WORD FREQUENCY LISTS:

RATIONALES, SELECTION, RECOMMENDATIONS, AND USES

1) RATIONALE: WHY WORD FREQUENCY IS IMPORTANT FOR LANGUAGE TEACHING

Vocabulary teaching is a core component of foreign language learning. Arguably, the most useful vocabulary for learners reflects the needs and interests of learners themselves, and is also informed by the frequency with which words occur in real language use (Nation & Meara, 2002). Language corpora (bodies of actual spoken, transcribed, and/or written language collected together from one or more sources) indicate that a relatively small number of high-frequency words represent a large proportion of the total words in a written text or speech. According to some estimates, the most common 2000 words represent around 80% of the words in any written text and an even greater percentage of the words in speech (Nation, 2001; Nation & Waring, 1997), highlighting the importance of word frequency in informing vocabulary learning.

The teaching of high-frequency words is particularly important during the early stages of language development. Over the last few decades in the UK MFL context, the choice of what words to teach has often been topic-driven, which leads to the teaching of specialised and rare vocabulary before teaching more commonly occurring words. For example, many of the words for pets or hobbies will be low frequency words which are not useful beyond those particular topics. This means that very precious time can be spent learning relatively rare words. If learners do not first learn the most frequent words, they will not be able to say or understand basic things. A further consequence of topic-driven vocabulary teaching is that learners may only recall a word when it is needed for one particular topic, even though the word is used across many different contexts. For example, pupils may learn the word 'voiture' when describing how they travel to school. The word may be strongly associated for them with that particular topic; they may then be unable to apply the same word in new contexts (e.g. when talking about pollution or fitness). Note that the AQA specification explicitly requires pupils to be able to use words across different contexts and topics in this way: '*Vocabulary listed under a particular theme should be considered transferable, as appropriate, to the other themes.*'

There is some evidence of an awareness of the importance of word frequency in existing materials. Textbooks, for example, sometimes provide lists of 'common words', and examination boards provide word lists. However, these lists are based on intuition or personal experience, which are not necessarily consistent with the data found in large language corpora (Biber & Reppen, 2002; see also Anderson, 2007a). Illustrating this problem, research has identified disparities between the words taught and used in textbooks and classroom environments on the one hand, and the frequency with which they occur in real-life use on the other (Anderson, 2007b; Biber & Reppen, 2002; Häcker, 2008; Holmes, 1988;).

To sum up, word frequency information can inform: a) teachers' choice of words to teach; b) textbook design; and c) test design. However, word frequency lists differ in how they are compiled and so it is important to be transparent about the criteria used to select a frequency list for language teaching.

2) SELECTION: CRITERIA FOR SELECTING A WORD FREQUENCY LIST FOR LANGUAGE TEACHING

During our search for lists of high frequency words in French, Spanish and German, we considered the following factors:

Size of corpus from which the word list is derived. In principle, a word list based on a bigger corpus is better. Because the word frequency data comes from a larger dataset, it is more likely that the frequency ranking of each word is accurate. However, most corpora tend to produce relatively similar lists of *high* frequency words (differences emerge in lists of *low* frequency words), meaning that corpus size was a less important criterion for us than others.

Unit of counting words. What counts as a word? Scholars recommend 'inclusive' definitions (i.e. counting multiple forms as one word for the purposes of estimating frequency). There are two main stages when grouping different word forms together. First, inflectional forms (e.g., 'playing', 'plays', 'played') are forms of the same lexeme 'PLAY'. Second, each LEXEME has different derivational forms (e.g. 'PLAY', 'PLAYER', 'PLAYABLE' etc.) which belong to a wider 'word family'. Learners are usually expected to understand multiple forms within the same 'word family', even if they cannot produce them (Nation, 2010).

The list should also explain how it counts multi-word units and expressions (e.g., 'by and large').

