Pearson

Research Series

Developing the GSE Vocabulary

April 2017

Dr Veronica Benigno
Prof. John de Jong

Global Scale of English
Fast-track your progress

Purpose of this document

This document is a report on an ongoing project to develop the GSE Vocabulary, a graded lexical inventory aligned to the Global Scale of English (GSE) and the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2001). This project was set up in response to the Council of Europe recommendation to the state members to create, for each regional and national language, inventories of linguistic forms known as Reference Level Descriptions (Council of Europe, 2005). The RLDs are "inventories of the linguistic realisations of general notions, acts of discourse and specific notions / lexical elements and morpho-syntactic elements" which are characteristic of each level (Council of Europe, 2005, p.5).

The GSE Vocabulary inventory aims to complement the guidance of the functional approach used in the CEFR and in the GSE Learning Objectives https://www.english.com/gse/learning_objectives by indexing and scaling the lexical exponents needed to acquire the competences described in the framework, with the ultimate goal of making language learning more efficient.

Contents

Executive summary

The GSE Vocabulary is a graded lexical inventory of general English for adult learners which indicates which and how many word meanings learners should be able to understand at different proficiency levels to successfully communicate in English. The underlying principle on which the inventory is based is efficiency: What vocabulary gives learners the highest chance of communicating with other speakers? What is the relative importance of vocabulary items to be able to participate in a general conversation?

It can be assumed that being able to talk, for example, about the weather would yield more frequent opportunities to participate in communication then being able to talk about the periodic table. Basic words about weather conditions are therefore more useful for general communication than basic names of chemical substances.

The main features of the GSE Vocabulary inventory can be summarized as follows:

- The GSE Vocabulary was created using a mixed methodology which combines corpus frequency analysis and teacher judgments on communicative usefulness.
- Each lexical entry in the GSE Vocabulary is a word meaning, not a lemma (a base word form and its inflected forms within the same part of speech) or a word family (a word and its related inflections and derived words). In this way, we are reflecting the fact that vocabulary learning takes place in context and that different meanings of polysemous words are most likely learned at different stages of proficiency.
- The GSE Vocabulary includes more than 37,000 word meanings (corresponding to about 20,000 lemmas), 80,000 collocations and 7,000 phrases.
- The database is searchable by keyword, part of speech, topic, subtopic, and proficiency level (on the CEFR and the Global Scale of English).
- It is aimed at learners, teachers, and materials designers with the purpose of helping them prioritize vocabulary.

The lemma *foot* (noun) includes the inflected forms *foot* (singular) and *feet* (plural). The word family of the lemma *foot* includes its derived words such as *footer*, *footing*, *footless*, etc. The lemma *foot* has more than one meaning in English, e.g., "the part of your body that you stand on" or "a unit for measuring length, equal to 0.3048 metres". These two meanings are expected to be learned at different proficiency levels on the GSE and CEFR, GSE 20 (<A1) and GSE 59 (B2) respectively.

Theoretical framework

A review of the studies on vocabulary acquisition, teaching and assessment (e.g., Bogaards and Laufer, 2004; Meara, 2009; Milton, 2009; Nation, 2001; Read, 2000; Schmitt, 2000; Schmitt and McCarthy, 1997) shows that although this field of investigation has evolved quite rapidly over the last few decades, there is still little agreement on which and how many words are needed to communicate efficiently at increasing proficiency levels. Attempts to relate vocabulary knowledge to proficiency levels have focused on quantitative aspects and frequency profiling methods have investigated learners' knowledge of single words (Laufer & Nation, 1995; Nation, 1990).

In fact vocabulary learning should not simply be regarded as a quantitative process (in terms of expansion of one's vocabulary size) but should also be considered from a qualitative point of view (in terms of vocabulary depth, e.g., knowledge of collocations and pragmatic rules).

Although many studies outline the importance of frequency of exposure as an objective criterion in deciding what to teach first (Ellis, 2002; Gyllstad, 2007; Nation & Beglar, 2007), frequency alone is not sufficient to identify pedagogically-relevant vocabulary. According to Widdowson (2003, p.83), "[...] prototypical prominence in the mind does not accord with frequency of actual occurrence". A purely frequency-based pedagogical list is necessarily biased by the nature of the corpus and would ignore low-frequency words which refer to basic concepts that are useful for communicative purposes but rarely spoken or written about by users of the language. As Stubbs (2002) points out, the definition of what is basic depends not only on frequency, but also on functional criteria such as communicative relevance or usefulness.

