along with the discovered MWEs.

This approach extracts a number of MWEs, however problems arise with verb tenses. Considering Sample Text 2, the past tense MWEs ‘laid claim’ and ‘held up’ were not extracted because they exist in the lexicon only in their infinitive forms (‘lay claim’ / ‘hold up’). To solve this problem, we integrate POS-tagging and lemmatizing into our algorithm to identify non-infinitive verbs in the text and consider them for an MWE match in their lemmatized forms.

**Vocabulary Item Extraction: Return Options**

We build in a number of options for returning the list of found vocabulary items. These can be returned with or without an accompanying POS tag (some features being POS-dependent) and with or without the filtering out of non-dictionary words (those not appearing in NLTK’s WordNet, nor in a list of stopwords, nor being a hyphenated word). We use this option throughout our analysis, as it is effective in filtering out ‘non-words’ such as proper nouns and company-specific acronyms.

**Initial Evaluation**

Figure 5 shows all the MWE vocabulary items identified as of interest by teachers for each text, ordered by their corresponding difficulty scores (highest scores being the most difficult and therefore arguably most import to extract). We observe that more than 50% of the 40 MWEs have been extracted or partially extracted (for example, for text 1 ‘sign off’ rather than ‘sign off on’ was extracted), including more than 70% of the top scoring median in the case of texts 1 and 2. For further discussion of potential improvements to be made to this algorithm see Chapter V: Evaluation, Part A.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Text 1** | | |  | **Text 2** | | |  | **Text 3** | | |
| **MWE** | **SCORE** | **EXTRAC- TED?** | **MWE** | **SCORE** | **EXTRAC- TED?** | **MWE** | **SCORE** | **EXTRAC- TED?** |
| fair-weather friend | 1 | Y | held up | 1 | Y | ahead of schedule | 0.83 | P |
| sign off on | 1 | P | sign up | 1 | Y | lie ahead | 0.83 | N |
| make up | 0.83 | Y | spur on | 1 | Y | retail sales | 0.33 | N |
| on track | 0.83 | Y | changing tide | 1 | N |  | | |
| run fire drills | 0.83 | P | follow suit | 0.83 | Y |
| in a row | 0.67 | Y | switch from | 0.83 | N |
| off target | 0.67 | N | lay claim | 0.83 | Y |
| booster projects | 0.5 | N | in the wake of | 0.83 | Y |
| press on | 0.5 | Y | on its books | 0.67 | N |
| letter of intent | 0.33 | N | top the list | 0.5 | N |
| pull out of | 0.33 | P | fossil fuels | 0.33 | Y |
| re-engage in | 0.33 | N | local authority | 0.33 | N |
| a few weeks back | 0.17 | P | move away from | 0.33 | N |
| crazy excited for | 0.17 | N | set a target | 0.17 | N |
| get to focus on | 0.17 | P | as well as | 0.17 | Y | ***Figure 5: MWEs extracted (Y), partially extracted (P) or not extracted (N) by the TextItems algorithm.*** | | |
| given the circumstances | 0.17 | N | vast majority | 0.17 | N |
| long term | 0.17 | Y | with regards to | 0.17 | N |
| might have | 0.17 | N |  | | |
| pushed for | 0.17 | N |
| rather than | 0.17 | Y |

Note: Going forward, in the case of vocabulary items such as ‘sign off’, where a significant partial match of the teacher-identified MWE ‘sign off on’ has occurred, the teacher score for the whole MWE will be used for the partial MWE.

## B. Scoring Features: i. Vocabulary Item Frequency

(See also: Supporting Material: *tool / FrequencyFeature.ipynb )*

Our FrequencyFeature class is instantiated with a dictionary which maps 681,861 unique vocabulary items (comprising both single words and MWEs) to their frequencies in the BNC. The class’s methods use this dictionary to map vocabulary items to BNC frequency scores, and from this calculate a normalised difficulty score.

**Incorporating MWEs into the Frequency Distribution**

NLTK’s Frequency Distribution package combined with the BNCCorpus Reader readily facilitates the production of a single word frequency distribution from the BNC. However, for our purpose it is also necessary to also incorporate MWEs. To achieve this, we used the same algorithm described in the previous section **Vocabulary Item Extraction: Multiword Expressions** to identify MWEs in the BSC and count their frequencies. 39,807 unique MWEs were thus identified as samples in the BNC, with a total of 4,536,599 outcomes.

**From Frequency Distributions to Frequency Scores**

The single word and MWE frequency distributions are combined and transformed into a dictionary. Each vocabulary item extracted from our input text is mapped to a frequency figure from this dictionary. The process for calculating the normalised frequency score is as follows:

* All frequency values are smoothed by adding 1 to each. This avoids a later problem of dividing by zero for those vocabulary items which do not appear in the BNC.
* The smoothed values are transformed into proportions:  
   Vocabulary Item Smoothed Frequency / Total Smoothed Frequency Distribution Outcomes
* We then compute the logarithms of these proportions (using log frequencies being more appropriate for working with such language data where outcomes for a relatively low number of common samples (words) are much higher than for the majority of samples in the distribution)
* SciKitLearn’s MinMaxScaler provides convenient functionality for normalising the logged values
* Finally, this value is inverted, giving each vocabulary item a normalised frequency score between 0 (most frequent, and therefore hypothetically ‘easier’) and 1 (least frequent, and therefore hypothetically ‘more difficult’).

**Initial Evaluation**

Figure 6 shows, for each of our three sample texts, the mean square error and the Pearson correlation co-efficient between the target scores and the normalised frequency scores:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Text 1** | **Text 2** | **Text 3** |
| **Mean Squared Error Rate** | 0.14 | 0.17 | 0.18 |
| **Pearson correlation co-efficient** | 0.49 | 0.32 | 0.31 |
| **pvalue:** | 0.0 | 0.0 | 0.0 |
| ***Figure 6: Mean Squared Error & Pearson Correlation Co-Efficient, Teacher Score/ Frequency Score***  See also: Supporting Material: Texts1to3\_Feature\_Evaluation\_and\_Linear\_Regression.ipynb | | | |

## B. Scoring Features: ii. POS-Tagged Frequency

(See also: Supporting Material: *tool / POSFrequencyFeature.ipynb)*

Measuring the frequency of a word alone has limitations: some words are significantly more common when used as one part of speech than as another. From Sample Text 2, we note that the word *drive* scored highly with teachers, though on the surface this would appear to be a common, not particularly difficult word. When we consider its context in the text, we note that it was used as a noun rather than the more commonplace and easily understood verb: ‘*The company’s* ***drive*** *to boost profits’*.

We therefore design a feature which samples vocabulary items for a frequency distribution in tandem with their POS tags. Our POSFrequencyFeature class uses this POS-tagged frequency distribution to compute its scores.

**POS-Tagged Frequency Distribution**

The BNC is in itself annotated with POS tags, however these do not map well to the different tagset used by NLTK in other parts of our algorithm. We therefore use the NLTK to POS-tag the BNC corpus in making our tagged frequency distribution, giving greater parity with the tags given to the vocabulary items extracted from our sample texts. MWEs found within the BSC were given a custom ‘MWE’ POS tag.

**From Frequency Distributions to POS-Tagged Frequency Scores**

The calculation of POS-Tagged Frequency scores from the frequency distribution values follows the same procedure as described in the previous section, though with samples taking the form of tuple-POS tag pairs.

**Initial Evaluation**

Figure 7 shows, for each of our three sample texts, the mean square error and the Pearson correlation co-efficient between the target scores and the normalised frequency scores:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Text 1** | **Text 2** | **Text 3** |
| **Mean Squared Error Rate** | 0.15 | 0.18 | 0.22 |
| **Pearson correlation co-efficient** | 0.48 | 0.28 | 0.28 |
| **pvalue:** | 0.0 | 0.0 | 0.0 |
| ***Figure 7: Mean Squared Error & Pearson Correlation Co-Efficient, Teacher Score/POS-Tagged Frequency Score***  See also: Supporting Material: Texts1to3\_Feature\_Evaluation\_and\_Linear\_Regression.ipynb | | | |

We do not note an improvement in the Mean Squared Error nor the Pearson correlation co-efficient compared with the non-POS tagged Frequency scores. It should be noted that including POS tags in the frequency distribution makes the data much sparser. Further, the use of NLTK’s POS tagger is not ideal for two reasons: firstly, its assigned tags are not always accurate, secondly, they are very fine grained (representing e.g. not simply ‘verb’ but ‘verb in the infinitive’, ‘verb in the past simple’, ‘verb in the past participle’ etc) leading to a sparser frequency distribution. A future iteration of this feature should strive to improve the accuracy of POS tagging our input text, and use a mapping to translate both these tags and those within the BNC frequency distribution to broader, more universal POS tags. Stemming and lemmatizing all words within the frequency distribution is another possibility to reduce sparsity, however as added prefixes and suffixes may affect the difficulty of a vocabulary item for a learner, we did not choose to stem or lemmatize for the design of this feature.

## B. Scoring Features: iii. L1 Similarity Feature

(See also: Supporting Material: *tool / L1SimilarityFeature.ipynb)*

Our L1Similarity class is instantiated with a dictionary of data from English-to-French Wiktionary, which maps 72,827 POS-tagged English words and MWEs to French translations. The class’s methods use this dictionary to find translations for each of the vocabulary items extracted from input text, and calculate the Levenshtein distance between the English and French strings. The resulting score is inverted, so that the more similar the two strings are, the lower their score will be (hypothesising that English words with similar French translations are easier for French L1 learners).