Representativeness. The word list should be representative of (or explicit about) the following:

- Modality: Written and/or spoken language (Nation, 2010; Nation & Waring, 1997) and sometimes these are examined separately, since spoken and written language typically differ in terms of the vocabulary they contain.
- Location: The geographical usage of the language (Nation & Waring, 1997).
- Genre: A range of text-types and genres, e.g., fiction, non-fiction, popular magazines. newspapers, academic journals, blogs (Nation, 2010; Nation & Waring, 1997).
- Time period: When the language was collected, with a preference, in many cases, for including recent language (Davies, 2019)

Part-of-speech tagging. The words in the list should be annotated with information regarding word class (noun, verb etc.). There should also be evidence of efforts to manually clean the tagging that is usually done by software first, so as to make the tagging as accurate as possible. (For example, software may not always identify whether a word is used as a noun or verb e.g., learning, meeting.)

Other information included with each word in the list/corpus. This might include: the core meaning of the word; variations of meaning; collocations (what other words often occur alongside it); the relative frequency of different meanings and uses; restrictions on the use of the word with regard to different contexts, such as register (formality, politeness) (Nation & Waring, 1997).

***Searchability*.** How user-friendly is the word list for a language learner or teacher? Can it present frequency lists in more than one way?

***Wider use of the word list*.** Has the word list been used for research, pedagogy, or testing?

Status. Has the list been recommended by researchers or practitioners?

3) RECOMMENDATIONS FOR WORD FREQUENCY LISTS

We considered a number of corpora and word frequency resources in French, Spanish and German. Some of these were open access websites, while others were dictionaries available as e-books or in printed editions.

Our recommendations are: *A Frequency Dictionary of Spanish: Core Vocabulary for Learners* (Davies & Davies, 2018); *A Frequency Dictionary of German* (Tschirner & Möhring, 2019) and *A Frequency Dictionary of French* (Lonsdale & Le Bras, 2009), published by Routledge.

As Table 1 shows, these three frequency dictionaries meet many of the criteria above. In particular, they all include a) words counted by lexeme, rather than by individual word form; b) words tagged according to part of speech, with a clear concern for accuracy of tagging; c) a substantial amount of written and spoken language; d) a range of genres. It is also useful that there are multiple ways to search the dictionary (by frequency rank, by spelling, by part of speech) and that each one contains a set of words grouped by thematic lists. In addition, our three choices have been used more widely than others for both academic and pedagogical purposes in different contexts. For example, words from *A Frequency Dictionary of Spanish* have been directly incorporated into a game on Memrise, which provides a free, ready-made resource for vocabulary learning that is informed by frequency data.

We do, however, note some shortcomings in our choices. For example, unlike the very recent Spanish Frequency Dictionary, the French and German dictionaries draw on (some) slightly dated corpus data. Furthermore, the German frequency dictionary does not contain English translations of its sample sentences, which may be a limitation for lower-level learners wishing to see and understand words in context. However, these are outweighed by the limitations of other potential word frequency resources[1].

4) USES OF VOCABULARY LISTS: SOME SUGGESTIONS FOR CLASSROOM PRACTICE

Our vocabulary lists have been tagged for frequency in line with the resources that we have recommended. Our lists are closely related to the lists produced by the major examination boards in England. However, critically, our lists also indicate the words that are highly frequent but are *not* on the examination board lists. Our lists also show which words on the examination board lists are not very frequent.

For each of the three languages, we provide a single excel file with the lists presented by topic area, in alphabetical order, and by parts of speech. We also provide three word documents with these individual lists.

