BRITISH COUNCIL

ASSESSMENT RESEARCH GROUP

KNOWLEDGE-BASED VOCABULARY LISTS USER MANUAL: A Supplementary Resource

Norbert Schmitt
Karen Dunn
Barry O'Sullivan
Laurence Anthony
Benjamin Kremmel

ISBN: 978-1-7397544-0-2
© BRITISH COUNCIL 2022

CONTENTS

OVERVIEW

The Knowledge-based Vocabulary Lists (KVL) are lists which indicate the probability that second language learners will be able to correctly spell a range of English words. They were compiled by testing large numbers of Spanish, German and Chinese learners of English using an online tool by the project researchers during 2018–19. A full description of the project is given in the British Council Monograph of the same name (Schmitt, Dunn, O'Sullivan, Anthony, and Kremmel, in press). Words which most learners knew are considered easier, and words which fewer learners knew are considered harder. The lists give 5,000 lemmas in sequential order, from easiest to hardest. The size of the list (5,000) was chosen because research shows that this amount of vocabulary allows leaners to operate in English to a considerable extent, and because it was the largest size the research study was able to produce within the time and financial constraints. The lists should be useful for pedagogical purposes in which it is beneficial to know whether learners are likely to be able produce and correctly spell the words they know. This User Manual accompanies the publication, *Knowledge-based Vocabulary Lists*, published by Equinox; the authors are: N. Schmitt, K. Dunn, B. O'Sullivan, L. Anthony and B. Kremmel.

What is a word?

Each KVL consists of lists of words. But these words represent more than just the word on the list. They represent *lemmas*. A lemma consists of a *base word* (also called *stem word* or *root word*), for example *deliver*, plus all of its grammatical inflections (*delivered*, *delivering, delivers*). Research has shown that learners are largely able to understand or produce these inflected forms if they know the base form, or vice-versa. However, they often have problems with *derivative* forms, where the word class is changed, e.g., *deliver* (v.)→delivery (n.) or *deliver* (v.)→*deliverable* (adj.). Thus, the lemma is the best sized 'package' to use with L2 learners when it comes to related words. Therefore, when using the KVL, remember that items on the list represent lemmas, and not just the individual base words.

What level of vocabulary knowledge do the KVL explain?

Many people and materials talk about learning and knowing vocabulary. But what does 'knowing' a word really mean? It could actually refer to many levels of knowledge. It could refer to merely knowing that a word exists and perhaps having some vague idea of its meaning and/or word form (pronunciation and spelling). This would be the level of knowledge at the very beginning of the learning process.

It could also refer to the ability to understand the word when it is heard or read. This is typically called *receptive knowledge*. The word might be known better, to a point where learners can pronounce and/or spell the word correctly but cannot necessarily use it appropriately in sentences and wider discourse.

Finally, the word can be fully mastered, where a learner can not only recognize the word when used in speech or writing, but can also use the word accurately and appropriately in any context they so desire (*productive knowledge*). Thus, knowledge of any word lies on a continuum, and we must be clear about which point of the continuum we are referring to.

When developing the KVL, we wished to test a level of mastery at the more advanced end of this continuum. It is impossible to test full productive mastery, as this entails being able to use a lemma correctly in a wide variety of appropriate contexts, and no practical test can measure this diversity. We opted for a more restricted level of productive mastery that was possible to test: *form-recall mastery*, i.e. being able to recall the form of a word when the meaning is given. The test we employed to measure this used the following format. The example for *house* was given to Spanish-speaking learners. The English translation for this item (which the learners did not see) is: *house I live in a large **house** that has three bedrooms*. Only fully correct spellings were accepted.

casa Vivo en una **casa** grande que tiene tres dormitorios.

h __ __ __ __

Thus, the level of knowledge which the KVL describes is the ability to accurately spell a lemma if the meaning is known. Research shows that this level of knowledge is relatively advanced, and is higher than receptive knowledge, as defined above. But the test format does not demonstrate that the lemmas can be used appropriately in sentences. So, the KVL cannot claim to represent full productive knowledge. Since the test was in a written mode, the KVL also makes no claims about the ability to employ the lemmas in listening or speech. It must also be noted that the test prompts gave the first letter and number of blanks for the tested lemmas, so learners may not always be able to spell the words independently without this information.

There were cases where the researchers retrospectively discovered that there were problems with the test items on which the KVL rankings were based. These lemmas appear in *Supplementary List C: KVL lemmas that are uncertainly ranked* at the end of this User Manual. They are also flagged in the KVL spreadsheets with a plus mark (+) in the column *Uncertain Ranking* to indicate that the learner knowledge may actually be higher than the KVL rankings indicate.

