Between limitations and boundaries: The need for a comprehensive study and approach to the fundamental lexicon of Spanish

Isabel Sánchez López
Universidad de Granada

Received: 2025-05-15 / Accepted: 2025-10-17
DOI: *https://doi.org/10.30827/portalin.viXV.33669*
Porta Linguarum ISSN paper edition: 1697-7467, ISSN digital edition: 2695-8244

ABSTRACT: This study highlights the importance of optimizing the process of learning a new language through a comprehensive analysis of the core Spanish lexicon. A holistic approach is proposed, addressing internal structural delimitation, semantic, morphological, and phonological complexity, as well as lexical networks and processing. The research examines word frequency, lexical acquisition in academic contexts, conceptualization tasks, and the quantitative and qualitative profitability of vocabulary learning. It also discusses lexical centrality and the need for an interdisciplinary perspective. Practical tools such as lexical analysis, frequency indexes, lexical availability studies, and other complementary works are presented. Finally, the lexical nomenclatures used in the Chinese educational system are introduced as a comparative model, underscoring the relevance of corpora and lexicometric tools in the analysis and selection of fundamental vocabulary.

Keywords: vocabulary, core lexicon, learning, frequency, corpus, lexicometry.

Entre limitaciones y delimitaciones: La necesidad de un estudio y abordaje integral del léxico fundamental del español

RESUMEN: Este trabajo aborda el proceso de aprendizaje de una nueva lengua a través de un estudio integral del léxico fundamental del español. Se propone un enfoque holístico que incluye la delimitación estructural interna, la complejidad semántica, morfológica y fonológica, y el análisis de redes léxicas y procesamiento. Se examinan la frecuencia de uso, el aprendizaje léxico en el ámbito académico y las tareas de conceptualización, así como la rentabilidad cuantitativa y cualitativa en la adquisición del vocabulario esencial. Además, se analiza la centralidad léxica y la necesidad de un enfoque interdisciplinario, a través de herramientas prácticas como el análisis léxico, los índices de frecuencia, la disponibilidad léxica y estudios complementarios. Finalmente, se expone el modelo de nomenclaturas léxicas en el sistema educativo chino como ejemplo comparativo, destacando la relevancia de los corpus y herramientas lexicométricas en el análisis y selección del vocabulario fundamental.

Palabras clave: vocabulario fundamental, adquisición, frecuencia, corpus, lexicometría.

1. Introduction: The importance of optimising the process of learning a new language

Communication between people is based on the words that make up our vocabulary, and that vocabulary is not only useful, it is essential. When we look at our society, we realise that people who need to improve their vocabulary come from a variety of backgrounds. The way we learn and use words plays a role in how we communicate, both for native speakers and for those who are learning a second language. Several studies have shown that effective vocabulary instruction can have a significant impact on our ability to communicate (Webb & Nation, 2017; Milton, 2009).

Therefore, one of the foundations of our approach is its empirical basis. We want to do something that responds to social needs in different areas: integration, education, professional development, among others. We believe that two key principles for studying vocabulary are cost-effectiveness and efficiency, which means that we should consider not only how many words we learn, but also the quality of our repertoire. Recent research has pointed to the importance of selecting words according to criteria of frequency and utility, as well as the effect of exposure and context on how much we retain words (Schmitt, 2019).

The study of words in educational and communicative settings, which are the contexts we want to focus on, presents several challenges that affect their teaching, learning and use. Two of these challenges are holistic and structural approaches. These ideas help us to understand why some words are more important than others in different situations (Nation, 2001; Schmitt, 2008). Furthermore, although technologies in the classroom aim to improve our digital skills, it is essential to reflect on their pedagogical use so that they do not undermine our reading comprehension and vocabulary learning (Robin, S. J., & Aziz, A. 2022).

1.1. Holistic approach

We refer to how we choose and organise vocabulary according to its relevance in different contexts, as well as to people's social, academic and occupational needs. This approach sees vocabulary as dynamic, varying according to place and situation (Halliday & Hasan, 1976). For example, the words needed in a technical-scientific field are very different from those used in everyday conversations, and teaching must adapt to these differences in order for learning to be more effective (Laufer & Nation, 1995).

In this sense, this approach responds to practical criteria, prioritising the inclusion of terms that promote effective communication in specific contexts (Coxhead, 2000; Dang & Webb, 2016). Word selection may be based on factors such as professional relevance, cultural appropriateness or usefulness in certain situations. However, this approach can also lead to some valuable words being left out, making it difficult to create a shared repertoire. Recent studies have highlighted the need to complement this approach with learning strategies that encourage flexibility in word use and the creation of broader semantic networks (Schmitt, 2019; Webb, 2020).

1.2. Structural or internal delimitation

When addressing structural or internal delimitation in the study of the lexicon, reference is made to the constraints imposed by the intrinsic properties of lexical units. These constraints include the semantic, morphological and phonological complexity of words. (Aitchison, 2012; Bybee, 2003). This approach allows us to understand that not all terms present the same degree of accessibility to speakers, as their acquisition and use are mediated by multiple cognitive and linguistic factors (Ellis, 1994; Brown, 2021).