In order to address some limitations of current research, the GSE Vocabulary was created using a mixed methodology combining frequency data and teacher judgements with the aim of producing a weighted measure to identify level-appropriate vocabulary.

In recent years, many studies have shown that language is formulaic with no rigid separation between vocabulary and grammar (Ellis, 2002; Wray, 2002). Words occur most frequently in a limited number of contexts, i.e. in co-occurrence with a limited number of other words producing collocations, chunks, ready-made phrases, and fixed units.

Other research has shown that partial acquisition of a word meaning is a very common stage in language development since learners encounter and use words in a number of predictable lexical environments and gradually extend their knowledge as their proficiency increases (Wolter, 2009). The GSE Vocabulary takes into account both areas of research, providing information on which words combine with each other (collocation) and distinguishing between different word meanings. The database was developed based on existing research evidence on vocabulary learning and acquisition and aims to create a model of lexical proficiency which integrates different dimensions (size and depth) of vocabulary knowledge.

Methodology and data

The process to create the GSE Vocabulary inventory consisted of four main steps:

- Identification of word frequency through corpus analysis
- Semantic annotation of the database by topic and subtopic
- Rating of vocabulary by teachers for communicative usefulness
- Alignment of word meanings to the Global Scale of English and the CEFR

Step one. The first step used corpus analysis to produce a word frequency list from three corpora of differing size and content:

- **LCN** (Longman Corpus Network) – a corpus of 330 million words created by Pearson and used as the basis for Longman dictionaries
- **UKWAC** http://wacky.sslmit.unibo.it/doku.php?id=corpora; Baroni et al., 2009, a 2 billion word corpus constructed by crawling the web.
- **COCA** http://corpus.byu.edu/coca, a corpus of contemporary American English of 450 million words. For our study, we selected the spoken component only (about 90 million words).

LONGMAN CORPUS NETWORK

The Longman Corpus Network consists of the following corpora:

- The Longman/Lancaster English language Corpus
- The Longman Spoken Corpus
- The Longman Learners' Corpus
- The British National Corpus
- The American National Corpus

The decision was taken to combine three different corpora to compensate for the limitations of using either a traditional corpus (which is balanced but tends to be more limited in size) or a web-based corpus (which is unbalanced but a powerful due to its size, authenticity, linguistic and socio-linguistic variety, and up-to-dateness).

Our selected data sample consisted of the top 10,000 lemmas occurring in the three corpora and the majority of the entries found in the Longman Active Study Dictionary of English (an intermediate learner's dictionary), creating a total of about 20,000 lemmas. The dictionary entries were added to the corpus-based frequency list to ensure low-frequency but pedagogically useful vocabulary was included in the database.

Step two. The resulting frequency list included more than 20,000 lemmas. For each lemma, the meanings were identified based on the contents of the 37,000 Longman dictionary database. The resulting list of around 37,000 word meanings was semantically annotated on the model of the Council of Europe Vantage Specifications (van Ek & Trim, 2001), which categorize vocabulary into Specific Notions, General Notions, and Functions. A team of expert lexicographers manually annotated all word meanings by topic and sub-topic. The resulting database was thus organised around pedagogical areas with different meanings of the same word assigned to different topic areas. For example, take the three different meanings of the noun "fork": "a small tool that you use for picking up and eating food"; "a place where a road or river divides into two parts"; and "a tool used for digging and breaking up soil". The first word meaning was tagged as topic "Food and drinks", subtopic "Utensils, appliances, and tableware"; the second word meaning as topic "Holidays, travel, and transportation", subtopic "Road or rail network"; and the third word meanings as topic "Sports, hobbies, and interests", subtopic "Gardening".

The model used in the Council of Europe Specifications (Waystage, Threshold and Vantage) distinguishes between categories of language functions and notions. **Language functions** refer to what people do with language, e.g. apologizing, making a request, complaining, etc. Notions refer to the concepts that people handle when they use language and can be general and specific. **General notions** refer to abstract, relational concepts, e.g. time or space. **Specific notions** refer to more concrete vocabulary, e.g. food and drinks or health and body.