Other lemmas affected by poor test items did not make the KVL, but may have if their test items functioned properly. These lemmas are listed at the end of this User Manual in *Supplementary List D: Lemmas that potentially could have been placed on the KVL*. It is possible that learner knowledge may actually have been high enough for these lemmas to make the KVL if their test items worked well.

In some cases, the Spanish and German equivalents for English words were exact translations, e.g., for English *jazz,* the Spanish and German words are also *jazz*. This meant that, for a small number of lemmas on the KVL test, the correct answer could be obtained by simply putting the translation from the prompt into the answer slot. It is likely that most respondents answered these test items correctly because they knew the lemma, either from exposure, instruction, or understanding the L1-L2 similarity.

But is possible that some respondents 'gamed' the test and inserted the prompt translation into the answer slot even though they did not know the English word. This possibility makes the knowledge rankings for these 'exact translation' lemmas less certain than we would like. These lemmas are listed at the end of this User Manual in *Supplementary List E: Lemmas which may not be known as well as the KVL rankings indicate*. They are also flagged in the column *Uncertain Ranking* in the KVL spreadsheets with minus signs (-) to indicate the possibility learners might not know these lemmas as well as the KVL indicate.

Regardless of these caveats and lists, the KVL reports a level of knowledge that is more advanced than most other lists based on receptive-only tests and so can be used as an indicator of relatively advanced levels of knowledge. Furthermore, they are likely to be the best resource currently available to use as a proxy for fully productive levels of knowledge.

Why are the KVL 5,000 lemmas long?

The KVL contains the best-known 5,000 lemmas for each of three language groups (Spanish speakers, German speakers, and Chinese speakers). The 5,000 figure was chosen for three reasons. The most important is that 5,000 lemmas supply enough vocabulary to be able to use English to a considerable extent. This much vocabulary should allow learners to understand everyday conversation, and much of more detailed, specific spoken discourse (e.g., lectures, radio talk shows, television programmes). It should also allow the reading of easy English texts (e.g., teenage novels), and will allow entry into everyday adult texts (e.g. newspapers, magazines and textbooks), although likely requiring some support from teachers.

The second reason related to practicality. It would have been ideal to create a list of up to 10,000 lemmas but this was simply not possible. In order to capture the best-known 5,000 lemmas, we tested 7,679 of the most common lemmas to see which were known by the greatest number of participants (although only 7,532 were included in the final analysis — see below). Each lemma was tested on between 125–150 learners, so the project was massive (e.g., 7,679 lemmas x 125 participants = 959,875 responses). This was multiplied by three languages, which resulted in about 3 million responses. The project took five years to complete, and so it was not possible to compile lists beyond 5,000 lemmas for each language.

The final reason is that a large number of learners never progress beyond learning a few thousand lemmas, and so the 5,000-lemma lists are likely to be suitable for the majority of students around the world learning English as a second language.

Despite the great care taken in compiling the list, no research method is infallible. It must be acknowledged that some relatively well-known lemmas may have been missed in our procedure and may not appear on the lists. However, we feel confident that any potential missing lemmas are few in number and can simply be added to the lists by practitioners if discovered.

Which words are not included on the KVL?

Not all of the best-known words are included on the KVL. Our test format required accurate spelling for a lemma to be counted as known. This caused problems for lemmas in which there were alternative spellings in British vs. American English (e.g. *colour/color*, *realize/realise*, *dialogue/dialog*). As we provided the exact number of blanks for the letters in the test lemmas, and the web application would not allow additional letters to be added, this meant that respondents were forced to spell according to the language variety they chose at the beginning of the test. For example, if they chose American English, the prompt for *color* would be 'c _ _ _ _'. If they happened to know the British spelling instead (*colour*), this would not fit into the blanks. The KVL intends to indicate whether learners can correctly spell lemmas, not whether they are able to predict whether their knowledge is of British or American spelling. This means we would have liked to accept either British or American spelling, but the limitations of the test format did not make this possible. As a consequence, we were not able to derive reliable rankings for lemmas with alternative British vs. American English spellings, and these lemmas are not included on the list. These lemmas are listed in *Supplementary List A: Lemmas with alternative British vs. American spellings that are not included in the KVL* at the end of the User Manual. Many of these lemmas are likely to be well-known and would have made the list if we were able to test them accurately.

There were also limitations in how many lemmas we could test, so we were forced to reduce the list of lemmas we tested. Lemmas were excluded from our tests for various reasons, and these lemmas are listed at the end of the User Manual in *Supplementary List B: Lemmas which are not included on the KVL*. There were various reasons for these exclusions. For example, the purposes of the KVL are pedagogically oriented, and so lemmas which were not likely to be taught were excluded (e.g., taboo words). Proper nouns were also excluded (e.g., *Atlantic Ocean*), because if we included these, the number of words to test would have become impossibly large. Function words (e.g., *do, can*) were excluded because they could not easily be defined on the test. More information for the reasons of exclusion is given in the Supplementary List.