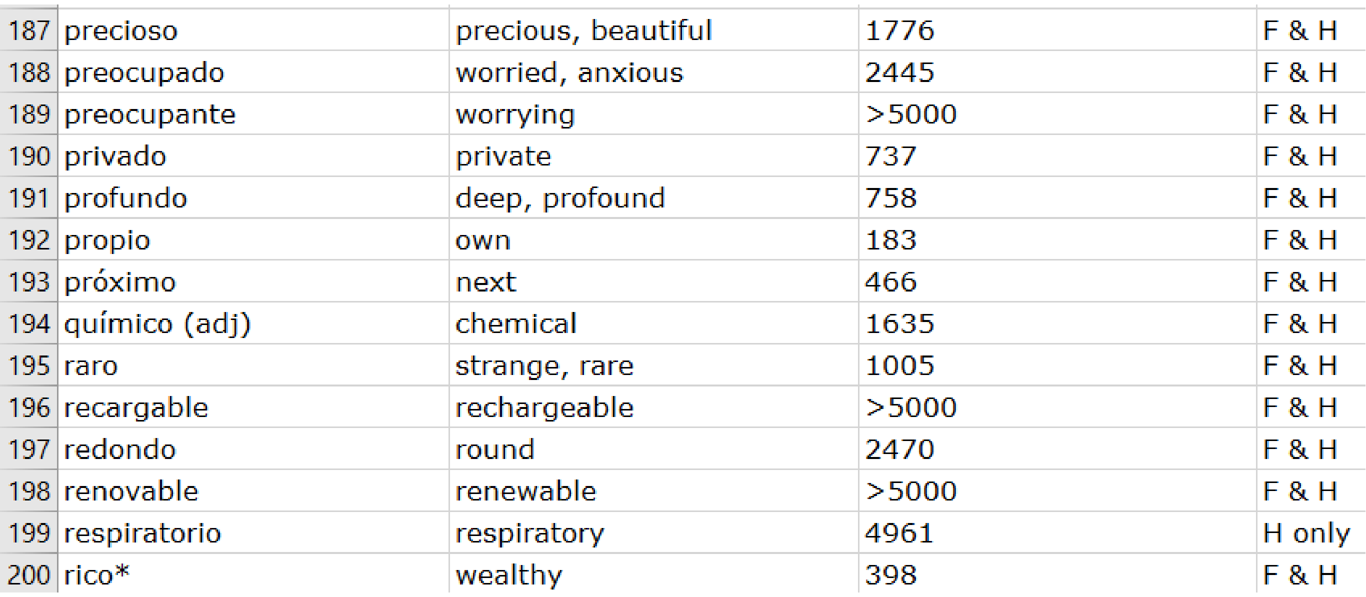
These resources can be used to:

i) check the selection of vocabulary to be taught and help plan a scheme of work
Learners cannot learn all of the vocabulary they encounter to the same level of mastery or depth. With limited teaching time, teachers have to make choices about which words receive the most teaching time and the most frequent revisiting. These word lists allow teachers to check their selections for frequency of use.

ii) stimulate interest
Pupils might be genuinely interested to know the frequency of the vocabulary they are learning, and this can be provided, as appropriate, on word lists given to pupils. They might

be interested to learn, for example, that 'aburrido' (boring) [3917] is by no means the most frequently used adjective in Spanish (see below).

Figure 1. Screenshot of Spanish word list for adjectives, with frequency rankings sourced from Davies and Davies (2018)



187	precioso	precious, beautiful	1776	F & H
188	preocupado	worried, anxious	2445	F & H
189	preocupante	worrying	>5000	F & H
190	privado	private	737	F & H
191	profundo	deep, profound	758	F & H
192	propio	own	183	F & H
193	próximo	next	466	F & H
194	químico (adj)	chemical	1635	F & H
195	raro	strange, rare	1005	F & H
196	recargable	rechargeable	>5000	F & H
197	redondo	round	2470	F & H
198	renovable	renewable	>5000	F & H
199	respiratorio	respiratory	4961	H only
200	rico*	wealthy	398	F & H

iii) teach multiple meanings of the same word

Very high frequency words often have multiple meanings. For example, 'rico' (Spanish for, approximately, 'rich') is probably highly frequent because it can mean both 'wealthy' and 'tasty'. Our excel lists mark words that appear more than once with an asterisk, a number of which have different meanings, e.g. 'tiempo' in Spanish, which can mean both 'time' and 'weather'. So, using the excel version, you can rank words in a particular word class (e.g. verbs) to illuminate multiple meanings. This enables teachers to help learners deepen knowledge of words (see, for example, the Spanish verbs 'pasar' (to happen, to go through, to spend (time), to pass), 'quedar(se)' (to stay, to agree), 'volver(se)' (to return, to do (something) again) etc.