**Preparation of the English-to-French Dictionary**

As previously noted, the dictionary can contain up to 26 suggested translations for a single English entry. Also, multiple entries exist in some cases for the same English word + POS tag combination, where it has different meanings (eg the noun *record* can mean both a vinyl disk, or a log of evidence). In dealing with these issues, we used the following principles:

* Where an entry had multiple translations, the Levenshtein distance between the English entry and each French translation was calculated and stored. The single translation among these with the highest Levenshtein distance to the English entry (i.e. the translation that most closely resembles the English) was preserved as the unique translation for that entry. If we consider the noun *responsibility,* which has the definition of *obligation*, two French translations are listed: *responsibilité* and *devoir.* With the former having a Levenshtein ratio of 0.9 (compared to *devoir*’s 0.3), and being sufficient for a French speaker to understand the word *responsibility,* only this translation is preserved in the dictionary.
* Where a vocabulary item with its part of speech tag has more than one entry (row), only the entry whose translation has the highest Levenshtein ratio is preserved.

Note: We acknowledge the necessary limitations of this approach in meaning that some words will be incorrectly translated, though observation of the translations and resulting L1 Similarity Scores suggests this is not a significant problem in our research.

**Use of the English-to-French Dictionary**

* The cleaned dictionary is then used to assign the appropriate score to each item in the inputted text’s vocabulary pool, based on looking up the vocabulary item with its POS tag to find its French translation and the Levenshtein distance between the two.
* **Use of the POS tag:** Without POS tags, translation errors occur, such as the adjective ‘sole’ in ‘*Our* ***sole*** *problem is money’* being translated to the French **‘*sole’*** – which is a correct translation for the noun *soul* (fish) but not the adjective *sole* (the correct translation being ‘*seul’*). Inclusion of the POS tag provides to a much better translation. It was however necessary to create a mapping for POS tags between the dictionary’s tag types and NLTK’s tags (see Appendix C).
* Where no entry is found in the dictionary for the vocabulary item with its POS tag, three further attempts are made:
  + The item is stemmed via NLTK’s Porter Stemmer (thus *dogs* becomes *dog)* and searched for with its POS tag in this form
  + The item is lemmatized using NLTK’s WordNet Lemmatizer (thus ***laid***becomes ***lay****)* and searched for with its POS tag in this form
  + The item is searched for independent of its POS tag (stemmed / lemmatized if necessary)
  + The Levenshtein Distance is inverted to provide the difficulty score for each item
  + Vocabulary items that remain unfound in the dictionary receive a score of 1.

**Initial Evaluation**

Figure 8 shows, for each of our three sample texts, the mean square error and the Pearson correlation co-efficient between the target scores and the L1 similarity scores:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Text 1** | **Text 2** | **Text 3** |
| **Mean Squared Error Rate** | 0.37 | 0.32 | 0.32 |
| **Pearson correlation co-efficient** | 0.21 | 0.12 | 0.04 |
| **pvalue:** | 0.0 | 0.07 | 0.56 |
| ***Figure 8: Mean Squared Error & Pearson Correlation Co-Efficient, Teacher Score / L1 Similarity Score***  *Supporting Material: Texts1to3\_Feature\_Evaluation\_and\_Linear\_Regression.ipynb* | | | |

## B. Scoring Features: iv. Part of Speech Tag Feature

(See also: Supporting Material: *tool / POSFeature.ipynb)*

|  |
| --- |
| **C:\Users\rowena\AppData\Local\Microsoft\Windows\INetCache\Content.MSO\13C81E65.tmp** |
| ***Figure 9: Mean teacher scores assigned to vocabulary items of each POS type, Texts 1 - 3*** |

When studying the vocabulary items which have received teacher scores of >0, we note that these tend to correspond to certain parts of speech: in other words they are more likely to be MWEs, nouns, verbs, adverbs and and adjectives than determiners, prepositions, pronouns etc. Figure 9 shows the mean teacher-assigned score across texts 1 to 3 for each type of vocabulary item (note that the POS tags have, for this purpose, been mapped to broader categories than those used by NLTK, see Appendix D).

|  |  |  |
| --- | --- | --- |
| **Broad POS Tag** | **Mean Teacher Score** | **Normalised POSTag Score** |
| MWE | 0.15 | 1.00 |
| NOUN | 0.08 | 0.55 |
| VERB | 0.08 | 0.51 |
| ADJ\_ADV | 0.06 | 0.42 |
| PRP\_NOUN | 0.02 | 0.13 |
| PREPS\_DETS\_CONJ | 0.00 | 0.03 |
| INTERJ\_FOREIGN | 0.00 | 0.00 |
| MODAL | 0.00 | 0.00 |
| PRONOUN | 0.00 | 0.00 |
| PUNC\_NUM\_OTHER | 0.00 | 0.00 |
| WH\_TH\_HOW | 0.00 | 0.00 |
| ***Figure 10: Mean teacher scores assigned to vocabulary items of each POS type, Texts 1 – 3*** *Supporting Material: tool/files/posscores.xlsx* | | |

**POSTagClass**

We normalise these average scores and use these values to provide a score for vocabulary items based on their broad POS tag. The POSTagClas thus maps all received vocabulary items to their broad POS tag, and gives them the appropriate normalised score whereby all MWEs receive a score of 1 from this feature, all nouns 0.55 etc.

**Initial Evaluation**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Text 1** | **Text 2** | **Text 3** |
| **Mean Squared Error Rate** | 0.25 | 0.22 | 0.20 |
| **Pearson correlation co-efficient (all pvalues 0.0)** | 0.24 | 0.30 | 0.12 |
| ***Figure 11: Mean Squared Error & Pearson Correlation Co-Efficient, Teacher Score / POS Tag Score*** *Supporting Material: Texts1to3\_Feature\_Evaluation\_and\_Linear\_Regression.ipynb* | | | |

## B. Scoring Features: v. Vocabulary Item Length

(See also: Supporting Material: *tool / LengthFeature.ipynb)*

As noted in section II iii, word length has in the past been considered a measure of vocabulary difficulty, with longer words supposed more difficult. Our LengthFeature class assigns scores to vocabulary items based on their length.

**Designing the Length Feature**

In designing this feature, we used the frequency distribution of the BNC corpus (for both single and MWEs) to create a new frequency distribution which this time sampled distinct word lengths rather than distinct words. For the purposes of this exercises, a MWE considered as a single word, with its total number of characters constituting its length.

Using numpy, the frequency distribution was binned into bins of 5% based on word lengths and corresponding frequencies. Each possible word length was assigned a normalised score based on its length and bin, thus longer vocabulary items receive higher scores from this feature than shorter ones.

**Initial Evaluation**

Figure 12 shows, for each of our three sample texts, the mean square error and the Pearson correlation co-efficient between the target scores and the length scores:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Text 1** | **Text 2** | **Text 3** |
| **Mean Squared Error Rate** | 0.04 | 0.05 | 0.05 |
| **Pearson correlation co-efficient** | 0.21 | 0.19 | 0.15 |
| **pvalue:** | 0.00 | 0.01 | 0.03 |
| ***Figure 12: Mean Squared Error & Pearson Correlation Co-Efficient, Teacher Score / Length Score***  *Supporting Material: Texts1to3\_Feature\_Evaluation\_and\_Linear\_Regression.ipynb* | | | |

## C. Feature Contribution: Assigning an Overall Score

(See also: Supporting Material: *Texts1to3 Feature\_Evaluation and Linear\_Regression.ipynb   
and tool / TextScorer.ipynb)*

Given the small size of our dataset, we must be cautious in using statistical methods to learn feature weightings. We nonetheless use SciKitLearn’s Linear Regression module to fit models to our data, (with a helper method to view coefficients (Sabas, 2014)) informed by domain expertise and our own research.

**Feature Selection**

Having observed a better correlation with targets from the Frequency feature than the POS-Tagged Frequency feature, and preferring independency between features, we discount the latter and use the remaining four features in our initial attempt. The L1 Similarity feature showed a correlation coefficient of 0.21 with target scores for Sample Text 1, 0.12 for Text 2 and only 0.04 for Text 3. However, research from our teachers nevertheless suggests the validity of L1 Similarity as a predictor for vocabulary difficulty, and given our sparse dataset (and the pvalues for the Pearson coefficients for particular feature for texts 1 and 2 suggesting a lack of statistical significance) we select the L1 Similarity Feature for inclusion, along with Frequency, Length and POS Tag.

**Learning Coefficients**

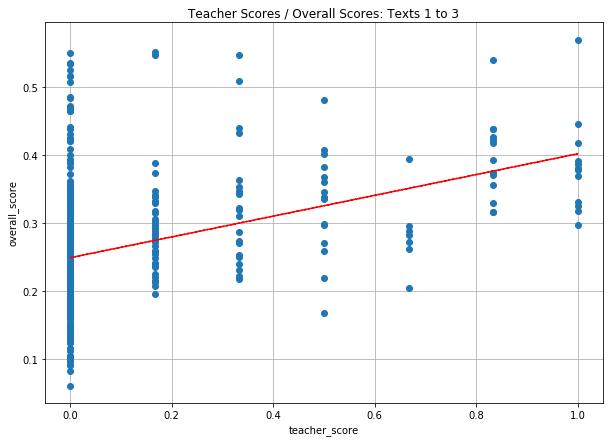
We combine the feature data and target scores from Texts 1 – 3 to form our X and Y arrays with numpy. Linear Regression returns the following co\_efficients:

|  |  |  |  |
| --- | --- | --- | --- |
| **Frequency Feature** | **L1Similarity Feature** | **Length Feature** | **POS Tag Feature** |
| 0.43 | 0.10 | -0.08 | 0.04 |

Interestingly, we observe a negative coefficient with our length feature, despite the feature having a positive correlation with target scores in its own right. Observing that there is likely to be a lack of independence between the Length and POS Tag features (as MWEs are likely to be longer in length with prepositions and determiners shorter in length) we discount this feature from our overall difficulty score. We use the learned coefficients applied to the Frequency Feature, L1Similarity Feature and POS Tag Feature to determine the overall difficulty score for each text.