How should I understand the sequencing of the KVL?

The order of the lemmas on the KVL were sequenced according to the number of participants who answered the lemmas correctly, and also taking into account the relative proficiency of the participants themselves. This resulted in a sequence from #1 to #5,000. There are several important points to remember when using this sequence.

First, the list is probabilistic in nature. This means that lemmas earlier in the list (e.g., 50) are more likely to be known than lemmas later in the list (e.g., 250). It does **not** mean that the earlier lemmas will *necessarily* be known before later ones, just that it is more likely that they will be for a majority of learners.

Second, every learner is different, and KVL will not describe the knowledge of any individual learner. But it should describe the knowledge of learners in general for each language group and should be a good approximate description for any individual learner.

The third point concerns the degree of precision. The lists are not precise enough to say that nearly adjacent lemmas differ in degree of difficulty. For example, we cannot say that #1,500 will be known before #1,505. Rather, the larger the gap between two lemmas, the better the chances are that the list will correctly describe the order of knowledge. In general, lemmas that have a difference of ±100 rank places or more can be seen as having different degrees of difficulty. For example, it is very likely that #1,500 will be known before #1,600, and even more likely that it will be known before #1,700.

Overall, it is better to think of the list in terms of 'blocks' of lemmas instead of as a one-by-one list. For example, most of the lemmas in Block 700–750 are likely to be known before most of the lemmas in Block 850–900 by most of learners, most of the time (i.e., it is a probabilistic list, but the minimum of ±100 spacing gives a good degree of confidence). We have refrained for setting the size of the blocks, as this will differ according to the needs of different practitioners. Some may wish to work with smaller blocks (e.g., 50 lemmas) and some might find it to better to use larger blocks (e.g., 250 lemmas). But in any block, there may be lemmas that are learned earlier or later than might be expected.

Users should feel free to move the order of the lemmas on these supplementary lists if the proposed sequence does not match up with their own experience with their particular learners.

Fourth, there were a limited number of lemmas for which it was difficult to establish knowledge rankings. For a small number of lemmas, their test items proved misleading. These lemmas might be known better than indicated on the KVL, and so they are indicated by a plus mark (+) in the column *Uncertain Ranking* in the KVL spreadsheets, and are listed in *Supplementary List C: KVL lemmas that are uncertainly ranked*.

Other lemmas might not be known as well as indicated on the KVL. This was mainly because of the difficulty in providing workable prompts for words which have very similar word forms in both English and the learners' L1 (i.e., cognates and loanwords). For example, *jazz*, is the commonly used word for this style of music in English, German and Spanish. No other prompt would adequately convey this meaning. Thus, for some lemmas, the L1 test prompt had the exact same spelling as the required English answer.

It is possible that some respondents might have simply copied the prompt when answering the test, although the evidence suggest this practice was not widespread. Nevertheless, for this type of lemma, it is possible that the KVL rankings are somewhat higher than they should be. In the KVL spreadsheets, these lemmas are indicated by a minus mark (-) in the column *Uncertain Ranking*, and are listed in *Supplementary List E: Lemmas which may not be known as well as the KVL rankings indicate*.

Fifth, the lists are language specific. Although we found broad similarities in the sequences of the Spanish, German and Chinese groups, the variation between groups was so large that each group needed their own KVL with its individual sequence of learning. This was largely caused by *cognates*. Cognates are words that have the same or a similar word form in two languages, often the result of coming from a common ancestor word. For example, one might not think that *generation* would be an especially well-known word. But it was for the Germans speakers (#24). This is because the German equivalent is exactly the same: *generation*. The Spanish also knew it very well (#251), and again we see a cognate similarity: *generación*. For these two cognate languages, *generation* was obviously easy to spell. The Chinese learners (from a non-cognate language using the ideograph 代) found it more difficult (#1,121).

Because of these kinds of differences, we present separate KVL for speakers of Spanish, German and Chinese. We do not yet know how well these three lists might describe the learning sequences of learners from other languages. Until further research is carried out, we can only suggest that practitioners select the list closest to their own language and explore whether it is useful for their context. Spanish would seem to be the most likely possibility for Romance languages, German for Germanic/Nordic languages, and Chinese for non-cognate languages.

Where do I find the KVL and what information is available?

The KVL will only be available online on this British Council KVL website:
https://www.britishcouncil.org/exam/aptis/aptis-expertise/knowledge-based-vocabulary-lists-kvl

This is because the KVL may be updated in the future as researchers receive additional test data from users which could lead to revisions. The revisions need to be done on a single authorized site, or else a range of different versions may proliferate on different websites, some revised and others not. Therefore, users should always check the British Council website to make sure the most up-to-date version is used. The first version of each of the three lists is KVL Version 1.00 (1 December 2021).