1.2.1. Semantic complexity

The semantic complexity of a word can greatly influences how we learn and use it. Some words have multiple or ambiguous meanings, which can make them more difficult to understand and use correctly. As Taylor (2003) says, words that have multiple meanings can cause confusion in learning because speakers need to identify the correct meaning based on the context. This problem occurs especially with abstract or figurative words, whose interpretation varies according to the speaker's prior knowledge and linguistic experience (Lakoff & Johnson, 1980). Recent studies have shown that polysemy and ambiguity not only affect comprehension, but also the speed with which speakers retrieve words when speaking or writing (Marrero & Palacios, 2023).

1.2.2. Morphological and phonological complexity

Words that have complicated morphological structures, such as those that include many affixes or irregular formations, tend to be more difficult to learn and remember (Plag, 2003). In Spanish, words with multiple suffixes can bring complications for non-native speakers, especially if the form does not follow regular patterns (Serrano et ali., 2009).

Similarly, the phonology of a word also affects its accessibility: words with unusual sound combinations or difficult-to-pronounce sequences can pose a greater challenge for speakers (Bybee, 2003). This is especially noticeable when learning a second language, where learners often avoid or modify complex phonetic structures to adapt them to what they know from their native language (Flege, 1995; Romero et alii. 2018).

1.3. Networks and lexical processing

The lexicon does not function in isolation, but within networks that organize words according to their conceptual relationships. Croft and Cruse (2004) mention that dense semantic nodes, that is, those terms that are highly interconnected with other concepts, can be an additional challenge for learners. This is because when activating a word in this semantic network, there may be interference from related terms, which slows lexical access and increases the cognitive load in language production (Aitchison, 2012).

Recent research has shown that the way our mental lexicon is organized affects the speed with which we access words, and that semantic connections can facilitate or complicate this access, depending on how terms are related (Sánchez-Saus and Álvarez, 2024).

1.4. Frequency of use and lexical learning

The frequency with which a word is used is a key factor in the internalization and acquisition of vocabulary. Biber et al. (1999) point out that words that are used a lot tend to be easier to learn and are used more naturally in everyday communication. On the other hand, words that are rarely used, such as technical or specialized vocabulary, require more mental effort to learn and remember (Nation, 2001).

This principle is supported by studies on language acquisition that have shown that both native and non-native speakers access more quickly words that are frequently used, compared to those that are less common in the language they receive (Ellis, 2002). In addition, research in the context of Spanish suggests that repeated exposure to infrequent terms, in contexts that make sense, can aid their long-term learning, especially in educational settings.

2. Vocabulary in the academic environment

Vocabulary is fundamental in the educational process, as it directly influences the progress of learning a foreign language and the subjects taught in that language, as is the case in Content and Language Integrated Learning (CLIL) approaches (Coyle, Hood, & Marsh, 2010; Dalton-Puffer, 2007; Pérez-Vidal, 2009). Learning an adequate vocabulary not only makes basic communication easier, but also, helps to understand and build knowledge in different academic contexts. If we do not have a solid lexical base, our development in linguistic and cognitive skills is limited, which can affect our competence in both the language we are learning and the content being taught (Nation, 2001; García Mayo & Zeitler, 202). In this context, it is relevant to make a distinction between learning in immersion environments and in home contexts.

In an immersion environment, vocabulary is acquired more naturally, thanks to continuous exposure to the language in authentic and meaningful situations (Cummins, 2000; Muñoz, 2012). On the other hand, in home contexts, where foreign language use is limited to specific situations, a more structured and explicit approach to teaching vocabulary is required. In this case, didactic planning should focus on prioritizing lexis that allows students not only to function in everyday conversations, but also to effectively access academic content (Gajo, 2007; Fernández Fontecha, 2024).

Therefore, understanding vocabulary as a priority educational need is key, since its development transversally affects success in language learning and related disciplines. Approaching lexical instruction in an intentional and contextualized manner enhances students' communicative and academic competence, facilitating their progress in multilingual and multicultural environments.

2.1. The academic task

Having a well-defined vocabulary is crucial in the academic world, especially in assessment and accreditation tasks because it ensures uniformity and accuracy in measuring language proficiency. Today, in a world where language learning and certification are key to academic and professional mobility, it is essential to have a set of basic words to serve as a reference in assessment tests (Hammrich, 2025; Council of Europe, 2001). This vocabulary should include terms that are the most common and necessary words for effective communication, which would allow for a fairer and more equitable assessment of language proficiency levels (Paul, 2019).

Although it is obvious that we need a unified vocabulary, we still do not have an agreed tool that clearly establishes which key words should be known by those assessed at each level (A1, A2, B1, etc.). This lack of uniformity generates notable differences between the various assessment tests, as each institution or entity can choose its own words according to its own criteria (Green, 2013; Milanovic, 2009). As a consequence, a proficiency level can be assessed in very different ways, which creates inequalities and makes it difficult to compare results.

If we had a unified fundamental vocabulary, assessment tests would be much more consistent and objective, as all assessors would be using the same lexicon. This would also help make language proficiency certificates more reliable and internationally standardized (Little, 2011; INAES, 2024). Without this tool, the accreditation process becomes less clear and equitable, affecting not only students and professionals seeking to validate their language level, but also institutions that need clear and consistent criteria for assessing these skills. Examining this would also help improve educational materials, allowing them to be designed in a more focused and effective way, which would reduce the dispersion of content and ensure that learning is focused on words and expressions that really matter for understanding the language (Nation & Webb, 2011).

3. Conceptualisation

We come from a tradition where vocabulary learning has often focused on lists of words. Many have used the semiological dictionary as a support tool, either in a general or more specific way. To understand it today, it is essential that we are willing to challenge old rules and traditions. It is important to recognise that vocabulary is not simply a collection of words, but is part of a dynamic system that is constantly changing (Béjoint, 2010).