**Initial Evaluation: Texts 1 to 3 (Overall Difficulty Scores)**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Text 1** | **Text 2** | **Text 3** |
| **Mean Squared Error Rate** | 0.06 | 0.07 | 0.06 |
| **Pearson correlation co-efficient *(all corresponding pvalues=0.0)*** | 0.52 | 0.38 | 0.31 |
| ***Figure 13: Mean Squared Error & Pearson Correlation Co-Efficient, Teacher Score / Overall Score*** | | | |

**⦁** vocabulary item ***Figure 14: Scatter graph and line of best fit: Teacher Scores / Overall Scores Texts 1 – 3.***

We note that for all three texts, using our learned coefficients to calculate an overall score from three features provides a better correlation co-efficient against teacher target scores than for any of the individual features in their own right.

It is also observable that the feature coefficients sum to 0.57 rather than 1, meaning that vocabulary items will receive difficulty scores in the range 0 to 0.57, with obtention of a 1 score being impossible. Notably, in the target data, a majority of vocabulary items have a teacher difficulty score of 0, giving our model a propensity to “aim low”. For the purposes of our exercise in predicting the difficulty scores as calculated from our training data, and/or extracting the most difficult vocabulary from a text, we continue to use the 0.57-summed coefficients. However, for an application where a full range of difficulty scores would be desirable, we would recommend scaling the coefficients to a more normalised distribution (Frequency: 0.75, L1: 0.18, POS: 0.07)

# V: EVALUATION

## Chapter Summary

In this chapter we turn to the three holdout texts: Texts 4, 5 and 6 to test our tool’s performance in respect to three categories:

**A. Vocabulary Item Extraction:** We consider to what extent the tool successfully extracts target single word and MWE vocabulary items from the three texts, and compare its performance to two other comparable online tools.

**B. Difficulty Scores:** We use Pearson’s Correlation Coefficient to evaluate the predicted difficulty scores for the three texts against the teachers’ difficulty scores, both for the overall difficulty score with our learned feature coefficients, and two notable features in their own right.

**C. Vocabulary Item Ranking:** Aside from predicting the specific difficulty scores derived from our teacher research data, we consider how effective our tool was at predicting the most difficult vocabulary items within each text.

### A. Vocabulary Item Extraction

(See also: Supporting Material: *Texts 4to6 Vocabulary Item Extraction Evaluation.ipynb)*

**Single Word Vocabulary Items**

We identify 89 single word items as having a teacher difficulty score of >0 across Texts 4 to 5. Of these 86 were successfully extracted by the algorithm. Single words not extracted and returned are ‘productizing’ and ‘impactful’, which are not recognised as valid words by Wordnet’s lemmatizer, and thus get filtered out by the dictionary check.

**MWE Vocabulary Items**

We noted that for the three sample tests used in the implementation stage, more than 50% of the 40 MWEs identified as to some extent ‘difficult’ were extracted or partially extracted, including more than 70% of the top difficulty-scoring median in the case of texts 1 and 2. Figure 15 now considers texts 4 to 6 in these respects:

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Text 4** | | |  | **Text 5** | | |  | **Text 6** | | |
| **MWE** | **SCORE** | **EXTRAC- TED?** | **MWE** | **SCORE** | **EXTRAC- TED?** | **MWE** | **SCORE** | **EXTRAC- TED?** |
| ('expand', 'the', 'pie') | 1.00 | N | ('throw', 'into', 'disarray') | 1 | N | ('social', 'venture') | 0.83 | N |
| ('scale', 'up') | 1.00 | Y | ('deal', 'with') | 0.67 | Y | ('track', 'progress') | 0.67 | N |
| ('break', 'down') | 1.00 | Y | ('going', 'forward') | 0.67 | Y | ('team', 'up', 'with') | 0.67 | Y |
| ('land', 'a', 'client') | 0.83 | N | ('cover', 'costs') | 0.50 | N | ('parental', 'leave') | 0.67 | Y |
| ('on', 'the', 'fence') | 0.83 | Y | ('along', 'with') | 0.33 | Y | ('set', 'goals') | 0.5 | N |
| ('on', 'track') | 0.83 | Y | ('due', 'to') | 0.33 | Y | ('date', 'back', 'to') | 0.33 | N |
| ('walk', 'you', 'through') | 0.67 | N | ('force', 'to') | 0.33 | N | ('leading', 'news') | 0.33 | N |
| ('light', 'at', 'the', 'end', 'of', 'the', 'tunnel') | 0.67 | Y | ('go', 'through') | 0.33 | Y | ('slow', 'down') | 0.33 | Y |
| ('on', 'the', 'side') | 0.67 | Y | ('set', 'restrictions') | 0.33 | N | ('mission', 'statement') | 0.17 | Y |
| ('spread', 'out') | 0.67 | Y | ('apply', 'for') | 0.17 | N | ('pulmonary', 'aerterial') | 0 | N |
| ('lie', 'around') | 0.67 | Y | ('illegal', 'activities') | 0.17 | N | ***Figure 15: Target MWEs for Texts 4 to 6, with teacher scores and extraction success (Y), failure (N) or partial success (P)*** | | |
| ('eat', 'the', 'fees') | 0.50 | N | ('out', 'of'', 'fear') | 0.17 | N |
| ('put', 'in', 'the', 'work') | 0.50 | P |  | | |
| ('check', 'out') | 0.50 | Y |
| ('on', 'and', 'on') | 0.50 | Y |
| ('on', 'your', 'own', 'time') | 0.33 | N |
| ('side', 'business') | 0.33 | N |
| ('start', 'off') | 0.33 | Y |
| ('focus', 'on') | 0.17 | N |
| ('master', 'class') | 0.17 | N |
| ('payment', 'plan') | 0.17 | N |
| ('cut', 'back') | 0 | Y |

Our program extracted or partially extracted 50% of the 44 MWEs identified across these two texts. Once more we note that it performed particularly well with those MWEs which scored 0.5 or greater with our teachers, extracting 62% of these.

We may also evaluate the success of MWE extraction by comparing our tool to two others mentioned in Chapter III Part D: Vocabulary.com’s *VocabGrabber* and Cambridge English Profile’s *TextInspector.* Of the 44 MWEs listed by our teachers for texts 1-3, VocabGrabber extracted 8 (16%) and TextInspector only 6 (14%). Our tool performs very well in terms of MWE extraction compared to these two offers.

**Vocabulary Item Extraction: Overgeneration**

When considering the vocabulary items extracted, we note a slight problem of over-generation, that is some vocabulary items being extracted which we would in fact wish to exclude from the output:

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Text 4** | **Text 5** | **Text 6** |
| Unique Vocabulary Items Extracted | 296 | 250 | 223 |
| Of which undesirable MWEs | 15 | 7 | 9 |
| Of which undesirable single words | 2 | 12 | 1 |
| %age ‘good’ extractions | 94% | 92% | 96% |
| ***Figure 16: Analysis of all vocabulary items extracted from Texts 4 – 6.*** | | | |

Single-word items identified as undesirable include a small number of tokens resulting from contractions (‘d’, ‘t’, ‘m’), as well as some proper nouns (‘Japan’, ‘Paris’, ‘Kyoto’, ‘Liza’) which would not be considered ‘vocbaulary’ in the context of ESL. In terms of MWEs, though we have established it can be hard to define what is and isn’t a ‘valid’ MWE, our lexicon included, and led to the extraction of, some entries which were of dubious use to our purposes, again not serving as ‘taught’ vocabulary items in ESL: ‘how do’, ‘of a’, ‘do it’, ‘there are’, ‘I want to know’.

**Improvements**

To perfect single-word vocabulary item extraction, a more effective filter for proper nouns is desirable: perhaps using more efficient POS tagging to filter out proper nouns (we noted NLTK tended to over-use the proper noun POS tag, and so did not wish to filter words with this tag out completely, as a number of non-proper nouns would be lost).

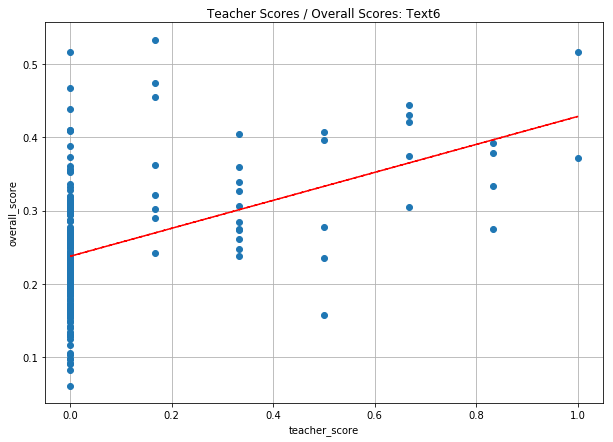
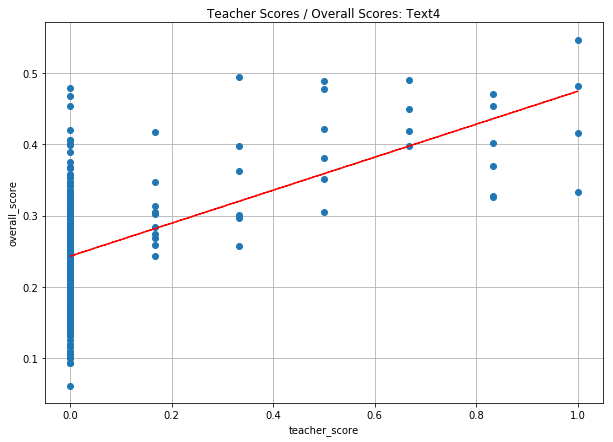
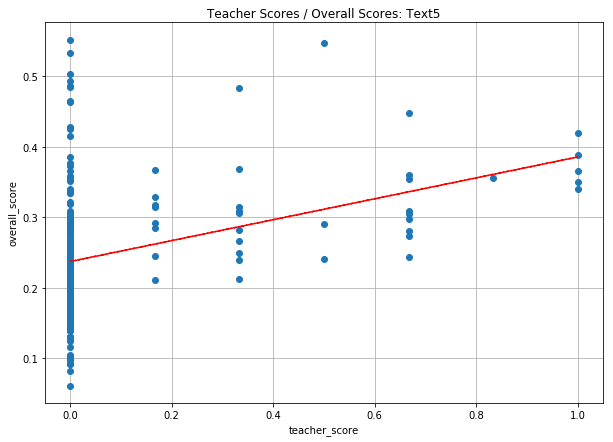
Potential improvements that could be made to the extraction of MWEs are as follows:

* **Refining the MWE lexicon:** The lexicon we use to look up MWEs is evidently not complete. From looking at the non-extracted MWEs in texts 4 to 6, we note that some fairly common expressions such as ‘focus on’, ‘master class’, ‘payment plan’ were not extracted due to absence from the lexicon. Adding such further entries to the lexicon would thus improve performance. Similarly, some fore mentioned entries in the lexicon which are not MWEs for ESL purposes could be worth deleting.
* **Non-contiguous MWEs:** Some MWEs were in the lexicon, but not identified as they were non-contiguous in the text: ‘walk through’ was in the lexicon, but appeared in text 4 as ‘walk you through’, with ‘date back also in the lexicon but appearing in text 6 as ‘dates as far back as’. To successfully identify these MWEs in their non-contiguous contexts, we can envisage either expanding the lexicon to include common non-contiguous variations of existing entries, or adding significant improvements to the algorithm to search for certain transitive or flexible MWEs in the lexicon in a variety of contexts in the text.
* **Accepting limitations:** With no hard and fast means of determining what is and what is not a MWE, neither in the field of NLP nor ESL, the identification of MWEs within a text is likely to always be an imperfect process, and as such it is necessary to accept limitations in any such algorithm.

### B. Difficulty Scores

(See also: Supporting Material: *Texts4to6 Difficulty Score Evaluation.ipynb)*

We evaluate the correlation of the final difficulty scores (generated using the coefficients of the Frequency, L1 Similarity and POS Tag features as described in Chapter IV Part C) with the teacher targets for Texts 4 to 6 as follows:

** ⦁** vocabulary item **  *Figure 17: Scatter graph and line of best fit: Teacher Figure 18: Scatter graph and line of best fit: Teacher   
 Scores / Overall Scores Text 4 Scores / Overall Scores Text 5***

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Text 4** | **Text 5** | **Text 6** |
| **Mean Squared Error Rate** | 0.07 | 0.07 | 0.06 |
| **Pearson correlation co-efficient** *(all corresponding pvalues=0.0****)*** | 0.54 | 0.34 | 0.45 |
| ***Figure 20: Mean Squared Error & Pearson Correlation Co-Efficient, Teacher Score / Overall Difficulty Score, Texts 4-6*** | | | |

***Figure 19: Scatter graph and line of best fit: Teacher   
 Scores / Overall Scores Text 6***  
We note a correlation co-efficient of between 0.34 and 0.54 for each text. It is informative to   
  
also evaluate the Frequency and L1Similarity Features in their own right for these texts:

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | **Text 4** | **Text 5** | **Text 6** |  | **Text 4** | **Text 5** | **Text 6** |
| **Mean Squared Error Rate** | 0.16 | 0.16 | 0.17 | 0.35 | 0.36 | 0.3 |
| **Pearson correlation co-efficient** | 0.49 (*pvalue 0.0)* | 0.37 (*pvalue 0.0)* | 0.42 (*pvalue 0.0)* | 0.24  (*pvalue 0.0)* | -0.02 (*pvalue 0.7)* | 0.13 (*pvalue 0.5)* |
| ***Figure 21: Mean Squared Error & Pearson Correlation Co-Efficient, Teacher Score / Frequency Score*** | | | | ***Figure 22: Mean Squared Error & Pearson Correlation Co-Efficient, Teacher Score / L1 Similarity Score*** | | |

When considering the learned scores using all three features, to the scores obtained from the Frequency feature alone, all three texts show an improvement in mean squared error rate, with Texts 4 and 5 show also showing an improvement in correlation, when all contributing features are used.

Isolating the L1 Similarity feature and evaluating its correlation to teacher scores (Chapter IV Part B ii), we observe a good correlation between these scores and teacher scores for Texts 1 and 4, with Texts 2, 3 and 6 having a weaker yet still positive correlation and Text 5 showing a negative correlation. Interestingly Texts 1 and 4 are both emails (Texts 2 and 4 being news articles, and 3 and 6 being company reports). With emails being more likely to contain informal language of a non-French/Latin etymology (Harries, 2013) it could be the case that this feature is more useful when working with such texts rather than the more formal company reports and news articles. We continue to evaluate the utility of the L1 Similarity feature in the next section.

**Improvements**

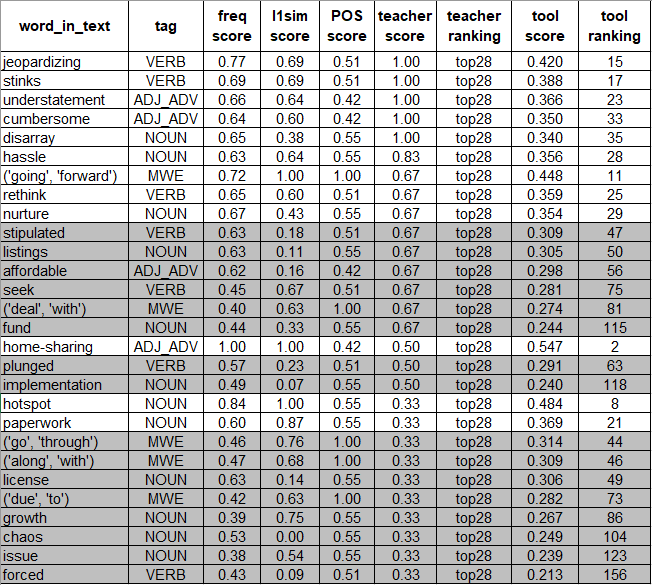
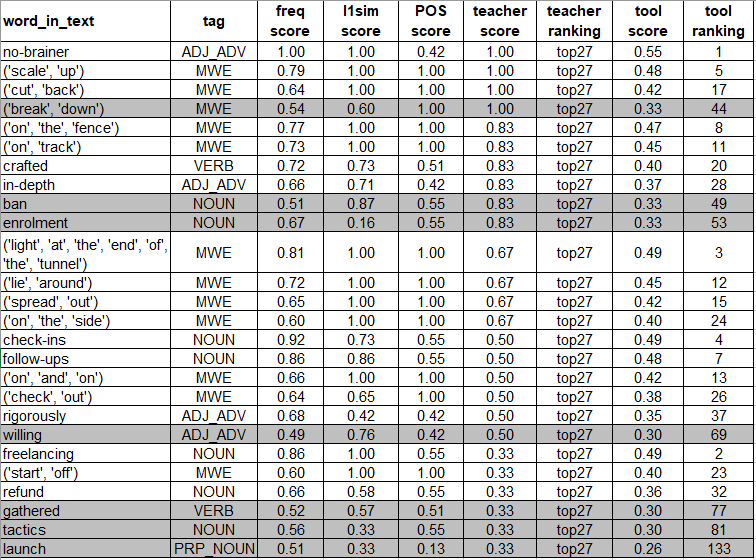
The clear way to improve our difficulty scoring algorithm would be to obtain more data for training and testing purposes: a greater selection of sample texts, with more participating teachers. In order to have an effective basis for scoring vocabulary items of a pre-intermediate level, we would also wish teachers to engage with the difficulty of all words in the sample text rather than only the most difficult. We discussed in Chapter III Part A some of the problems in obtaining such data.

Another potential for improvement would be the extraction of further features from which to learn target difficulty scores. In working with the English to French dictionary, we noted that some English entries had multiple translations: capturing the number of translations for a single word could be in itself a useful measure of difficulty.

### C. Vocabulary Item Rankings

(See also: Supporting Material: *tool / files / vocab\_ranking\_evaluation.xlsx)*

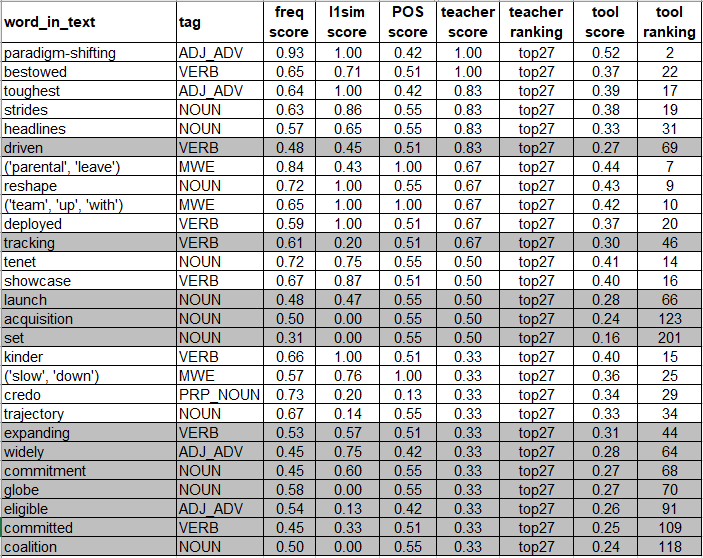
We examine the top ~30 items from each text in terms of teacher difficulty ranking, and consider how many are within the top ~40 items predicted by our tool

****For **Text 4**, 67% of the top-ranking items by teacher score are within our tool’s top 40. Notably, some items with a ~0.5 frequency score (indicating averagely common words) such as ‘ban’, ‘willing’, ‘gathered’ and ‘launch’ were included within our teachers’ selection, which impacted on the overall score assigned by the tool. The mean L1 similarity score for items in the teachers’ top 27 is 0.78, in contrast to a text average of 0.56.

**Figure 23 : Top ranked vocabulary items by teacher difficulty (the top 27 includes all vocabulary items with a score of 0.33 or more)**Items outside the tool’s prediction of the top 40 items are highlighted in grey

For **Text 5**, 43% of the top-ranking items by teacher score are within our tool’s top 40, 61% within the top 50. Here, we see the weaker correlation between L1 similarity and teacher-perceived difficulty, with the average L1 score being comparable to the text average.