The main KVL for each language are presented in Excel spreadsheets named *KVL-Spanish*, *KVL-German* and *KVL-Chinese*. Each of these spreadsheets contains the following columns of information: the target lemma; word class; the knowledge rank order (derived from the analysis in this project); the frequency rank order based on frequency counts from the *Corpus of Contemporary American English* (COCA) in 2018; uncertain KVL rankings, as described in Supplementary lists C and E.

There is another set of spreadsheets available for researchers which contain more technical information about the lemmas' characteristics, and the design of the test items on which the knowledge rankings are based (*KVL-Spanish-Technical*, *KVL-German-Technical* and *KVL-Chinese-Technical*). Descriptions of the columns in these technical spreadsheets are given on the second sheet labelled '*Key*'.

Uses for the KVL

The KVL will often be used in conjunction with frequency lists, and so it is important to understand the most advantageous use of each. Frequency lists report how often words appear in written and/or spoken discourse, that is, how 'common' or frequent each word is. This makes frequency lists good sources for information about which words learners are likely to come across in English and need to know to operate in English. Thus, the lists are particularly useful in regard to the receptive skills of reading and listening.

But they are not particularly good at revealing which words learners actually know. This is where the KVLs have value. They were compiled based on the direct testing of English language learners and show the likelihood of these kinds of learners knowing individual lemmas to the level of being able to spell them accurately. Thus, the KVLs should be useful when it is advantageous to understand which lemmas learners are likely to know to a productive written level.

There are a range of possible applications for this knowledge-based information. We list a few possibilities to illustrate the uses of the KVL, but there will be many more.

- One main application will be informing the likelihood of learners achieving form-recall mastery of vocabulary. Current lists used to predict the knowledge/difficulty of words are usually based on frequency or on receptive measures of vocabulary knowledge. But the KVL should provide much better predictions about the sequence in which learners achieve spelling control over a range of lemmas. This information should be useful for writing teachers, and test developers who assess writing ability.
- In selecting reading materials, it is often useful to grade the readings to match the abilities of learners. This is currently done by frequency profiles. But frequency is only a crude proxy for knowledge. Using the KVL should give a better idea of whether learners know the words in particular texts or not. While the KVL are based on form-recall tests, research shows that if words are mastered to a form-recall level, learners can typically also understand the words when they see them. This makes the KVL potentially suitable for reading-based applications.

The sequencing of the KVL should provide a baseline for understanding which lemmas learners know. For example, if learners know many lemmas at the 1,000–1,500 level, it can be inferred that they will also know most of the other lemmas in that band, and also in the 1–1,000 range as well. While frequency lists also do this to some extent, the KVL are customized to each of the three language groups (Spanish, German and Chinese speakers), and so take account of words which are relatively easy for each group due to cognateness. Frequency lists do not take account of cognateness, and so the KVL should be a much better representation of learner knowledge than frequency lists.

- Because the KVL are language-specific, they can provide teachers with information about cases where lemmas are highly frequent, but less likely to be known by learners of a particular L1. That is, lemmas that learners might find unexpectedly difficult. Conversely, the lists can indicate lemmas, which though relatively infrequent, are likely to be known by learners because of L1 similarities. For example, *caramel* is a low-frequency lemma (#14,900 in frequency ranking). Yet is easy for German learners (#876 in knowledge ranking), because it is *karamell* in German.
- In testing, we often attempt to measure or discover which words learners know. To do this, we first need to build a pool of which words learners *might* know. Frequency lists have typically been drawn upon to build this pool. However, as frequency does not predict knowledge of individual words very well, target words drawn from frequency lists may not match learner knowledge very closely, which makes for inefficient and potentially misleading tests. Drawing on the KVL for pools of test words should give test developers a better chance of targeting the words on their tests to the level of their test-takers. Whereas, if the purpose of a test is to describe learner knowledge of words which the learners *need to know*, then frequency lists may be a better source, as they describe which words occur most commonly in discourse. Thus, frequency lists may be more suitable for *prescriptive* testing purposes (learners need to know frequent words), while the KVL may be better for *descriptive* testing purposes (understanding the inventory of words learners already know).
- In psycholinguistic experiments, a range of factors affect the processing of vocabulary. This makes it crucial that target words are selected which are controlled in terms of the word characteristics which make the words easier-to-more difficult to process. Frequency has been shown to be a robust word characteristic which affects processing. However, frequency does not account for cognateness, and so target words may be exceptionally easier or more difficult than frequency might suggest for particular language groups. The KVL provides psycholinguists a valuable alternative source of information about potential word knowledge/difficulty to use in building their experiments.
- Because vocabulary knowledge relates so strongly with virtually all aspects of language proficiency, vocabulary tests can be usefully employed as placement tests. The KVL can be used to select words of the proper difficulty for these placement tests.