One of the first steps in conceptualising vocabulary is to identify its component units. This type of vocabulary is not limited to single words; it also encompasses larger units, such as collocations and phraseological structures, which are key to facilitating the most important communicative interactions (Sinclair, 1991; Higueras, 2016). This is a necessary understanding if we are to achieve the first objective: improving communicative quality.

3.1. Quantitative and qualitative profitability

When we make decisions about words, combinations and so on, this remains a common thread throughout the process. In making these choices, it is essential to prioritise empirical questions, valuing their usefulness as a tool rather than remaining anchored in rigid rules or concepts (Tono, 2001).

Quantification has been implicit in every lexical project over time. It has been a commercial resource and has guaranteed the quality of the work, among other things. Presenting its nomenclature in a denatured form and in alphabetical order is commonplace (Hartmann & James, 1998). In this context, it is presumed that there is a careful analysis of their cost-effectiveness, both on a large scale and in detail. We do not quite understand how we can consider a useful vocabulary simply as a list of words, described in an article, exemplified and completed with a couple of usage features (Hausmann, 1990). When a lexical inventory, whether small or large, general or specific, is presented without further information, it leads us to question whether communicative profitability has really been the mainstay of that compilation (Lew, 2013).

We want to emphasise that the concept of cost-effectiveness is not new. It has been an aspect that has come up frequently in dictionary writing (Bergenholtz & Tarp, 2003). This leads us to reflect in several directions:

Not everything has happened lexicography. Applied Linguistics in Spanish Language Teaching has also played an important role. The principle of communicative affordance has been central since its inception (Carter, 2012). However, it is discouraging to see realities such as the lexical inventory that accompanies the Cervantes Institute's Curriculum Plan, which has not been updated since 2006 and whose lexicon deserves an in-depth analysis and a detailed review due to its importance (Vine Jara et. alii. 2016).

3.2. Lexical centrality and core vocabulary

Lexical centrality is not only about the frequency with which certain words are used; it also refers to how certain words function as *key nodes* in the semantic network of a language. From a more technical approach, we can analyse lexical centrality using network models. The words that have the highest centrality are those that are most connected to others, making them essential for the language to function (Ferrer-i-Cancho & Solé, 2001; Liu, 2008; Solé et al. 2010). These studies show that the vocabulary of a language is not just a random collection of terms, but an interconnected structure where certain words play a fundamental role in conveying and understanding meaning.

Words that have high centrality often share specific characteristics, such as polysemy, which is the ability to have several meanings or to be used in different contexts (Bybee, 2011). A good example is the verb hacer, which is not only common because of its frequency, but also because it can be combined with a wide range of nouns and adjectives to create expressions with diverse meanings (making the bed, asking a question, playing sport). This shows its flexibility both syntactically and semantically (Tamariz, 2011).

Moreover, studies on lexical centrality have found that these words tend to be more resistant to linguistic change. While more specialised or less commonly used vocabulary may evolve or even disappear, core words tend to remain stable because of their role in language structure (Pagel et al. 2007). This phenomenon has been documented in different languages and shows that the most central and frequent terms retain their form and function over time.

These words are also easy to learn, especially for children who are acquiring language and people who are studying a foreign language (Ellis, 2002). In fact, it has been observed that second language learners tend to incorporate words with higher centrality first, as these facilitate access to key syntactic and semantic structures (Nation, 2001).

Lexical centrality relates to the ability to anticipate what is coming next in language processing. In a text, central words help us anticipate more likely terms to follow, which improves our language comprehension and production (Christiansen & Chater, 2018). This is critical for fluency in communication, as much of human language is based on patterns and structures that we can predict (MacWhinney, 2005). This concept is closely linked to the classification of thematic and athematic words, which distinguishes between terms with strong semantic content and those with more abstract grammatical functions (Forest, 2015).

3.3. Vocabulary and interdisciplinarity

In order to develop a core vocabulary of Spanish, collaboration between various disciplines is key. This ensures that the words selected are not only frequently used but are also relevant and representative of the various contexts in which the language is used.

Linguistics provides the grammatical and semantic analysis necessary to define how words function within the language system. *Sociolinguistics* helps to understand how language is used in different social groups and communication contexts (Labov, 2001). At the same time, *psycholinguistics* explains how people process and acquire words effectively, an aspect that is fundamental to vocabulary learning (Aitchison, 2012). In this regard, Peronard's (2005) research highlights the importance of metacognitive processes, which enable individuals to manage their own cognitive strategies when learning a language.

One of the key themes is *language didactics*, which transforms knowledge into pedagogical tools that actually work. Through this perspective, we seek to improve the teaching of vocabulary in the classroom, ensuring that the words we choose are useful and applicable to developing communicative skills (Celce-Murcia, 2001). This is where lexical statistics come into play, providing us with objective data on the frequency of use and distribution of vocabulary in different contexts and forms of communication (Nation, 2001).

In addition, the development of core vocabulary has been enhanced by advances in *language technologies*, which have facilitated automatic lexical analysis and the creation of useful digital resources for both teachers and learners (Jurafsky & Martin, 2021). With these technological tools, it is possible to identify usage patterns in large linguistic databanks and develop interactive applications for language teaching (Manning, 2022).