iv) teach vocabulary in multiple word class clusters

There is some research evidence that teaching vocabulary in topic-based lists of words with a single word class (e.g. a list of just nouns) is not optimal (Häcker, 2008). Using the full list by topic allows teachers to build mixed word class vocabulary lists, based on the highest frequency. For example, a list of 10 words for Year 7 Spanish pupils to learn, when embarking on the theme of describing their town, might be:

	Word	**Frequency ranking**	**Word class (Part of speech)**
1	la plaza (square)	806	noun
2	la iglesia (church)	437	noun
3	el teatro (theatre)	605	noun
4	ser (to be)	7	verb
5	grande (big)	66	adjective
6	pequeño/a (small)	202	adjective
7	estar (to be)	21	verb
8	cerca (de) (near (to))	1042	adverb
9	lejos (de) (far (from))	833	adverb
10	el museo (museum)	1114	noun

v) teach relations between words within word families

Teachers often want to help their pupils spot the patterns between words of different word classes with the same root meaning. You can use the excel list A-Z to search for all instances of a particular word stem, e.g. 'val-' in the Spanish list, identifies '*¿cuánto vale?* (how much does it cost?), *la evaluación* (assessment), *vale* (ok), *valer la pena* (to be worth the trouble), *valiente* (brave).

vi) teach cognate patterns

The AQA specification identifies common cognate patterns between English and the target language, and carries the expectation that pupils will be able to understand such cognates, in both speech and writing. This includes cognates that are not listed in the GCSE word list. However, teachers can usefully search for these cognate patterns within the excel list, as a useful point of departure for learners. For example, searching for '-ción' in the Spanish list yields more than 30 nouns, and a search for '-oso' yields 18 adjectives (not all of which are cognates). Importantly, such searches highlight links between words from different topics and allow teachers to make inter-topic vocabulary links that they would otherwise only be able to do using their intuition.

vii) highlight differences between the highest frequency words and the examination board lists.

The examination boards do not include all of the most frequent 2000 words in their lists. We have therefore indicated these additional words in a separate tab.

Conversely, very many words on the examination board lists are not among the most frequent 5000 words. We have indicated these with a [>5000] symbol.

Notes

[1] We also considered *Corpus del español del siglo XXI* (Real Academia Española), *Sketch engine*, *Corpora from the web* (COW), and *LexEsp corpus.* The following were returned by our search but we were unable to access them: *Spanish key words: The basic 2000-word vocabulary arranged by frequency* (Pedro Casal), *Spanish-English frequency dictionary- essential vocabulary: The 2500 most used words & 468 most common verbs* (J. Laide), *2000 most common Spanish words in context: Get fluent & increase your Spanish vocabulary with 2000 Spanish phrases* (Lingo Mastery) and *The 8000 most frequently used Spanish words: Save time by learning the most frequently used words* (Neri Rook).

[2] Tschirner and Möhring (2019) count words by lemma, while also acknowledging that "these entries are in some cases forms of the word or lemma and not the base word or dictionary form. For our purposes, it was useful to combine the various forms of the definite article […], for example, into a single entry" (p. 4).

References

Anderson, B. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*, 383-409.

Anderson, B. (2007). Pedagogical rules and their relationship to frequency in the input: observational and empirical data from L2 French. *Applied Linguistics*, *28*, 286-308.

Biber, D. & Reppen, R. (2002). What does frequency have to do with grammar teaching? *Studies in Second Language Acquisition*, *24*, 199-208.

Davies, M. (2019). Word frequency data. Retrieved 04.03.2019 from www.wordfrequency.info

Davies, M., & Davies, K. (2018). *A frequency dictionary of Spanish: Core vocabulary for learners* (2nd ed.). London: Routledge

Häcker, M. (2008). Eleven pets and 20 ways to express one's opinion: the vocabulary learners of German acquire at English secondary schools. *Language Learning Journal*, *36*(2), 215-226.