- The increased use of artificial intelligence in language learning and assessment also suggests a potential application for the KVL. Currently, AI-driven systems rely on traditional vocabulary lists. This may well introduce bias into the system by assuming a particular level of difficulty that cannot be supported by evidence beyond frequency. Exploratory research into potential bias would therefore benefit the valid application of AI-driven assessment engines in learning and assessment.
- The KVL may also be of value in human-to-machine communication where we increasingly find automated dialogue systems in use, e.g., in automated computer, phone or car systems. Where a user's first language is known, then communication can be tailored by including vocabulary that is likely to be known to that user.
- Ultimately, it is probably most useful for teachers, materials writers, syllabus designers and test developers to use both KVL and frequency lists in conjunction, as long as they understand the strengths and limitations of each type of list.

KEY POINTS TO REMEMBER

✗ The KVL can be used with learners of English from any first language.
✓ **There are three separate lists suitable for Spanish, German and Chinese speakers.** While they may prove useful for speakers of other first languages, this remains to be demonstrated. Therefore, use them with caution for these other languages and determine how well they work in your own contexts.

✗ The items on the list represent all forms of a word, i.e. the complete word family (e.g., *persist, persisted, persisting, persists, persistence, persistent, persistently*).
✓ **The items represent lemmas, which include the base form of a word and its inflections** (e.g., *persist, persisted, persisting, persists*).

✗ The KVL include all of the best-known lemmas in English.
✓ **Limitations in the scope of the KVL project and of the test format means that it was not possible to provide accurate rankings for every well-known lemma**. The lemmas which were not tested are given in the Supplementary Lists at the end of the User Manual. Your learners may well know these lemmas, even though they are not included on the lists.

✗ The lists represent the order in which lemmas are fully learned and can produced with complete fluency, accuracy and appropriacy.

✓ **The list represents the order in which learners can spell the lemmas accurately if the meaning is given or is known** (i.e. a form-recall level of mastery). Learners may or may not be able to produce the lemmas appropriately in sentences and wider discourse. However, it is highly likely that they will be able to understand the meaning of the lemmas when reading.

✗ The list gives information about the order in which oral vocabulary is learned.

✓ **The list gives information about the likelihood that learners can spell English words correctly.** The extent to which the list describes the likelihood of being able to produce oral vocabulary remains to be established.

✗ The sequence of the lemmas on the list is precise and learners will always know the lemmas in this order.

✓ **The sequence of the lemmas is probabilistic.** This means that earlier lemmas on the list are likely to be known earlier than later lemmas on the list, for most learners, most of the time.

✗ Lemma #50 will be learned before lemma #51.

✓ **The KVLs do not have this level of precision.** The bigger the gap between lemmas (e.g. #50 and #250), the greater the likelihood that an earlier lemma will be known before a later lemma. In general, a difference of ±100 or more places should allow for a fair amount of confidence concerning differences in difficulty. Thus, it is better to think in terms of clusters of words being more likely to be known (e.g. lemmas #50–100 more likely than lemmas #151–200) than individual lemmas (e.g., lemma #50 more likely than lemma #60. Also, a limited number of lemmas produced uncertain rankings which may not reflect your learners' knowledge. These are given in the Supplementary Lists at the end of this User Manual.

✗ The KVL can describe the probabilities of particular learners knowing words.

✓ **The KVL describe the average knowledge of the groups of learners that we tested.** Therefore, the list describes the knowledge patterns of groups of learners. The lists are unlikely to describe the knowledge of any particular learner, as every learner has their own idiosyncratic exposure to English, school syllabus materials, and study habits.

✗ The KVL are fixed and will never change.

✓ **The KVL may be revised and updated as the research team receives additional test data from people who are using the lists.** If they are updated, a new version number will be assigned. The latest version will always be available on the British Council KVL website.

✗ The KVL can only be used in the ways suggested in the User Manual.

✓ **The KVL are intended as a resource which provide information about the probabilities of learners knowing English vocabulary to a form-recall level of mastery.** Users can employ the lists in various ways which they may find beneficial, but they should always do so with the limitations stated in this User Guide in mind.