Together, all these disciplines help to make the core vocabulary of Spanish practical, representative and culturally relevant. This provides a solid foundation for both learning and using the language in everyday life.

4. Fundamental vocabulary. The tools

4.1. Introduction. Practical applications of lexical analysis

A detailed vocabulary analysis, together with lexicometric tools, has applications in different areas, such as linguistics, technology and education.

In teaching, lexicometric analysis can be really useful for designing more effective educational curricula. By identifying the most frequent and relevant words in different contexts, teachers can focus on vocabulary that really has an impact on everyday and professional communication. For example, a study on teenagers' digital writing on WhatsApp used lexicometric analysis to study their linguistic and non-verbal communication particularities, showing how this methodology helps to understand language use in specific contexts (Vázquez et. alii., 2015). In addition, research at university level has applied this technique to analyse students' conceptions of learning, revealing significant differences in their approaches.

In the development of technologies with linguistic involvement, lexicometrics also plays a key role. Tools such as machine translators, speech recognition systems and virtual assistants require robust lexical databases to function well. The ability to analyse lexical and semantic patterns on a large scale helps to improve the accuracy of these technologies, making them more useful and accessible to speakers of different variants of Spanish. In fact, some studies have used text mining techniques to identify thematic foci in scientific research, underlining the usefulness of lexicometric analysis in organising large volumes of information. Moreover, lexicometrics has helped to redefine concepts in productive pedagogical projects, contributing to a better organisation of knowledge and identifying trends in the educational field (Romero et ali. 2018).

4.2. Lexicometric tools in vocabulary analysis

These tools allow the examination of large amounts of textual data, helping to identify patterns of usage, frequency and lexical distribution that would otherwise be difficult to see. With techniques such as data mining and frequency analysis, we can discover a variety of phenomena related to the most common words in the language (Biber, 2011). In addition, lexicometric tools allow us to go deeper than simple word counts and explore more complex relationships within the lexicon. For example, co-occurrence analysis can show us how words are combined in different contexts (Gries, 2021), and computational semantic models can help us understand how the meanings of terms change over time (Hamilton, Leskovec & Jurafsky, 2016). All this not only gives us a clearer picture of current vocabulary, but also opens up the possibility of diachronic studies that allow us to follow the evolution of Spanish in a globalised world.

4.2.1. Corpus

Corpora have traditionally been tools which bring together a large volume of texts for analysis for different purposes. In the case of the lexicon, they can provide interesting data on lexical units and how they are used, allowing them to be placed within a context from collected textual samples (Sinclair, 1991). They are essential resources in lexical projects as they provide processed information that facilitates vocabulary search and learning (Tognini-Bonelli, 2001).

The massive availability of online texts has facilitated the building of large-scale linguistic corpora through automated processes of collecting and downloading materials. This possibility represents a considerable advance in linguistic research, allowing the efficient and agile analysis of huge volumes of textual data. However, when corpus compilation is based on automatic downloading without strict selection criteria, problems related to representativeness, balance and diversity of the corpus may arise, which in turn compromises the validity and applicability of the results.

Furthermore, it should be made clear that the corpus is neither a direct tool for lexical acquisition nor the only means of obtaining the key vocabulary we are analysing. Their complex way of processing information, data encoding and often unintuitive interfaces make them more suitable for linguistic research than for consultation, learning and teaching for people who are not specialised in metalinguistic tasks (McEnery & Hardie, 2012).

Another issue to consider is the nature of the results, which is related to the selection of the sources from which the samples were drawn. In general, a corpus works with a limited number of sources: oral, written, press, etc. This aspect does not detract from its value, but it is important to take it into account when assessing how representative your data are (Biber, 2011).

Finally, a third drawback that we may encounter, especially nowadays with the rise of new technologies and social networks, is the reliability of the material collected. Nowadays, we can carry out large-scale studies thanks to the ease with which information can be extracted and collected from digital sources (Kilgarriff, 2012). However, this also raises questions about the quality of the information collected, as the digital world can be full of hearsay and unverified data, which affects the quality of linguistic samples (Lindquist, 2009).

Several platforms have emerged that offer extensive and specialised corpora, useful for exploring different linguistic phenomena. A prominent example is the Corpus del español, which includes a web-based corpus (CdE-web) with around 2 billion words, and another focused on recent news (CdE-NOW), which exceeds 5 billion. For its part, the Sketch Engine tool provides access to Spanish corpora such as esTenTen18, as well as corpora in dozens of languages, some of which reach sizes close to 20 billion words. In the case of English, the enTenTen20 corpus even reaches 40 billion words.

These infrastructures represent fundamental resources for contemporary linguistic analysis. However, their optimal use requires a critical look at their methodological limitations, particularly with regard to the criteria for selecting, cleaning and categorising the texts included.

4.2.2. Frequency indices

Frequency indices are essential tools when it comes to descriptive statistics and data analysis. These metrics help us to better understand how data are distributed in a given set, facilitating the identification of patterns and trends that, at first glance, might go unnoticed (Biber et al. 2020). Its usefulness is broad, spanning across disciplines such as sociology, economics, psychology and biology, among others (Gries, 2021). In simple terms, a frequency rate tells us how many times a particular event occurs, and this can be expressed as absolute frequency, cumulative frequency or relative frequency (McEnery & Hardie, 2012).

In the field of corpus linguistics, frequency indices have been used to identify the most representative words of a language to create dictionaries that are very useful in language teaching (Kilgarriff et al. 2012). More recently, it has been found that combining frequencies with semantic distribution models can enrich the analysis of lexical patterns in different discourse contexts.