**Figure 24: Top ranked vocabulary items by teacher difficulty (the top 28 includes all vocabulary items with a score of 0.33 or more)**

For **Text 6****,** 67% of the top-ranking items by teacher score are within our tool’s top 40. We note that the relatively common verb ‘driven’ rates highly with teachers: in the text the word is used metaphorically rather than literally: this type of word-sense ambiguity and the difficulty it introduces is not currently accounted for by our tool.

**Figure 25: Top ranked vocabulary items by teacher difficulty (the top 27 includes all vocabulary items with a score of 0.33 or more)**Items outside the tool’s prediction of the top 40 items are highlighted in grey

For Texts 5 and 6 we can also learn from the **tool’s** top ranked words by difficulty, which for both texts include a number of words with American English spellings (Text 5 is an article from an American news site, Text 6 is an American company report): ‘*neighbors’, ‘honoring’, ‘canceled’*. These words had very low frequency in the BNC, it being a corpus of British text, and thus are effectively perceived as ‘more difficult’ by our tool.

**Improvements**

Using a corpus such as the International Corpus of English, which contains English texts from a wide variety of countries (The ICE Project, 2016) would enable a more accurate scoring of non-British texts.

With data for different texts suggesting different utility of the L1 similarity feature when it comes to predicting the most difficult word, it is clear that it would be useful to do further research with French students themselves to ascertain their perceived difficulty of such words and further probe teachers as to why they rated certain words that seem to be highly similar in the French, as ‘difficult’.

# VI: APPLICATIONS & EXTENSIONS

Many possible applications exist for a tool such as ours, some of which some are discussed in this section:

* **Web Tool:** Having noted the existence of tools such as Vocabulary.com’s *VocabGrabber* and Cambridge’s *TextInspector,* the addition of a user-friendly web interface to our tool could provide similar functionality for teachers and learners to input raw text and receive vocabulary items as output. Our tool provides enhanced MWE identification compared to others, and the ability to choose a number of sort / filter options including native language similarity would be another unique element.
* **Integration into existing e-learning platforms and article subscription services:** ESL e-learning platforms often integrate reading texts based on current news articles, including the forementioned Market Leader Premier Lessons and One Stop English. Our tool could aid in the process of extracting vocabulary items from these reading texts and help to automate content creation.
* **Exercise Generation:** There is a clear case for the automated generation of exercises for teachers to give learners (or learners to use as self-study tools), to practice and learn the top-ranking ‘difficult’ vocabulary items extracted from the text input. This would in effect provide a completely automated alternative to online services such as Market Leader and One Stop English, and be an extremely useful tool for teachers, who would no longer need to write their own exercises or source them from elsewhere, but simply input a text and receive a complete worksheet with a vocabulary list and exercises ready for use in lessons.   
  + **The problem of word sense disambiguation:** The key stumbling block to the above is the problem of word sense disambiguation, i.e. the fact that the same word can have multiple different meanings. Considering the MWE *break down* in text 4, this word has four possible definitions: *to separate into smaller parts*, *to stop working* or *to fail because of a problem* or disagreement (Cambridge Dictionary, 2018). To create, for example, a simple ‘match the word to its definition’ exercise, we must detect which definition is being used. This problem is multiplied if we wish to create cloze exercises (a popular exercise type in ESL, also known as ‘gap fill’) as we must be sure of the word’s usage in both the external sentence and within the input text. Further work on word sense disambiguation, perhaps using NLTK’s Lesk Algorithm, would therefore be necessary for this step.
* **Overall Text Readability / Difficulty Assessment:** As discussed in Chapter II Part B, much work has already been carried out in the field of NLP on predicting the overall difficulty level of a text. However, we could propose the utility of our tool in predicting how difficult a given text is likely to be for a reader of a particular L1, rather than for a generalised reader base.

# VII: CONCLUSION

In terms of vocabulary item extraction, our tool successfully extracts 97% of single words and over 50% of MWEs (over 60% of ‘difficult’ MWEs) from our sample texts. This is a favourable performance compared to other ESL text extraction tools. Some adjustments to the MWE lexicon, along with adapting the algorithm to enable extraction of non-contiguous MWEs, would improve it further.

We have provided a viable means of difficulty-scoring vocabulary items in a text, with a frequency measure proving to be the most significant predictive feature for teacher-perceived word difficulty. Information from domain experts and running a linear regression on our sample data suggests that L1 similarity measured via the Levenshtein Edit Distance between an English word and its French translation, could account for 18% of the overall difficulty level of a vocabulary item. Due to the sparse nature of our dataset, and varying results from using L1 similarity as a predictive feature for the difficulty of individual texts, more research and data gathering is required to ascertain the full utility of L1 similarity for our purposes, and we would suggest that carrying out research with learners themselves would be a useful next step.

There are a number of potential applications for our project, which could usefully serve as a web tool to aid teachers in identifying vocabulary items within texts for use in lessons: our feature scores would provide multiple possibilities for filtering and sorting vocabulary and predicting difficulty. Extension to provide automated exercise-generation from vocabulary items would increase the value of such a tool further, both for teachers and learners.