Step three. Each of the 37,000 word meanings was rated by 10 raters out of a pool of 19 English teachers using an overlapping design. Teachers were asked to rate words by meaning using a pre-defined scale based on the principle of usefulness. The scale (see table 1 below) ranges from 1 (the most essential vocabulary) to 5 (words language users would need only occasionally). In addition to the 1 to 5 scale, teachers were given the possibility to use the arbitrary value "99" when they found it impossible to rate a word meaning/phrase because they had never encountered the word/phrase before or because they could not decide between widely different ratings. However, they were asked to use this value sparingly. Teachers were provided with a written briefing and received online training.

RATING SCALE	
Essential	"Essential" items are the words/phrases that learners would want to acquire first. They are essential for basic communication.
Important	"Important" items are words/phrases that become necessary at a next stage; they are still very common. They are perhaps a little more detailed or a little more specific in their meaning.
Useful	"Useful" items are words/phrases that expand the user's vocabulary enabling more detailed and specific language use.
Nice to have	"Nice to have" items are for users to express themselves accurately and precisely.
Extra	The "extra" category is for items that some language users will use occasionally, but they are not needed for everyday communication.

Step four. Information on the frequency of words (i.e. lemmas; based on the corpus) and on the usefulness of words (i.e. meanings; as rated by the teachers) was combined to produce a weighted value to rank vocabulary. A total of 372,265 teacher ratings were analysed by means of descriptive statistics in order to identify deviant ratings in the data set. After removal of 19,784 deviant ratings (5.4%), the remaining data set contained 347.033 (94.6%) ratings. Removal of deviant ratings reduced variance among raters and improved the average correlation between the individual raters and the average over all raters from 0.77 to 0.84. After the data cleaning, teacher ratings and frequency values of each word meaning were combined to obtain a ranking. In combining the frequency and the rating data, the rating data was considered primary, unless the degree of agreement between the raters was too low, in which case the frequency information was given relatively more weight. In a following step, a relation was sought between the combined measure and findings from current research evidence on vocabulary size (e.g., Hazenberg & Hulstijn, 1996; Nation & Beglar, 2007) thereby linking word meanings to the Global Scale of English and the CEFR. Thus the approach used to align word meanings to the CEFR and the GSE is based on existing research about the vocabulary size needed at increasing levels of proficiency. Since the available research on vocabulary size mainly makes assumptions concerning the amount of vocabulary needed to understand and not to produce language, the GSE values and CEFR levels assigned to each word meaning refer to receptive knowledge.

Final considerations: What GSE values mean

The dominant approaches to vocabulary teaching and assessment research have used corpus frequency as the main selecting criterion to produce ordered lists of single words to serve as teaching or assessment targets. The GSE Vocabulary is the first large-scale project to combine frequency data with qualitative ratings of vocabulary usefulness. The project is part of a wider initiative by Pearson to align learning, teaching, and assessment content and provides a clear description of the intended vocabulary goals for adult learners of general English.

The GSE values assigned to each word meaning are based on a probabilistic model and should by no means be interpreted as prescriptive. The GSE values ought to be used as an indicative value of the stage at which a particular word meaning is likely to be useful in order to communicate efficiently in English. The GSE values refer to the relative importance of vocabulary in language: for example, they are helpful in prioritizing vocabulary items within the same topic. They are meant to inform and to guide content creation and help practitioners find realistic teaching and assessment targets for vocabulary.

Knowing a word receptively at a given GSE level means having 50% probability of being able to understand it. A learner at 25 on the GSE has 50% probability of understanding a vocabulary item which is at that level of difficulty (25), whereas they have a lower probability of understanding words which are higher on the GSE and a higher probability of understanding words which are lower on the GSE. It is generally accepted that receptive knowledge is larger than productive knowledge, although the threshold at which receptive vocabulary becomes productive vocabulary is not clear. It can therefore be expected that a word which is known receptively will gradually be known productively – provided that it has been encountered a number of times in a number of meaningful contexts. It is also important to keep in mind that at the very low levels, the distinction between receptive and productive vocabulary is less pronounced.