Analysing frequency indices can provide valuable information for decision-making and data interpretation. For example, in market research, these indices can reveal which products or services are consumers' favourites. In education, they help us discover which topics are most frequently covered in exams (Nation, 2013). And in scientific research, they allow us to examine the frequency of certain characteristics or phenomena within a sample (Brezina, 2018). However, care must be taken not to build a fundamental vocabulary on the basis of frequency rates alone, something we have already discussed with regard to corpora. The documentary sources from which these indices are obtained may lead to erroneous conclusions about the empirical values of a core lexicon. Moreover, although manipulating this information is very useful in research, it is not necessarily a practical tool for the common language user, who is likely to have limited knowledge of the subject (Tognini-Bonelli, 2001).

4.2.3. Lexical availability

For decades, one of the topics of study in statistical lexical analysis has been lexical availability. In the 1950s, Gougenheim, Michéa and Sauvageot developed a research approach focusing on the lexicon available among French speakers, with the aim of creating a book entitled *Français Élémentaire*. This work was aimed at teaching French in former colonies (Gougenheim et al. 1956). Their research was based on an initial criterion of frequency of use, which gives a practical character to the results. From his findings, we are left with the ideas of athematic word and thematic word, which we mentioned at the beginning of this text.

Thematic words are words that are directly linked to a specific topic or group of concepts. They are usually nouns, adjectives or verbs that have a key meaning in a particular context. In contrast, athematic words are somewhat more general; they are not tied to a specific topic. They tend to be commonly used terms that can appear in many contexts. Words such as “thing”, “do” or “good” are considered athematic, as their meaning is not restricted to a particular domain (Bartol, 2006). The study of lexical availability seeks to identify how many units make up lexical centrality in a given group of speakers. To this end, empirical research is carried out in order to find out a person’s active vocabulary. Thus, we are not only assessing a speaker’s lexical richness, but also his or her ability to process and activate vocabulary. This implies that there are other factors at play than just language learning (álvarezHerranz and Gómez, 2022). An example of this is the Pan-Hispanic Lexical Availability Project, developed by López Morales in 2011, which focused on the lexical analysis of a school population, closely linked to the academic environment. Most of the studies were conducted in educational settings, which could give an idea of the lexical richness of students.

Lexical availability provides valuable information. The nuclearity value of many of its units, obtained through different extraction methods and tools, helps us to identify terms that could form part of a core lexicon (González, 2014).While it is important to recognise the advantages of these projects, it is also crucial to reflect on their usefulness in relation to our specific objectives.

The value of data as a sample of a language is quite relative. This varies depending on the tools used and the characteristics of the person being analysed. All such studies are affected by how the population is defined, which may be influenced by factors such as age, social group or other criteria. On the other hand, they tend to focus on an active vocabulary that covers a limited range of topics, resulting in inventories that are not fully complete (Moreno, 2022).

Another important point is the way in which the results are obtained, which is deeply connected to the data collection process. The goal here is to identify the words that the brain activates when talking or thinking about a specific topic. This often leads to the collection of very simple lexical units, which restricts lexical production by not considering more complex word combinations or units.

Finally, the very nature of the fields of interest, which are often somewhat outdated, may result in data that are not useful for projects that need the lexical units to be relevant and current. Here the dilemma arises as to whether the topics of interest really reflect the social reality in which the study is being carried out, or whether the study is instead adapted to a social reality that already existed (Reyes et. alii. 2021).

4.2.4. Other work

a. Curricular Plan of the Instituto Cervantes

It includes an inventory of lexical units which should not be considered as a fundamental vocabulary of Spanish in its broadest sense. The Plan, in its section on communicative notions and functions, provides a list of words and expressions organised by theme, but these do not represent an exhaustive and adequate core vocabulary to guarantee full communicative competence in Spanish.

The inventory of lexical units in the curriculum is based on a series of notions of language use, such as family, work or health, which correspond to topics that students should know according to their level of learning. However, these lists are incomplete and do not reflect the full range of vocabulary that learners really need in order to develop a solid linguistic competence.

Core vocabulary, in a broader sense, should include all words and expressions commonly used in different registers, genres and communicative situations. However, the Curriculum Inventory is limited in several respects. First, it is an inventory that prioritises the use of standard and academic terms, leaving out colloquial expressions, regionalisms, jargon and other important elements of real language that are used on a daily basis in various contexts, such as in informal social interactions or in more technical situations.

In addition, the curriculum inventory is also outdated in many cases, as it does not include neologisms, technological terms and words that are relevant in today's social, political and cultural context. In the contemporary world, vocabulary related to new technologies, social networks, digitalisation and other developments is underrepresented, leaving students without the necessary tools to engage in conversations about these topics. This inventory reflects a limited view of the language that is not fully adapted to the real communicative needs of learners, nor to the diversity of contexts and situations faced by speakers of Spanish in today's world.

b. Lexical nomenclatures in the Chinese education system

The Gaokao (高考), or National University Entrance Examination in China, is a standardised assessment that determines access to higher education for millions of students. Although most opt for English as a foreign language, Spanish is also offered as an option, which has led to the creation of specific lexical inventories for this test.

However, these Spanish lexical inventories used in the Gaokao have not been updated in decades, posing significant challenges. Lack of renewal may result in teaching and assessment that do not reflect contemporary language use, affecting learner readiness and test validity.