# REFERENCES

1. Aluisio, L., Gasperin, C., and Scarton, C. (2010). *Readability assessment for text simplification*. In Proceedings of the NAACL HLT 2010 Fifth Workshop on Innovative Use of NLP for Building Educational Applications (IUNLPBEA '10). Association for Computational Linguistics, Stroudsburg, PA, USA, 1-9.
2. Baldwin, T. and Kim, S. (2010). *Multiword Expressions* in *Handbook of Natural Language Processing*, 2nd edition. CRC Press, Taylor and Francis Group, pages 267–29
3. Bird, S., Klein, E., Loper, E. (2014). ‘Natural Language Processing with Python’ [ONLINE] <https://www.nltk.org/book/ch02> Accessed: 01/08/2018
4. Breland, H. (1996). *‘Word Frequency and Word Difficulty: A Comparison of Counts in Four Corpora*’. Psychological Science, 7(2), 96-99. [ONLINE] Retrieved from <http://www.jstor.org/stable/40062917>
5. British Council (2011). *Stages of a reading lesson.* [ONLINE] <https://www.teachingenglish.org.uk/sites/teacheng/files/Hard%20Times_nothing%20but%20facts_worksheet%202.pdf> Accessed: 10/8/2018.
6. Buchmeier, M. / Wiktionary, (2018). *Bilingual Dictionaries for Offline Use*. [ONLINE] <https://en.wiktionary.org/wiki/User:Matthias_Buchmeier#English-French> Accessed: 5/7/2018
7. Cambridge University Press / Cambridge English Language Assessment (2015). *English Vocabulary Profile*. [ONLINE]. Available at: <http://www.englishprofile.org/wordlists> [Accessed: 20 July 2018]
8. Cambridge, (2018). *Cambridge Dictionary.* [ONLINE] <https://dictionary.cambridge.org> Accessed: 20/08/2018
9. Chen, X., and Meurers, D. (2018) *Word frequency and readability: Predicting the text‐level readability with a lexical‐level attribute*. Journal of Research in Reading, 41: 486–510. <https://doi.org/10.1111/1467-9817.12121>.
10. Constant, M., Eryiğit, G, Monti, J, Van Der Plas, L, Ramsich, C, Rosner, M, Todirascu, A (2017). *Multiword Expression Processing: A Survey.* Computational Linguistics 2017 43:4, 837-892
11. Curto, Mamede & Baptista (2015). *Automatic Text Difficulty Classifier - Assisting the Selection Of Adequate Reading Materials For European Portuguese Teaching.* Conference: Conference: 7th International Conference on Computer Supported Education.
12. Dudley-Evans, Tony (1998), *‘An Overview of ESP in the 1990s’,* The Japan Conference on English for Specific Purposes Proceedings, Aizuwakamatsu. November 8, 1997. Available at: <https://files.eric.ed.gov/fulltext/ED424775.pdf> [Accessed: 26 July 2018]
13. Ellis, R (2003). *‘Task-based language learning and teaching’*, Oxford: Oxford University Press
14. Financial Times / Pearson (nd). *Market Leader Premier Lessons,* Pearson. Available at: [www.market-leader.net/news.asp](http://www.market-leader.net/news.asp) [Accessed 17 August 2018]
15. Harries, E (2013). How English has been shaped by French and other languages  
    In: Language Update, Volume 9, Number 4, 2013. [ONLINE] <http://www.btb.termiumplus.gc.ca/tpv2guides/guides/favart/index-fra.html?lang=fra&lettr=indx_autr8O1_IvjM70bY&page=9PK3TnrHndLI.html>
16. Hedge, P (2000). *‘Teaching and Learning in the Language Classroom’*, Oxford Handbooks for Language Teachers, Oxford: Oxford University Press, p67
17. Heilman, M., Collins-Thompson, K., and Eskenazi, M. (2008). *An analysis of statistical models and features for reading difficulty prediction.* In Proceedings of the Third Workshop on Innovative Use of NLP for Building Educational Applications (EANL '08). Association for Computational Linguistics, Stroudsburg, PA, USA, 71-79.
18. internationalteflacademy.com (nd) How large is the job market for English teachers abroad?. [ONLINE] Available at: <https://www.internationalteflacademy.com/faq/‌bid/102201/‌‌‌How-large-is-the-job-market-for-English-teachers-abroad>. [Accessed 26 July 2018]
19. The ICE Project (2016). *International Corpus of English*. [ONLINE] <http://ice-corpora.net/ice/>
20. Kerr, P (2016)*. ‘Personalization of language learning through adaptive technology: Part of the Cambridge Papers in ELT series.’* [pdf] Cambridge: Cambridge University Press
21. Kosterina, Anna P. (2007). ‘*The influence of the grammatical structure of L1 on learners' L2 development and transfer patterns in ESL academic writing: a comparative study (a case of Chinese and Czech speakers)’*. Retrospective Theses & Dissertations. 15086. <https://lib.dr.iastate.edu/rtd/15086>
22. Losnegaard, G., Sangati, F., Parra Escartín, C., Savary, A., Bargmann, S. et al (2016). *PARSEME Survey on MWE Resources*. 9th International Conference on Language Resources and Evaluation (LREC 2016), May 2016, Portorož, Slovenia. LREC 2016 Proceedings, pp.2299-2306
23. NLTK (2018). ‘*NLTK 3.3 Documentation’.* [ONLINE] Natural Language Toolkit <https://www.nltk.org> Accessed: 18/8/2018
24. Pypi.org (2014) ‘Levenshtein: Python extension for computing string edit distances and similarities’. [ONLINE] <https://pypi.org/project/python-Levenshtein/> Accessed: 18/18/2018
25. Rodgers, T. (2009). *On measuring vocabulary difficulty. An analysis of item variables in learning Russian-English vocabulary pairs*. IRAL - International Review of Applied Linguistics in Language Teaching, 7(4), pp. 327-344. Retrieved 17 Aug. 2018, from doi:10.1515/iral.1969.7.4.327
26. Sabas, A. (2014)*. ‘Selecting good features – Part II: linear models and regularization’* in Diving Into Data [ONLINE] <https://blog.datadive.net/selecting-good-features-part-ii-linear-models-and-regularization/> Accessed: 17/8/2018
27. Sag, I. A., Baldwin, T., Bond, F., Copestake, A., & Flickinger, D. (2002). *Multiword expressions: A pain in the neck for NLP*. Proceedings of Computational Linguistics and Intelligent Text Processing: Third International Conference: CICLing-2002, Lecture Notes in Computer Science, 2276, 1-15.
28. Sanmuganathan, K. (2014). *Impact of L1 on ESL (English as a Second Language) Writings of the Undergraduates of University of Jaffna*. Proceedings of Jaffna University International Research Conference (JUICE-2012), pp. 171181, published: March 2014, Sri Lanka
29. Schneider et al, (2014) ‘*English Multiword Expression Lexicons EnWikt*’, 2014 [ONLINE] <http://www.cs.cmu.edu/~ark/LexSem/mwelex/README.md>
30. Schneider, N., Onuffer, S., Kazour, N., Danchik, E., Mordowanec, H., Smith, N. (2014). *‘Comprehensive Annotation of Multiword Expressions in a Social Web Corpus’.* In Proceedings of the 9th Linguistic Resources and Evaluation Conference, Reykjavík, Iceland, May 2014.
31. Scikit-learn, (2017). ‘SkLearn.linear\_model. LinearRegression’ [ONLINE] <http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html> Accessed: 18/18/2018
32. Spencer, K. (2002). *English spelling and its contribution to illiteracy: word difficulty for common English words*. [ONLINE] Reading, 36: 16-25. doi:[10.1111/1467-9345.00180](https://doi.org/10.1111/1467-9345.00180)
33. Springer Nature Limited (2000-2018). *One Stop English,* Macmillan Education.[ONLINE] Available at:<http://www.onestopenglish.com/> [Accessed 17 August 2018]
34. StatSoft, Inc. (2013). Electronic Statistics Textbook. Tulsa, OK: StatSoft. [ONLINE]: http://www.statsoft.com/Textbook/StatisticsGlossary/P/button/p#Pearson%20Correlation
35. The British Council / Paul Howson (2013). *The English Effect.* [ONLINE] Available at: <https://www.britishcouncil.org/sites/default/files/english-effect-report-v2.pdf> [Accessed: 26 July 2018]
36. The British National Corpus*,* version 3 (BNC XML Edition). (2007). Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: [ONLINE] <http://www.natcorp.ox.ac.uk/>
37. ThinkMap Inc (1998-2018). *Visual Thesaurus Vocab Grabber,* Visual Thesaurus. [ONLINE] Available at: <https://www.visualthesaurus.com/vocabgrabber/> [Accessed: 20 July 2018]
38. Tremblay A., Broersma, M., Coughlin C., Choi, J. (2016). *‘Effects of the Native Language on the Learning of Fundamental Frequency in Second-Language Speech Segmentation’.* Frontiers in Psychology 7 985.
39. Trilling, B. & Fadel, C. (2009). *‘21st Century Skills’*, San Francisco: Wiley.
40. Twinword (2018). *Twinword Language Scoring API.* [ONLINE] Available at: [www.twinword.com/api/language-scoring.php](http://www.twinword.com/api/language-scoring.php) [Accessed: 20 July 2018]
41. Vocabulary.com (1998-2018). *Vocab Grabber,*[ONLINE]Available at [www.vocabulary.com/lists/vocabgrabber](http://www.vocabulary.com/lists/vocabgrabber) [Accessed: 20 July 2018]
42. West, M. (1953). *A General Service List of English Words*. London: Longman, Green and Co.
43. Wood, D. (2004). *An empirical investigation into the facilitating role of automatized lexical phrases in second language fluency development*. Journal of Language and Learning. 2. 27-50.
44. Yoon S., Bhat, S., and Zechner, K. (2012). *Vocabulary profile as a measure of vocabulary sophistication*. In Proceedings of the Seventh Workshop on Building Educational Applications Using NLP (NAACL HLT '12). Association for Computational Linguistics, Stroudsburg, PA, USA, 180-189.

# APPENDIX A: SAMPLE TEXTS

|  |  |  |
| --- | --- | --- |
| **Text** | **Reference & Link** | **Text Extract & Word Count** |
| Text 1:   Internal Email (SparkToro) | SparkToro, 2012. ‘7 Ways We Try To Make Internal Emails Better: *[A tough email]*’ [ONLINE] <https://sparktoro.com/blog/7-ways-we-try-to-make-internal-emails-better>.  Direct link to text [Accessed: 3/3/2018] :  [https://images.sparktoro‌.com/blog/wp content/‌uploads/‌2017/10/mozzy‌-email-rand.gif](https://images.sparktoro‌.com/blog/wp%20content/‌uploads/‌2017/10/mozzy-email-rand.gif) | *… If you've got questions or want to chat more feel free to grab me, I'll be in the office for the remainder of the week and Monday of next week.…*  **Overall Word Count:** 472 |
| Text 2:  Newspaper Article (Renewable Energy) | The Guardian, 2018. ‘More than 100 cities now mostly powered by renewable energy, data shows’ [ONLINE] Accessed: 3/3/2018  Direct link to text [Accessed: 3/3/2018] :  [https://www.theguardian.com/cities/2018/‌feb/27/cities-powered-clean-energy-renewable](https://www.theguardian.com/cities/2018/feb/27/cities-powered-clean-energy-renewable) | *… The number of cities reporting they are predominantly powered by clean energy has more than doubled since 2015, as momentum builds for cities around the world to switch from fossil fuels to renewable sources…*  **Overall Word Count:** 518 |
| Text 3:  Company Report (Coca Cola) | CocaCola, 2016. ‘16 Coca-Cola Stories That Made Headlines in 2016’ [ONLINE]  Direct link to text [Accessed: 3/3/2018] :<https://www.coca-colacompany.com/‌stories/16-coca-cola-news-headlines-from-2016> | *… Marking a significant shift in its marketing strategy, Coca-Cola announced in January 2016 that for the first time, all Coke Trademark brands will be united in one global creative campaign: “Taste the Feeling.” …*  **Overall Word Count:** 479  (note: only an extract comprising the first 479 words of the report was used) |
| Text 4:  Sales Email (Earn1K) | Sumo.com [nd], ‘Sales Email Template: Today I’m Opening Earn1K’ [ONLINE]  Direct link to text [Accessed: 13/6/2018] : <https://sumo.com/stories/sales-email-templates> Alternative Repository: <goo.gl/zd8dxT> | *… “How do I join if I don’t have that kind of money lying around right now?”*  *If you have credit card debt, do not join – I’ll cancel your membership and ban you for life. Start by checking out my book from the library for free instead.….*  **Overall Word Count:** 777 |
| Text 5:  Newspaper Article (AirBNB) | Huffington Post, 2018. ‘New Rental Law Forces Airbnb To Cancel Thousands Of Reservations In Japan’ [ONLINE]  Direct link to text [Accessed: 13/6/2018] : [https://www.huffingtonpost.com/entry/thousands-airbnb-reservations-canceled-japan\_us\_‌5b1a5115e4b0‌9d7a3d711049](https://www.huffingtonpost.com/entry/thousands-airbnb-reservations-canceled-japan_us_5b1a5115e4b09d7a3d711049) | *… Airbnb said on Thursday it’s been forced to cancel thousands of bookings in Japan after the government unexpectedly announced the immediate implementation of a law that regulates home sharing…*  **Overall Word Count:** 426 |
| Text 5:  Company Report (Johnson & Johnson) | Johnson & Johnson, 2018. ‘A Look Back on Our Work in 2017’ [ONLINE]  Direct link to text [Accessed: 13/6/2018] : <https://www.jnj.com/2017-year-in-review> | *… In 2016, Johnson & Johnson launched a new set of five-year objectives, known as the Health for Humanity 2020 Goals, which include providing better access to healthcare worldwide, making the places we live and work in healthier, and teaming up with partners …*  **Overall Word Count:** 467 (note: only an extract comprising the first 467 words of the report was used) |