Holmes, J. (1988). Doubt and uncertainty in EFL textbooks. *Applied Linguistics*, *9*, 21-44.

Lonsdale, D. & Le Bras. Y. (2009). *A frequency dictionary of French: Core vocabulary for learners*. London: Routledge

Nation, P. (2001). How good is your vocabulary program? *ESL Magazine*, *4*(3), 22-24.

Nation, P. (2010). *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

Nation, P. & Meara, P. (2002). Vocabulary. In N. Schmitt (Ed.) *An introduction to applied linguistics* (pp. 35-54). Arnold: London

Nation, P. & Waring, R. (1997). Vocabulary size, text coverage and word lists. In N. Schmitt., & M. McCarthy (Eds.) *Vocabulary: Description, Acquisition and Pedagogy* (pp. 6-19). Cambridge: Cambridge University Press.

Tschirner, E. & Möhring, J. (2019). *A frequency dictionary of German: Core vocabulary for learners* (2nd ed.). London: Routledge

Table 1. Characteristics of three recommendations for word frequency lists

	Spanish Frequency Dictionary: **Davies & Davies (2018)**	***French Frequency Dictionary:*** **Lonsdale & Le Bras (2009)**	***German Frequency Dictionary:*** **Tschirner & Möhring (2019)**
Number of words	5000	5000	5009
Name and size of corpus	Corpus del Español, 2 billion words	French word corpus, 23 million words	*New Leipzig Corpus of Contemporary German*, 20 million words
How words are counted	By lemma	By lemma	By lemma[2]
Are words 'tagged' for word class?	Yes	Yes	Yes
Is the corpus representative of… ***1) Speech and writing***	33.3% spoken; 66.6% written	50% spoken; 50% written	25% spoken; 75% written
2) A range of genres	*Written:* fiction and non-fiction. New edition includes language from websites and blog posts (50% of total corpus). *Spoken:* conversations, lectures, sermons, sports broadcasts.	*Written:* newswire stories (3 million); newspaper stories (2,015,000), literature (4,734,000), popular science magazine articles (434), newsletters, tech reports & user manuals (1,317,000). *Spoken:* conversations (175); Canadian parliament debates (3,750,000), interviews/transcripts (3,020,000); phone calls (855); theatre dialogue/monologue (470); film subtitles (2,230,000).	*Written:* 15 million drawn from newspapers (5 million); literature (5 million); academic and instructional material (5 million). *Spoken:* 5,900,000 including a range of television material, public speeches, oral interviews, radio shows, podcasts and presentations.
3) Time period the types of language are from	First edition (2006): largely based on written texts from 1900s and spoken texts from 1970-2000 (most of which were from the 1990s). Second edition (2018): Language	No material was included from before 1950. We were unable to track which periods between 1950 and 2009 the language was from.	Material spans 30 years between 1990 and 2019, but focuses on 2015-2019.

	collected from internet web pages in 2014-15 was added to the original corpus.		
4) Geographical usage	78% Latin America; 22% Spain	Corpus did not try to balance data in this way.	No overall figures Language drawn from Germany, Austria and Switzerland.
Main information with each entry	Frequency rank Head word Part of speech (word class) English gloss Sample sentence (& English translation)	Frequency rank Head word Part of speech (word class) English gloss Sample sentence (& English translation)	Frequency rank Head word Part of speech (word class) English gloss Sample sentence (& English translation)
Further information included	Raw frequency of entry in (a) fiction; (b) spoken; fiction; (c) non-fiction; (d) web corpus Indication of register variation	A ranking of a word often occurs across different material within the corpus Raw frequency Indication of register variation	Raw frequency Dispersion Multi-word units and dominant word forms
Other indexes	Alphabetical index Part of speech index Thematic lists	Alphabetical index Part of speech index Thematic lists	Alphabetical index Part of speech index Thematic lists