The need to update these inventories is clear. The Spanish language has undergone significant changes in recent decades, and it is crucial that teaching and assessment materials reflect these developments. In addition, an update would allow for better cultural and linguistic adaptation for Chinese learners, facilitating a deeper and more relevant understanding of the language.

In short, updating the lexical inventories of Spanish in the Gaokao is essential to ensure fair and effective assessment. This would not only benefit learners but also strengthen educational and cultural ties between China and Spanish-speaking countries.

5. Conclusions

Despite advances in applied linguistics and language technologies, there is still a notable absence of a consolidated and universally accepted proposal as to which are the most recurrent and relevant words in Spanish. This contrasts with other languages, such as English, which have developed widely disseminated lists of essential vocabulary, such as the *General Service List or the Academic Word List*. In the case of Spanish, this shortcoming reflects not only a technical or methodological issue, but also the complexity of the internal and external factors involved in the development of such a proposal.

The scarcity of large-scale comparative studies that comprehensively address the use of Spanish in different countries and contexts is a reality. Although linguistic corpora do exist, such as the RAE's Corpus del Español del Siglo XXI (CORPES XXI), they have not yet been used consistently to establish lists of essential words. Efforts to date are often fragmented, either because they focus on a specific audience (e.g. learners of Spanish as a second language) or because they are limited to analysing data from a particular region or variant of the language. The lack of collaboration and standardisation between projects limits the possibility of building a comprehensive and widely accepted approach.

Furthermore, the construction of a consolidated list of essential vocabulary in Spanish requires a combination of sources, methodological approaches, etc. that have not always been applied in a coordinated manner. It is necessary to analyse the frequency of words in the language with the consequent updates, which implies processing large volumes of linguistic data of different nature, a complex work necessarily supported by data engineering. On the other hand, it is also essential to consider factors such as the usefulness of words in specific communicative contexts, their ability to generate or understand other related terms, and their relevance as an educational or cultural resource. Integrating these dimensions is not a simple task and requires considerable technological and human infrastructure.

Added to this is the fact that the teaching of Spanish, both as an L1 and as an L2 or FL, has not always had a unified approach to vocabulary. Although there are reference frameworks such as the *Cervantes Institute's Curriculum Plan or the Common European Framework of Reference for Languages (CEFR*), these do not provide exhaustive or up-to-date lists of vocabulary, but rather general guidelines on levels of proficiency. This contrasts with other languages, such as English, where basic vocabulary lists are more integrated into teaching systems. The lack of a consolidated tradition in this area has slowed down efforts to define an essential Spanish vocabulary that is practical and applicable in educational contexts.

The constant evolution of the language also represents a significant challenge. Spanish, like any living language, is constantly changing, incorporating new words and expressions as others fall into disuse. This dynamism makes it difficult to create a list that remains relevant over time. In addition, the influence of technology and global media has accelerated the incorporation of linguistic borrowings and neologisms, further complicating the task of determining which words should be included in an essential list.

Finally, it is important to consider the interests and perspectives of the academic and cultural institutions responsible for promoting Spanish. The *Royal Spanish Academy* (RAE) and the language academies of Spanish-speaking countries have played a key role in the standardisation of the language, but their traditional focus has been more on grammatical and orthographic issues than on the definition of basic vocabularies. At the same time, educational institutions and language technology developers have made progress in creating specific resources, but these efforts are often isolated and lack the collaboration necessary to generate a consolidated approach.

In this context, the lack of a universal list of the most recurrent and relevant words in Spanish is not the result of a lack of interest or effort, but rather of the complexity of the task itself as well as of the language and the multiple factors that influence its use. Overcoming this challenge requires a collaborative approach that integrates academic, technological and educational institutions, and that is able to address the diversity of Spanish in an inclusive and representative manner.

At first glance, this may seem an unattainable task. Seen from the perspective of time, the publishing world or the linguistic reality, we may think that we are facing something utopian. Perhaps it is time to change the focus of lexicographic works. We are not dealing with a dictionary, we are not producing a text for students, we are not going to create a corpus, we are not trying to sell anything. We are dealing with a research project, an empirical work born from the careful and conscientious study of institutions that do not intend to make a profit. We do not do what has already been done, a phenomenon all too common in research in our speciality.

We are aware of the need to undertake this work, there are realities in the world of Spanish that demand it, and this is how the interest in this project was born. We still do not have this fundamental vocabulary, but we continue to evaluate, accredit and test speakers who demand these certifications as a condition for progressing in their studies, in their work; it is not a trivial objective. Repeating projects or tools that have already been developed would be to continue with the real demand. We have to ask many questions, we have to think about the social and not the economic value, and we have to work in a multidisciplinary way to be able to walk along this path. You have to make mistakes, you have to keep looking for answers, all of which will form the skeleton of the final work.

In his 1997 work, Freeman Dyson introduces an important distinction in the history of science, differentiating between two types of scientific revolutions: conceptual revolutions and instrumental revolutions. Conceptual revolutions: Dyson describes conceptual revolutions as changes in our understanding of scientific phenomena that redefine the way we interpret existing data and theories. These revolutions focus on the introduction of new ideas or paradigms that offer a fresh perspective on familiar issues. Instrumental revolutions refer to significant changes in science and technology that result from the development and introduction of new tools and techniques. These revolutions do not focus on changing our theoretical understanding of scientific phenomena, but on expanding our abilities to observe, measure and experiment with the natural world. Advances in scientific tools and techniques often lead to unexpected discoveries and the opening up of new areas of research.