SUPPLEMENTARY LIST A:
Lemmas with alternative British vs. American spellings that are not included in the KVL

American English (AE)	British English (BE)
adapter	adaptor
aging	ageing
aluminum	aluminium
analyze	analyse
apologize	apologise
archeological	archaeological
archeologist	archaeologist
armor	armour
artifact	artefact
authorization	authorisation
authorize	authorise
authorized (adj)	authorised
behavior	behaviour
behavioral	behavioural
blond	blonde
burned (adj)	burnt
capitalization	capitalisation
catalog (n)	catalogue
categorize	categorise
center (n)	centre
center (v)	centre
centimeter	centimetre
centralization	centralisation
centralized (adj)	centralised
characterize	characterise
civilized	civilised
color (n)	colour
color (v)	colour
colored (adj)	coloured
colorful	colourful
coloring (n)	colouring
counseling (n)	counselling
counselor	counsellor
criminalization	criminalisation

American English (AE)	British English (BE)
criticize	criticise
customize	customise
customized (adj)	customised
defense	defence
dialog	dialogue
disk	disc
disorganized	disorganised
emphasize	emphasise
encyclopedia	encyclopaedia
endeavor	endeavour
energize	energise
enroll	enrol
equalize	equalise
equalizer	equaliser
equalizing (adj)	equalising
favor (n)	favour
favor (v)	favour
favorable	favourable
favorite (adj)	favourite
favorite (n)	favourite
fiber	fibre
finalize	finalise
flavor (n)	flavour
fulfill	fulfil
generalization	generalisation
generalize	generalise
glamor	glamour
globalization	globalisation
gray	grey
harbor (n)	harbour
harmonize	harmonise
honor (n)	honour
honor (v)	honour
honorable	honourable

American English (AE)	British English (BE)
hospitalize	hospitalise
humor (n)	humour
hypnotize	hypnotise
idealized	idealised
initialize	initialise
jewelry	jewellery
judgment	judgement
labor (n)	labour
legalization	legalisation
legalize	legalise
license	licence
licensed (adj)	licenced
localization	localisation
localize	localise
localized	localised
maximize	maximise
memorize	memorise
minimize	minimise
mom	mum
monolog	monologue
motorized	motorised
multicolored	multicoloured
nationalization	nationalisation
nationalize	nationalise
naturalize	naturalise
naturalized	naturalised
neighbor	neighbour
neighborhood	neighbourhood
neighboring (adj)	neighbouring
neutralization	neutralisation
neutralize	neutralise
normalize	normalise
odor	odour
offense	offence
optimization	optimisation
optimized	optimised
organization	organisation
organizational	organisational
organize	organise
organized (adj)	organised

American English (AE)	British English (BE)
organizer	organiser
personalization	personalisation
personalize	personalise
practice (v)	practise
practice (n)	practise
practiced (adj)	practised
program (n)	programme
program (v)	programme
realize	realise
recognizable	recognisable
recognize	recognise
reorganization	reorganisation
reorganize	reorganise
rumor	rumour
skeptical	sceptical
socialization	socialisation
socialize	socialise
specialization	specialisation
specialize	specialise
specialized	specialised
spoiled (adj)	spoilt
stabilization	stabilisation
stabilize	stabilise
stabilizer	stabiliser
sterilization	sterilisation
summarize	summarise
symbolize	symbolise
sympathize	sympathise
synchronization	synchronisation
synchronize	synchronise
synchronized (adj)	synchronised
theater	theatre
traveler	traveller
tumor	tumour
unauthorized	unauthorised
uncivilized	uncivilised
unrecognizable	unrecognisable
unrecognized	unrecognised
utilize	utilise

SUPPLEMENTARY LIST B:
Lemmas which are not included on the KVL

We did not include the following categories of words on our test, and therefore these types of lemma are not included in the KVL.

- Compounds (e.g., the race term *African American*) and all hyphenated words (*full-time, hi-tech*, and *long-term*), as 1) the meaning of these can usually be gained from understanding the meanings of the individual words (*full-time* = *full* + *time*), and 2) these items would have been difficult for our test format.
- Function (grammatical) words like *do* and *can*, which cannot easily be defined.
- Exclamations like *ha, huh, mm, hmm, oh, uh, yeah.*
- Taboo words like *bastard*, *bitch*, *fuck*, *fucking*, and *shit* because the purpose of the KVL is pedagogical and it is unlikely these would ever be part of pedagogical materials.
- When possible, we put together compound words with a space: *health care* → *healthcare.*
- Abbreviations like *i.e.*, *e.g., vs.,* and *re.* but we kept *AM* and *PM* as they seemed more like individual items with a meaning connected with time.
- We included the cardinal directions (*north south, east, west*) and their adjectives (*northern, southern, eastern, western*), but not sub-compass points (e.g., *northwest, southeast, northwestern, southeastern*).
- We did not include *n't.*
- We kept the most basic numbers (*one, two … ten, twelve, hundred, thousand, million*), but deleted others as quantity of number words would soon have become unmanageable.
- Proper names, such as *Atlantic* (as in ocean).
- Measure terms, e.g., *mile, gram, kilogram.*
- Month or day names, e.g., *January, Sunday.*
- Holidays: *Christmas, Easter, Halloween.*
- Comparatives and superlatives: *bigger, biggest.*
- Money units: *dollar, cent* (except *Euro*).
- Very colloquial lemmas: *dude, papa, mummy.*
- Technical vocabulary (which is obviously difficult and/or specific): *electromagnetic, electrotherapy, multicellular.*
- Words which do not make much sense alone, but are used as compound words, e.g., *keeping* which is mainly used in words like *beekeeping* and *peacekeeping*. Similarly, *shaped*, which is mainly used in compounds like *pear-shaped*. Other examples of words like these included *kept* (*well-kept, best kept*) and *lived* (*short-lived, long lived*).
- Plurals. We assumed that if learners know a noun like *need*, they also know the plural *needs*.
- Prefixed lemmas: *Un-* is transparent, and if a learner knows *acceptable*, then she probably knows *unacceptable* as well. Also, almost any word can be negated with *un-*, so we could not test every higher-frequency negative *un-* lemma, or it would have pushed out too many other content lemmas that we wanted to test. However, some words with the prefix *un-* are very frequent, and very well-known. In order to accommodate both of these contrasting points, a selection of lemmas (which were both of the highest frequency AND among the best-known) were kept (e.g., *incorrect, informal, unpopular, unusual*), but the rest were not included.