The extent to which vocabulary is selected within or above a targeted GSE range will depend a great deal on the intended purpose. Below we would like to offer some teaching recommendations that we think are also relevant for content creation and for setting assessment targets.

If the teaching purpose is to help a learner at a given ability understand vocabulary which is appropriate at his/her level, then we recommend that word meanings are mainly selected within and below the targeted range. However, teachers should allow themselves some flexibility to choose a small percentage of vocabulary above the targeted range, for example to ensure that the text that the learner is presented with is coherent and authentic.

However, if the teaching purpose is to help a learner at a given ability level to produce vocabulary which is appropriate at his/her level, then we recommend that teachers mainly select word meanings within the targeted range, preferably starting from the bottom of the range. Teachers should think in terms of relative importance of vocabulary: words at the bottom of the range will have a higher chance of being produced by learners than words in the middle or at the top of the range (which will be less frequent or useful in language communication).

The GSE Vocabulary database is freely available at https://www.english.com/gse/teacher-toolkit**. Users can search vocabulary items by keyword, part of speech, topic/subtopic, CEFR level/GSE range and have access to more than 80,000 collocations and 7,000 phrases.**

References

Baroni M., Bernardini S., Ferraresi A., & Zanchetta E. (2009). *The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora*. In Language Resources and Evaluation 43 (3), 209-226

Bogaards P. & Laufer B. (eds) (2004). *Vocabulary in a Second Language*. John Benjamins, Amsterdam

Council of Europe (2001). *The Common European Framework of Reference for Languages: learning, teaching, assessment.* Cambridge University Press, Cambridge, UK

Council of Europe (2005). *Reference Level Descriptions for National and Regional Languages (RLD). Guide for the production of RLD: Version 2. Retrieved* at https://www.coe.int/t/dg4/linguistic/Source/ DNR_Guide_EN.pdf

Ellis N. (2002). *Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition*. In Studies in Second Language Acquisition 24, 143–188

Gyllstad H. (2007). *Testing English collocations* (Unpublished doctoral dissertation). Lund University, Lund

Hazenberg S. & Hulstijn J. H. (1996). *Defining a minimal receptive second language vocabulary for non-native university students: An empirical investigation. In Applied Linguistics* 17(2), 145-163

Laufer B. & Nation I.S.P. (1995). *Vocabulary size and use: Lexical richness in L2 written production*. In Applied Linguistics, 16, 307-322

Meara P. (2009). *Connected Words: Word Associations and Second Language Vocabulary Acquisition*. John Benjamins, Amsterdam

Milton J. (2009). *Measuring second language vocabulary acquisition*. Multilingual Matters, Bristol

Nation I.S.P. (1990). *Teaching and learning vocabulary*. Rowley, MA, Newbury House

Nation I.S.P. (2001). *Learning vocabulary in another language.* Cambridge University Press, Cambridge, UK

Nation I.S.P. & Beglar D. (2007). *A vocabulary size test.* The Language Teacher, 31(7), 9-13

Pearson (2000). *Longman Active Study Dictionary of English* (LASDE). Harlow, UK

Read J. (2000). *Assessing vocabulary*. Cambridge University Press, Cambridge, UK

Schmitt N. & McCarthy M. (eds) (1997). *Vocabulary: Description, Acquisition, and Pedagogy*. Cambridge University Press, Cambridge, UK

Schmitt N. (2000). *Vocabulary in language teaching*. Cambridge University Press, Cambridge, UK

Stubbs M. (2002). *Words and phrases: corpus studies of lexical semantics*. Blackwell Publishing, Oxford

van Ek J. & Trim J. L. M. (2001). *Vantage*. Cambridge University Press, Cambridge, UK

Widdowson, H. (2003). *Defining issues in English language teaching*. Oxford University Press, Oxford, UK

Wolter B. (2009). *Meaning-last vocabulary acquisition and collocational Productivity*. In Fitzpatrick T. & Barfield A. (Ed.), Lexical Processing in Second Language Learners: Papers and Perspectives in Honour of Paul Meara (Second Language Acquisition), Multilingual Matters, Bristol 128-140

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press, Cambridge, UK

Be yourself
in English.

Pearson