6. References

Aitchison, J. (2012). *Words in the mind: An introduction to the mental lexicon*. John Wiley & Sons.

Bartol Hernández, J. A. (2006). La disponibilidad léxica. *Revista Española de Lingüística*, 36 / 1, 2006, (pp. 379-384).

Béjoint, H. (2010). *The lexicography of English*. Oxford University Press.

Bergenholtz, H. & Tarp, S. (2003). Two opposing theories: on H.E. Wiegand's recent discovery of lexicographic functions. *Lexicographica, 19*, pp. 1–20.

Biber, D. & Gray, B. (2020). Corpus-based discourse analysis. In K. Hyland and B. Paltridge (eds.) *The Continuum Companion to Discourse Analysis*. Continuum, pp. 97-110.

Biber, D. & Reppen, R. (eds.). (2011). *Corpus Linguistics; volume I: lexical studies*. Sage.

Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Bosque, I. (2015). *Las categorías gramaticales*. Síntesis.

Brezina, V. (2018). *Statistics in corpus linguistics. A practical guide*. Cambridge University Press.

Brown, A.V., Paz, Y.B. & Brown, E.K. (2021). *El léxico-gramática del español: Una aproximación mediante la lingüística de corpus*. Routledge.

Bybee, J. (2003). *Phonology and language use*. Cambridge University Press.

Bybee, J. (2011). *Language, Usage and Cognition*. Cambridge University Press.

Carter, R. (2012). *Vocabulary: Applied Linguistic Perspectives*. Routledge.

Celce-Murcia, M. (2001). *Teaching English as a second or foreign language*. Heinle & Heinle.

Christiansen, M. H. & Chater, N. (2018). *Creating Language: Integrating evolution, acquisition, and processing*. MIT Press.

Council of Europe. (2001). *Common european framework of reference for languages: learning, teaching, assessment*. Council of Europe Publishing.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34* (2), pp. 213-238.

Coyle, D., Hood, P., & Marsh, D. (2010). *CLIL: Content and language integrated learning*. Cambridge University Press.

Croft, W. & Cruse, D. A. (2004). *Cognitive linguistics*. Cambridge University Press.

Cummins, J. (2000). *Language, power and pedagogy: bilingual children in the crossfire*. Multilingual Matters.

Dalton-Puffer, C. (2007). *Discourse in content and language integrated learning (CLIL) classrooms*. John Benjamins.

Dang, T. N. Y. & Webb, S. (2016). Evaluating lists of high-frequency words. *International journal of applied linguistics,* 167/2, pp. 132-158.

Ellis, N. C. (1994). *Implicit and explicit learning of languages*. Academic Press.

Ellis, N. C. (2002). Frequency effects in language processing. *Studies in Second Language Acquisition, 24* (2), (pp. 143-188).

Fernández Fontecha, A, Jiménez Catalán, R. M. & James Ryan, J. (2024). Lexical production and organisation in L2 EFL and L3 EFL learners: a distributional semantic analysis of verbal fluency. *International journal of multilingualism,* 21/1, (pp. 60-77).

Ferrer-i-Cancho, R. & Solé, R. V. (2001). The small world of human language. *Proceedings of the Royal Society B: Biological Sciences,* 268 (1482), (pp. 2261-2265).

Flege, J. E. (1995). Second language speech learning: theory, findings, and problems. En W. Strange (ed.), *Speech perception and linguistic experience: issues in cross-language research*, pp. 233-277) York Press.

Gajo, L. (2007). Linguistic knowledge and subject knowledge: How does bilingualism contribute to subject development? *The international journal of bilingual education and bilingualism,* 10 (5), pp. 563-581.

García Mayo, M. P. (2021). Vocabulary, corrective feedback and intervention studies. *Language teaching research*, 25 (2), pp. 153-158.

Gougenheim, G., Michéa, R., Sauvageot, A. & Rivenc, P. (1956). *L'élaboration du français élémentaire*. Didier.

Green, A. (2013). *Exploring Language Assessment and Testing: Language in Action*. Routledge.

Gries, S. T. (2021). *Quantitative corpus linguistics with R: a practical introduction*. Routledge.

Halliday, M. A. K. & Hasan, R. (1976). *Cohesion in English*. Longman.

Hamilton, W. L., Leskovec, J. & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1489-1501.

Hartmann, R. R. K. & James, G. (1998). *Dictionary of Lexicography*. Routledge.

Hausmann, F. J. (1990). The dictionary as an encyclopedic inventory of words. In F. J. Hausmann, O. Reichmann, H. Wiegand, E. & Zgusta, L. (Eds.), *Wörterbücher. Dictionaries. Dictionnaires*. De Gruyter, pp. 1811-1835.

Herranz Llácer, C. V. & Gómez-Devís, M.ª B. (2022). La investigación en disponibilidad léxica infantil: aplicaciones para la enseñanza de ELE. *Cultura*. 28. pp. 83-101.

Higueras García, M. (2016). *La enseñanza de colocaciones léxicas en español como lengua extranjera: teoría y práctica*. Biblioteca Virtual Miguel de Cervantes. https://www.cervantesvirtual.com/nd/ark:/59851/bmcq83b9

Instituto Nacional de Educación Superior [INAES]. (2024). *Normativas para la evaluación de competencias lingüísticas*. INAES.