SUPPLEMENTARY LIST C:
KVL lemmas that are uncertainly ranked

The knowledge rankings for these lemmas were found to be based on test items that were potentially misleading. This means that accurate rankings would likely be somewhat higher than indicated by the KVL. The researchers were unable to determine what these accurate rankings might be, so users should be aware that learners might know these lemmas better than indicated by their rankings on the KVL.
In the KVL spreadsheets, these lemmas are indicated by a plus mark (+) in the column *Uncertain Ranking*.

Spanish-KVL	***German-KVL***	***Chinese-KVL***
not	painting (n)	big
		mostly
		trip (n)

SUPPLEMENTARY LIST D:
Lemmas that potentially could have been placed on the KVL

The knowledge rankings for these lemmas were found to be based on test items that were potentially misleading. Their knowledge rankings were beyond the 5,000 level, and so these lemmas were not placed on the KVL. However, accurate rankings would likely be somewhat higher than indicated, and so some of these lemmas could potentially have found a place on the KVL if they were based on better test items. The researchers were unable to determine what these accurate rankings might be, so users should be aware that learners might know these lemmas better than their non-inclusion on the KVL would indicate, i.e., these lemmas could potentially have appeared on the KVL.

Spanish-KVL

drug (n)	mostly	public (n)
easy (adv)	nature	super
grand	near	tight
hurt (v)	off	
most	public (adj)	

German-KVL

action	deal (v)	move (n)
afraid	grand	near
become	involve	position (n)
break (v)	lose	spot (n)

Chinese-KVL

clear (v)	lose	pride
false	mainly	reach (v)
freeze (v)	might (v)	simply
grand	plain (adj)	true

SUPPLEMENTARY LIST E:
Lemmas which may not be known as well as the KVL rankings indicate

Some lemmas are *cognates* (words which have similar spellings in both the L1 and L2), especially in German-KVL and Spanish-KVL. While we tried to avoid using cognates as the test prompts, in some cases the only appropriate prompts were cognates. An example of this is the lemma *jazz*, which is the commonly used word for this style of music in English, German and Spanish. Any other prompt would have been misleading. Thus, for some lemmas, the L1 test prompt had the exact same spelling as the required English answer, as in the *jazz* example. It is possible that a limited number of respondents on our test simply copied the prompt when answering this type of test item. Therefore, there is the potential that lemmas based on cognate prompts with the exact same spelling may be ranked somewhat higher than they should be. In the KVL spreadsheets, these lemmas are indicated by a minus mark (-) in the column *Uncertain Ranking*. They are listed below.

Spanish-KVL (101)

altar
amnesia
autism
bacterial
ballet
bikini
bingo
bisexual
blog (n)
bravo
brownie
cable (n)
campus
canal
capital
casino
chocolate
civil
club (n)
cobra
collage
colonial
conceptual
continental
crisis
cultural
delta
detector
diabetes
digital
dimensional
disco
eclipse (n)
editorial (adj)
electoral
euro
experimental
fax (n)
federal
fiscal
general (n)
golf (n)
hardware
heterosexual
hockey
homosexual
hotel
industrial
instrumental
invisible
jaguar
jazz
jet (n)
karaoke
karate
karma
kickboxing
kiwi
knockout (n)
liberal (n)
literal
mental
metal
microchip
mineral
molecular
mozzarella
multicultural
multimedia
natural
nazi
ninja
no
oral
paintball
panda
pasta
piano
pitbull
plasma
polar
portal
propaganda
radar
radio
reactor
safari (n)
social (adj)
softball
suite
sushi
tango (n)
taxi (n)
terminal (n)
topless
trauma
tsunami
variable
virus
vodka
yoga