# APPENDIX B: TEACHER RESEARCH DOCUMENT

***Note****: The following is the document for the first round of research (Texts 1 – 3) given to participating teachers. The document used for the second round of research (Text 4 – 6) was identical, but for the links to the texts themselves. Teachers were not asked to re-state the secondary information they had already filled in during round one.*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Page 1/3**  **RESEARCH EXERCISE: ESL TEACHERS’ CREATION OF VOCABULARY LISTS FROM TEXTS**  **Consent**  This is part of a project by conducted by Rowena Jones, a candidate for the MSc in Computing & Information Systems at Queen Mary University of London.You are invited to participate in this research because you are a teacher of English as a Second Language to adults. However, your participation is voluntary. You may choose not to participate.  The following two pages ask you to complete some information about yourself, read three short texts and produce vocabulary lists from them, and to answer some other associated questions. It is expected that all of this should take you a maximum of 60 minutes. Your responses will be anonymised and your name and contact details will not be used.  The results of this research will be used for scholarly purposes only and may be shared with Queen Mary University of London representatives.  If you have any questions about the research, please contact Rowena Jones (see contact information below).  By completing this exercise, you agree that you have read the above information, that you voluntarily agree to participate, that you are at least 18 years of age and that your responses may be used in the project.  **Instructions on the completion of this document**  Please either:   1. complete this document electronically (type your answers into this document), save it and email it as an attachment to [rowenajones1@gmail.com](mailto:rowenajones1@gmail.com). The formatting of the document is unimportant – don’t worry about presentation as long as your answers are clear. 2. print the document and write in your answers using any additional pieces of paper as needed. You may, if you wish, print the three texts and highlight your selection of vocabulary items directly on them. Scan the document/s and email as an attachment to [rowenajones1@gmail.com](mailto:rowenajones1@gmail.com) | **Page 2/3**  **Part 1: Complete this information about yourself.**  1a.How many years’ experience do you have teaching English as a foreign language to adults?  1b. What is the native language of the students you normally teach?  When carrying out the following exercise, please do so imagining that it is for an adult student of the native language stated above.  \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_  **Part 2:** **Read texts & produce vocabulary lists**  Imagine that one of your adult students wishes to use a particular text in their English lesson. They have identified this text as of interest or use to them, i.e. it is on a topic (or from a context) relevant to their job or interests. They wish to use it as a basis for increasing their vocabulary.  **Please read each text linked below and identify the vocabulary from each that you would select to ‘teach’ to the student, that is to say vocabulary items you think the student would not be likely to already have a reasonable understanding of.**  Consider each text separately (i.e. they could be brought by three different students on three different occasions, they wouldn’t be used the same lesson).  Link to text 1 [Link to text 2](https://www.theguardian.com/cities/2018/feb/27/cities-powered-clean-energy-renewable) [Link to text 3](http://www.coca-colacompany.com/stories/16-coca-cola-news-headlines-from-2016)  **Further information**   * Assume that the student has a minimum of a B1 CEFR level * You may, though you are not obliged to, differentiate between vocabulary that you would expect to have to teach to a B1/B2 student but not a C1/C2 level student. Do this by marking the **less** difficult vocabulary from your selection with a \* * You may include a mix of single words and multi-word expressions. * You may identify as many vocabulary items as you wish, but please identify a minimum of 10 and a maximum of 30 (per text)  |  |  |  |  | | --- | --- | --- | --- | | ***Example vocabulary list (for an imaginary text)*** | **Text 1: Email** | **Text 2: Renewable Energy Article** | **Text 3: Coca Cola Report** | | *get to grips with*  *scholarship\**  *phase out\**  *messy\**  *cull*  *forthwith\**  *scramble\**  *put into perspective\**  *drought*  *o badger somebody*  *[CUT]* |  |  |  | |
| **Page 3/3**  **Part 3:** **Comments on the exercise (optional)**  Please write any comments you have on completing the exercise here, e.g.:   * Caveats or qualifications to your selection of vocabulary ⦁ Any difficulties or uncertainties you had in completing the task * An outline of how you made the selection of vocabulary (did you go on instinct, experience, teaching theory, knowledge of the student’s native language, something else…?)   **Part 4:** **Please answer the following questions:**  4a. Have you taught English as a foreign language in a country where your students had a different native language to your current students? 4b. If yes, which language did they speak?  4c. If yes, do you think **a student of comparable level in that country** would find the same vocabulary difficult as your current students, or would your selection be different? Please explain briefly.  **Part 5: As an ESL teacher, which of the following tools (if any) would you be interested in using? Please rank in order of interest starting from 1 (most interesting) and stopping when you feel no more options would be interesting.**  \_\_\_ A tool which reads a text and gives it an overall difficulty rating in terms of vocabulary  \_\_\_ A tool which reads a text and automatically generates exercises based on vocabulary in the text (e.g. match   word to the definition, gap fill exercises etc)  \_\_\_ A tool which reads a text and gives it an overall difficulty rating in terms of grammar  \_\_\_ A tool which reads a text and extracts / states different grammar structures present in the text  \_\_\_ A tool which reads a text and automatically generates exercises based on grammar in the text (e.g. identify the   tense, conjugate the verb, fill in the preposition etc)  \_\_\_ (Optional) Any other ideas of your own: ………………………………………………………...………………………………………………  **Thank you very much for participating in this research exercise.** | |

# APPENDIX C: TEACHER RESEARCH RESULTS & RESULTING DIFFICULTY SCORES

The results for Text 1 are shown in full. The items included by each teacher in their vocabulary lists are marked with a 1. The average score for a given difficulty item is calculated as the overall score to that item. Any vocabulary items not included in the list below were not selected by any teachers, and will thus have a difficulty score of 0. Abridged results for the five other texts follow.

**Text 1: Text 2: Text 3:**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **Vocab Item** | **Score** | **T1** | **T2** | **T3** | **T4** | **T5** | **T6** |  | **Item** | **Score** |  | **Item** | **Score** |
| a few weeks back | 0.17 |  |  | 1 |  |  |  | ambition | 0.17 | achieved | 0.17 |
| adamant | 0.50 |  |  | 1 | 1 |  | 1 | among | 0.17 | ahead of schedule | 0.83 |
| board | 0.33 | 1 |  |  |  | 1 |  | as well as | 0.17 | announced | 0.17 |
| booster projects | 0.50 | 1 |  | 1 | 1 |  |  | attributed | 0.16 | appeal | 0.67 |
| bullish | 0.83 | 1 | 1 | 1 |  | 1 | 1 | biomass | 0.17 | aspirational | 0.50 |
| concern | 0.33 |  | 1 |  |  |  | 1 | boosted | 0.50 | beverage | 0.33 |
| conservatively | 0.33 |  |  | 1 |  |  | 1 | both | 0.17 | bold | 1.00 |
| crazy excited for | 0.17 |  |  | 1 |  |  |  | changing tide | 1.00 | bottling | 0.50 |
| estimates | 0.17 |  |  | 1 |  |  |  | commitment | 0.33 | brand | 0.17 |
| extraordinary | 0.17 | 1 |  |  |  |  |  | comprehensive | 0.33 | campaign | 0.17 |
| fair-weather friend | 1.00 | 1 | 1 | 1 | 1 | 1 | 1 | covenant | 0.67 | commitment | 0.33 |
| fully | 0.17 |  |  | 1 |  |  |  | disparate | 0.16 | diet | 0.17 |
| funding | 0.33 |  |  | 1 |  | 1 |  | drive | 0.50 | emerging | 0.50 |
| get to focus on | 0.17 |  |  | 1 |  |  |  | entirely | 0.17 | equity | 0.17 |
| given the circumstances | 0.17 |  |  | 1 |  |  |  | evidence | 0.17 | equivalent | 0.17 |
| grab | 0.83 | 1 | 1 | 1 |  | 1 | 1 | follow suit | 0.83 | expanding | 0.17 |
| growth | 0.17 | 1 |  |  |  |  |  | fossil fuels | 0.33 | flagship | 0.83 |
| howdy | 0.50 |  | 1 |  | 1 |  | 1 | fossil-free | 0.17 | generate | 0.17 |
| in a row | 0.67 | 1 |  | 1 |  | 1 | 1 | fully | 0.33 | headlines | 0.83 |
| launch | 0.17 |  |  |  |  | 1 |  | geothermal | 0.17 | higher-margin | 0.33 |
| letter of intent | 0.33 |  | 1 |  |  |  | 1 | held up | 1.00 | higher-return | 0.17 |
| long term | 0.17 | 1 |  |  |  |  |  | in the wake of | 0.83 | lie ahead | 0.83 |
| make up | 0.83 | 1 |  | 1 | 1 | 1 | 1 | instigated | 0.50 | lifestyle | 0.17 |
| might have | 0.17 |  |  |  |  |  | 1 | lay claim | 0.83 | market | 0.33 |
| off target | 0.67 |  | 1 | 1 | 1 | 1 |  | local authority | 0.33 | milestones | 1.00 |
| on track | 0.83 | 1 | 1 | 1 | 1 | 1 |  | momentum | 0.33 | noteworthy | 0.83 |
| please | 0.17 |  |  |  | 1 |  |  | move away from | 0.33 | pipeline | 0.17 |
| press on | 0.50 | 1 |  | 1 |  |  | 1 | not-for-profit | 0.33 | range | 0.50 |
| profitable | 0.17 |  |  | 1 |  |  |  | on its books | 0.67 | refranchising | 0.17 |
| pull out of | 0.33 |  |  |  | 1 | 1 |  | powered | 0.17 | replenish | 1.00 |
| pushed for | 0.17 |  |  | 1 |  |  |  | predominantly | 0.50 | retail sales | 0.33 |
| rather than | 0.17 |  |  |  | 1 |  |  | reassuringly | 0.17 | shift | 0.83 |
| re-engage in | 0.33 |  |  | 1 |  |  | 1 | recruits | 0.17 | solid | 0.33 |
| relieved | 0.17 |  |  |  |  |  | 1 | renewable | 0.33 | succeed | 0.67 |
| remainder | 0.50 | 1 |  | 1 |  |  | 1 | set a target | 0.17 | suits | 0.67 |
| round | 0.50 | 1 |  | 1 |  |  | 1 | shift | 1.00 | sustainability | 0.50 |
| run fire drills | 0.83 | 1 | 1 | 1 | 1 |  | 1 | sign up | 1.00 | trademark | 0.33 |
| sign off on | 1.00 | 1 | 1 | 1 | 1 | 1 | 1 | sourced | 0.50 | underscores | 0.83 |
| softness | 0.83 | 1 | 1 | 1 | 1 |  | 1 | spur on | 1.00 | unveiled | 1.00 |
| sole | 0.67 | 1 | 1 | 1 | 1 |  |  | states | 0.17 | whichever | 0.33 |
| startups | 0.17 |  |  |  | 1 |  |  | supply | 0.17 |  |  |
| suitors | 0.83 | 1 | 1 | 1 | 1 |  | 1 | switch from | 0.83 |  |  |
| throughout | 0.33 |  |  | 1 | 1 |  |  | top the list | 0.50 |  |  |
| upside | 1.00 | 1 | 1 | 1 | 1 | 1 | 1 | vast majority | 0.17 |  |  |
| whether | 0.33 |  | 1 |  |  |  | 1 | widespread | 0.67 |  |  |
|  |  |  |  |  |  |  |  | with regards to | 0.17 |  |  |
|  |  |  |  |  |  |  |  | withdraw | 1.00 |  |  |
|  |  |  |  |  |  |  |  | zero-carbon | 0.17 |  |  |
|  |  |  |  |  |  |  |  |  |  |  |  |