Jurafsky, D. & Martin, J. H. (2021). *Speech and language processing*. Pearson.

Kilgarriff, A. (2012). *Getting to know your corpus*. In S. Gries & J. Newman (Eds.), *Corpus-based approaches to Construction Grammar* (pp. 35-51). John Benjamins.

Labov, W. (2001). *Principles of Linguistic Change: Social Factors*. Wiley-Blackwell.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Laufer, B. & Nation, P. (1995). Vocabulary Size and Use: Lexical Richness in L2 Written Production. *Applied Linguistics*, 16, pp. 307-322.

Lew, R. (2013). Online dictionaries of English. *International Journal of Lexicography, 26* (2), pp. 189-210.

Lindquist, H. (2009). *Corpus linguistics and the description of English*. Edinburgh University Press.

Little, D. (2011). The Common European Framework of Reference for Languages: A research agenda. *Language Teaching, 44* (3), pp. 381-393.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9* (2), pp. 159-191.

López Morales, H. (2011). *Proyecto Panhispánico de Disponibilidad Léxica*. Asociación de Academias de la Lengua Española.

MacWhinney, B. (2005). A Unified Model of Language Acquisition. *Handbook of bilingualism: Psycholinguistic approaches*. 4967, pp. 50-70.

Manning, C. D. (2022). *Foundations of Statistical Natural Language Processing*. MIT Press.

Serrano Zapata, M. & Calero Fernández, Mª. A. (eds.) (2021). *Aplicaciones de la disponibilidad léxica*. Tirant Humanidades.

Marrero. M. & Palacios., Y. (2023). Las relaciones lexicales: Un acercamiento desde la polisemia y la homonimia. *Luz*, 22 (4), pp. 106-121.

McEnery, T. & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Milanovic, M. (2009). *Cambridge English exams - the first hundred years: a history of English language assessment from the University of Cambridge, 1913–2013*. Cambridge University Press.

Milton, J. (2009). *Measuring second language vocabulary acquisition*. *Multilingual matters*. De Gruyter.

Muñoz, C. (2012). The significance of intensive exposure as a turning point in learning a foreign language. En C. Muñoz (Ed.), *Intensive exposure experiences in second language learning* (pp. 1-11). Multilingual Matters.

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

Nation, I. S. P. & Webb, S. (2011). *Researching and analyzing vocabulary*. Heinle Cengage Learning.

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge University Press.

Pablo Núñez, L. (2019). Metodologías para la enseñanza del léxico en el aprendizaje de lenguas extranjeras: un recorrido histórico. *Foro de profesores de E/LE*, 15, pp. 161-177.

Pagel, M., Atkinson, Q. D. & Meade, A. (2007). Frequency of word-use predicts rates of lexical evolution throughout Indo-European history. *Nature,* 449 (7163), pp. 717-720.

Pérez-Vidal, C. (2009). The integration of content and language in the classroom: A European approach to education (the second time around). En E. Dafouz & M. Guerrini (eds.), *CLIL across educational levels: Experiences from primary, secondary and tertiary contexts* pp. 3-16. Richmond Publishing.

Peronard, M. (1998). La metacognición como herramienta didáctica. *Revista Signos: estudios de lingüística, 57*, pp. 61-74.

Plag, I. (2003). *Word-formation in English*. Cambridge University Press.

Robin, S. J., & Aziz, A. (2022). The Use of Digital Tools to Improve Vocabulary Acquisition. *International Journal of Academic Research in Business and Social Sciences*, 12 (1), pp. 2472–2492.

Romero-Pérez, I., Alarcón-Vásquez, Y. & García-Jiménez, R. (2018). Lexicometría: enfoque aplicado a la redefinición de conceptos e identificación de unidades temáticas. *Biblios*, 71, pp. 68-80.

Sánchez-Saus Laserna, M. & Álvarez Torres, V. (2024). Influence of learning contexts on the mental lexicon: lexical productivity and semantic networks in SFL students. *Revista De Lingüística y Lenguas Aplicadas*, 19, pp. 204–217.

Schmitt, N. (2008). *Review article: Instructed second language vocabulary learning*. *Language Teaching Research,* 12 (3), pp. 329-363.

Schmitt, N. (2019). *Vocabulary in language teaching*. Cambridge University Press.

Serrano-Dolader, D., Martín Zorraquino, M. A., & Val Álvaro, J. F. (coords.). (2009). *Morfología y español como lengua extranjera (E/LE)*. Universidad de Zaragoza.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Solé, R. V., Corominas-Murtra, B., Valverde, S. & Steels, L. (2010). Language networks: Their structure, function, and evolution. *Complexity,* 15 (6), pp. 20-26.

Tamariz, M. (2011). The evolution of linguistic complexity: The case of word frequencies. *Language and Cognition,* 3 (2), pp. 141-175.

Taylor, J. R. (2003). *Linguistic categorization*. Oxford University Press.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.

Tono, Y. (2001). *Research on dictionary use in the context of foreign language learning: Focus on reading comprehension*. Max Niemeyer.

Vine Jara, A. & Ferreira Cabrera, A. (2016). Propuesta de un modelo para una prueba con fines específicos académicos en ELE. *Literatura y Lingüística*, 33, pp. 369-390.

Webb, S. (2020). *The Routledge handbook of vocabulary studies*. Routledge.

Webb, S. & Nation, P. (2017). *How vocabulary is learned*. Oxford University Press.