German-KVL (328)

absurd
adoption
adverb
agent
airbag
album
algebra
alligator
alpha
alphabet
altar
alternative (n)
analyst
android
aquarium
arena
arm (v)
aspirin
asteroid
astronaut
avatar
baby (n)
ball
bar (n)
baseball
basketball
bass (n)
bikini
bitter
blind (adj)
block (v)
blocker
blog (n)
blogger
bodybuilder
bronze (n)
brownie
browser
brutal
budget (n)
burger
butter (n)
cheerleader
cheeseburger
cocktail
collage
comic (n+adj)
computer
countdown
cousin
cowboy
cupcake
deck (n)
delegation
delta
demo (n)
deodorant
depression
design (n)
designer
desktop (n)
digital
dimensional
diplomat
dna
dock (n)
dollar
dominant
drama
echo (n)
ego
elegant
element
elite
ensemble
euro
evolution
exhibitionist
experiment (n)
explosion
export (n)
eyeliner
fan (n)
fantasy
feminist
festival
film (n)
filter (n)
finalist
finger
fit (adj)
fitness
flamingo
form (n)
format (n)
forum
fossil
franchise
freak
frustration
fundamentalist
garage
gas (n)
general (n)
generation
generator
gladiator
global
gold
golden
golf (n)
gorilla
gospel
grapefruit
hacker
hamburger
hammer (n)
hand (n)
handball
hardware
heroin
hi
hit (n)
hobby
hotdog
hotel
hotline
hunger (n)
ideal (adj)
idol
illegal
imperial
improvisation
in
individualist
inflation
information
insider
inspiration
installation
instrument
integration
interface (n)
international
internet
interpretation
interview (n)
intolerant
intuition
investor
irrelevant
isolation
Israeli
jackpot
jaguar
jazz
journalist
joystick
karaoke
karate
karma
kiwi
land (n)
laptop
laser (n)
latex
layout
leopard

live (adj)
lobbyist
magnesium
mailbox
major
manipulator
marathon
massage (n)
materialist
matrix
mediation
meditation
medium (n)
memo
mentor
million
mineral
minimal
minimalist
minimum
minister
minus
minute
modem
modern
moment
motel
motivation
motivator
mozzarella
muffin
multimedia
museum
mutation
name (n)
nation
national
navigation
navigator
neon
nest (n)
networking (n)
neutral
neutron
ninja
norm
nylon
okay
olive
online
operation
opportunist
orange (n)
organ
outfit (n)
paintball
panda
panorama
parallel (adj)
park (n)
patent (n)
patient (n)
patriot
pause (n)
pc
person
phase (n)
pilot (n)
pitbull
pizza
planet
plasma
poker
pony
popcorn
portrait
post (n)
radar
realist
remix (n)
ring (n)
ritual (n)
rose (n)
sand (n)
scanner
sensor
server
setup (n)
shooter
shuttle (n)
simulator
single (n)
skateboard (v)
ski (n+v)
skyline
smartphone
snowboard (n)
softball
software
solo (n)
spaghetti
spam (n)
spoiler
spray (n)
sprint (n)
status
steak
streaming (n+adj)
studio
suite
superman
supermodel
superstar
surfer
sushi
synonym
system
tampon
tango (n)
taxi (n)
tennis
terminal (n)
terrorist
text (n)
thermometer
thriller
tiger
toaster
tolerant
tourist
tradition
traditionalist
trainer
transporter
troll (n)
tunnel (n)
uniform (n)
urgent
variable (n)
veteran (n)
vibrator
video (n)
virus
vitamin
volleyball
warm (adj)
webcam
website
wild (adj)
wind (n)
winter
wolf
workshop
yoga
zebra
zombie
zoo

British Council
Assessment
Research Group

The Assessment Research Group was formed in 2013 to support the British Council's work in assessment and testing across the world. The team is responsible for ensuring that all new assessment products and new uses of existing products are supported by the most up-to-date research. They also continuously evaluate the quality of British Council assessment products.

KNOWLEDGE-BASED VOCABULARY LISTS USER MANUAL: A Supplementary Resource

Norbert Schmitt
Karen Dunn
Barry O'Sullivan
Laurence Anthony
Benjamin Kremmel

Published by British Council
1 Redman Place, Stratford
London E20 1JQ
United Kingdom

© British Council 2022
The British Council is the United Kingdom's international organisation for cultural relations and educational opportunities**.**

ISBN: 978-1-7397544-0-2

www.britishcouncil.org/aptis/research