**T = Teacher**

**Text 4: Text 5: Text 6:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **Item** | **Score** |  | **Item** | **Score** |  | **Item** | **Score** |
| ban | 0.83 | (tourist) hotstpot | 0.33 | (across the) globe | 0.33 |
| beyond | 0.17 | affordable | 0.67 | (parental) leave (benefits) | 0.67 |
| break down | 1.00 | along with | 0.33 | (purpose) driven | 1.00 |
| built in | 0.17 | apply for | 0.17 | (social) venture | 0.83 |
| check ins | 0.50 | chaos | 0.33 | acquisition | 0.50 |
| check out | 0.50 | cover costs | 0.50 | address [verb] | 0.33 |
| convenient | 0.17 | cumbersome | 1.00 | best-in-class | 0.17 |
| craft | 0.83 | deal with | 0.67 | bestow | 1.00 |
| delighted | 0.17 | due to | 0.33 | citizenship | 0.17 |
| eat the fees | 0.50 | force to | 0.33 | coalition | 0.33 |
| enrolment | 0.83 | fund | 0.67 | commitment | 0.33 |
| ensure | 0.17 | go through | 0.33 | committed to | 0.33 |
| expand the pie | 1.00 | going forward | 0.67 | credo | 0.33 |
| focus on | 0.17 | growth | 0.33 | date back to | 0.33 |
| follow up | 0.50 | hassle | 0.83 | debut | 0.17 |
| freelancing | 0.33 | home sharing | 0.50 | deployed | 0.67 |
| gather | 0.33 | host | 0.17 | enable | 0.00 |
| in depth | 0.83 | illegal activities | 0.17 | expand | 0.33 |
| land a client | 0.83 | implementation | 0.50 | hypertension | 0.17 |
| launch | 0.33 | issue | 0.33 | impactful | 0.67 |
| light at the end of the tunnel | 0.67 | jeopardize | 1.00 | kinder | 0.33 |
| lying around | 0.67 | license | 0.33 | launch | 0.50 |
| master class | 0.17 | listings | 0.67 | leading news | 0.33 |
| no-brainer | 1.00 | localities | 0.17 | make headlines | 0.83 |
| on and on | 0.50 | might | 0.17 | mission statement | 0.17 |
| on the fence | 0.83 | notify | 0.17 | paradigm shifting | 1.00 |
| on the side | 0.67 | nurture | 0.67 | pathogens | 0.17 |
| on your own time | 0.33 | out of fear | 0.17 | reshape | 0.67 |
| payment plan | 0.17 | paperwork | 0.33 | serve | 0.17 |
| productizing | 0.50 | plunge | 0.50 | set (goals) | 0.50 |
| put in the work | 0.50 | regulate | 0.17 | showcase | 0.50 |
| raise | 0.17 | rethink | 0.67 | slow down | 0.33 |
| rates | 0.17 | seek | 0.67 | sustainability | 0.17 |
| refund | 0.33 | set restrictions | 0.33 | team up with | 0.67 |
| rigorously | 0.50 | stinks | 1.00 | tenet | 0.50 |
| scale up | 1.00 | stipulate | 0.67 | toughest | 0.83 |
| side business | 0.33 | surround | 0.17 | track (progress) | 0.67 |
| spread out | 0.67 | throw into disarray | 1.00 | trajectory | 0.33 |
| start off | 0.33 | understatement | 1.00 | widely | 0.33 |
| stay on track | 0.83 | unexpectedly | 0.17 |  |  |
| step by step | 0.17 |  |  |  |  |
| tactics | 0.33 |  |  |  |  |
| walk you through | 0.67 |  |  |  |  |
| willing | 0.50 |  |  |  |  |
| work one on one | 0.17 |  |  |  |  |

# APPENDIX D: POS TAG MAPPING

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **NLTK POS Tag** | **Broader POS Tag** |  | **Eng-French Wiktionary Tag** | **Broader POS Tag** |
|  |  |  |  |
| CD | PUNC\_NUM\_OTHER | <abbr> | PUNC\_NUM\_OTHER |
| LS | PUNC\_NUM\_OTHER | <acronym> | PUNC\_NUM\_OTHER |
| SYM | PUNC\_NUM\_OTHER | <adj> | ADJ\_ADV |
| JJ | ADJ\_ADV | <adv> | ADJ\_ADV |
| JJR | ADJ\_ADV | <article> | PREPS\_DETS\_CONJ |
| JJS | ADJ\_ADV | <conj> | PREPS\_DETS\_CONJ |
| RB | ADJ\_ADV | <contraction> | PREPS\_DETS\_CONJ |
| RBR | ADJ\_ADV | <determiner> | PREPS\_DETS\_CONJ |
| RBS | ADJ\_ADV | <initialism> | PUNC\_NUM\_OTHER |
| NN | NOUN | <interj> | INTERJ\_FOREIGN |
| NNP | PRP\_NOUN | <n> | NOUN |
| NNPS | PRP\_NOUN | <num> | PUNC\_NUM\_OTHER |
| NNS | NOUN | <particle> | PREPS\_DETS\_CONJ |
| VB | VERB | <phrase> | MWE |
| VBD | VERB | <prefix> | PUNC\_NUM\_OTHER |
| VBG | VERB | <prep> | PREPS\_DETS\_CONJ |
| VBN | VERB | <pron> | PRONOUN |
| VBP | VERB | <prop> | PRP\_NOUN |
| VBZ | VERB | <proverb> | MWE |
| MD | MODAL | <suffix> | PUNC\_NUM\_OTHER |
| WDT | WH\_TH\_HOW | <symbol> | PUNC\_NUM\_OTHER |
| WP | WH\_TH\_HOW | <v> | VERB |
| WP$ | WH\_TH\_HOW |  | |
| WRB | WH\_TH\_HOW |
| CC | PREPS\_DETS\_CONJ |
| DT | PREPS\_DETS\_CONJ |
| PDT | PREPS\_DETS\_CONJ |
| IN | PREPS\_DETS\_CONJ |
| EX | PREPS\_DETS\_CONJ |
| POS | PREPS\_DETS\_CONJ |
| RP | PREPS\_DETS\_CONJ |
| TO | PREPS\_DETS\_CONJ |
| PRP | PRONOUN |
| PRP$ | PRONOUN |
| UH | INTERJ\_FOREIGN |
| FW | INTERJ\_FOREIGN |
| (MWE) | MWE |

# APPENDIX E: GUIDE TO SUPPORTING MATERIAL

**supporting\_material***A series of jupyter notebooks showing code used in the creation of objects used within the tool, and in evaluating its performance.*

* Creating a MWE lexicon.ipynb
* Creating an Eng-to-French dictionary from Wiktionary data.ipynb
* Creating frequency distributions from the BNC corpus.ipynb
* Texts 1to3 Feature Evaluation and Linear Regression.ipynb
* Texts 4to6 Difficulty Score Evaluation.ipynb
* Texts 4to6 Vocabulary Item Extraction Evaluation.ipynb
* **supporting\_material / tool***Contains .py class files for running the tool, and equivalent jupyter notebooks*
  + Analysis Helper [a class used in the evaluation of the tool]
  + FrequencyFeature
  + L1SimilarityFeature
  + LengthFeature
  + POSFeature
  + POSFrequencyFeature
  + TeacherScores [Performs a mapping of teacher scores to vocabulary item scores for sample texts]
  + TextItems [Extracts vocabulary items from a text]
  + TextScorer [Calls TextItems and the Feature classes to extract vocab items from text and give scores]
  + **supporting\_material / tool / files***Contains the sample texts in .txt form, and Excel workbooks used by / produced by the tool*
    - sample\_texts [folder of .txt files]
    - lengthscores.xlsx [mapping of lengths to scores for the LengthFeature]
    - posscores.xlsx [mapping of POS tags to scores for the POSFeature]
    - tag\_mapping.xlsx [mapping of NLTK POS tags to broad tags]
    - tag\_mapping\_forfrenchdict.xlsx [mapping of French dictionary tags to broad tags]
    - teacher\_scores.xlsx [holds teacher scores from primary research]
    - vocab\_extraction\_evaluation.xlsx [contains all vocabulary extracted from texts 1 to 4]
    - vocab\_ranking\_evaluation.xlsx [contains all scored vocabulary extracted from texts 1 to 4]
  + **supporting\_material / tool / obj***Contains a number of large dictionary and list objects used by the tool, in pickle format*
    - combined\_bnc\_freq\_dict\_lower.pkl [BNC - derived frequency distribution]
    - combined\_tagged\_bnc\_freqdict\_lower.pkl [BNC-derived POS-tagged freq distribution]
    - french\_dict\_pos.pkl [Eng-French Wikipedia dictionary, with pos tags]
    - french\_dict\_simple.pkl [Eng-French Wikipedia dictionary, without pos tags]
    - mwes\_list.pkl [List of all MWEs from our MWE lexicon]