

*Individual Lexical Breadth and its associated measures. A contribution to the calculation of individual lexical richness**

1. Introduction

I

Studies on lexical richness conducted to date mainly focus on written texts, as their linguistic variety can be accessed with relative ease (Jarvis 2002; Malvern and Richards 2000; Malvern et al. 2004; Vermeer 2004). Basing such studies on spoken productions usually involves a more complex process, due to the highly individual nature of speech. Whilst some studies have concentrated on controlled lexical productions of children and adolescents (Ávila 1988; Verhoeven, van Leeuwe, and Vermeer 2011), others have developed theoretical indices based on algorithms that predict the lexical richness of an entire city (Ávila-Muñoz 2016). Currently, several parameters are being used for determining the concept of 'lexical richness', among others, content density, lexical sophistication, lexical accuracy and lexical diversity (Jarvis 2013: 13-44; Castañeda Jiménez and Jarvis 2013: 498-513).

In this paper, we present an original model for the measurement of the proposed concept of 'individual lexical breadth'. With this concept we try to represent lexical richness on a solid and non-intuitive basis so that a particular subject's lexical breadth is understood as the quantitative reflect of the amount of richness he or she has. Thus, lexical breadth is the quantitative, statistically proved, characteristic of lexical richness. Such a model is based on both size and specificity of the vocabulary that speakers produce on lexical tests sessions—i.e their available lexicon. On the basis of this theoretical model, several measurement indexes are

* This work has been made possible thanks to the support of the Spanish National Research Agency Project “Agenda 2050: The Spanish of Malaga: Processes of Spatial and Social Variation and Change” (PID2019-104982GB-C5-2). A preliminary Spanish version of some parts of this chapter was published in Ávila & Villena (2010), 37-81 and 177-286.

proposed, notably, Index of Lexical Decentralization (ILD) and Lexical Amplitude (ILA). Our idea is that these indexes can be used as complementary to the above-mentioned parameters for calculating lexical richness.

Our model of individual lexical breadth incorporates a quick, simple method and uses data collected via word association tests. The model is based on a revised interpretation of the traditional and well-known concept of “lexical availability” and has been built on rigorous methodological foundations from the field of socio- and psycholinguistic research (see Ávila-Muñoz 2016). It has proved useful, despite the simplicity with which it allows for linguistic information to be extracted (López Morales 1999; Hernández-Muñoz, Izura, and Ellis 2006; Ávila-Muñoz and Villena-Ponsoda 2010: 177-279).

II

Research on lexical availability originally dates back to 1951, when the French Ministry of National Education appointed a special committee for bringing about a basic lexicon with teaching purposes (Gougenheim et al. 1954). Apparently, the primary objective was obtaining a list of basic words to be accessible for students of French as a foreign language. However, the pioneering researchers on lexical availability were very soon aware that previous work was merely based on frequency. To obtain the word frequency of a particular language, linguists use methods that favor accessing the statistically most stable lexicon. In other words, lexical frequency lists basically include the words speakers use when building a message regardless of the topic the message is about. As is well known, we tend to use a reduced number of lexical and structural units which are, of course, the most frequent. In fact, the variable “frequency of use” is one of the most relevant factors on language processing both in production and decoding. Furthermore, frequency dictionaries exist for the main western languages and are broadly used. For instance, word frequency lists are very useful for native and foreign language teaching (Davies 2005; Lonsdale and Le Bras 2009).

Nevertheless, frequency lists include, almost exclusively, “athematic” words, i.e., words without a true lexical meaning—grammatical words—whilst “thematic” words with lexical content are missing or, at least, quantitatively insignificant. These thematic words are, however, very common in every speaker's language performance whenever the required communicative conditions for their use are fulfilled.

“Thematic” words are related to specific topics, refer to entities and objects and are, mainly, nouns. Additionally, these *thematic* words frequently cooccur and, almost always, tend to arise whenever certain topics are addressed (de Kock 1983: 59–60).

Then, available dictionaries can turn out to be adequate complements to balance frequency repertoires—excessively biased towards non-thematic lexicon—since they reflect the lexical flow as used by speakers in specific communicative situations. Such lists are based on the concept of “frequent situation”, which is quite different from the one underlying lists of “frequent words”. Thus, the idea of the available lexicon is founded on the motto that certain very commonly used words in a particular language are related to and constrained by the occurrence of certain topics.

To obtain the available vocabulary—which, as stated, is *not* the most frequent—we start carrying out association test sessions based on a certain number of centres of interest (CI) around which specific related vocabulary emerges. It is assumed that this vocabulary is the speakers' potential lexicon included in their active lexical competence, and they use it whenever a particular topic turns up during conversation. The overall number of words within the list of available vocabulary for a specific CI is constrained by the speakers' reaction time—two minutes for each association stimulus as far as both our own research and background are concerned (Samper Padilla 1998). It seems apparent that words that come the earliest to the memory will be the most available and, hence, the ones occupying the first positions in the

availability lists. In sum, speakers' available lexicon is part of their mental lexicon but it does not occur in everyday linguistic interactions unless a specific topic does establish so.

To obtain available lexicon lists, a mathematical formula based on the so-called Lexical Availability Index (LAI) is used. The LAI is a numeric parameter that relates frequency and ranking. The mathematical calculation the LAI is based on weights frequency of a certain word's occurrence in a specific CI and its position in the corresponding list of lexical availability. The most available words—i.e., those attaining LAI scores close to 1 within an interval between 0 and 1—are the ones occurring most times among the first positions in the lists. Though this formula has undergone changes through time, its mathematical foundations are available in the work of Strassburger-Frías and López Chávez (2000). (See also Strassburger-Frías and López-Chávez 1987; López Morales 1989).

Whenever a person is asked for writing a list of availability lexicon based on a particular stimulus, an individual cognitive task is triggered where several mechanisms regarding information processing are intertwined. An individual is asked to spend a pre-defined amount of time—commonly two minutes—writing down vocabulary they consider related to an initial cognitive stimulus, such as parts of the body, clothing, means of transport, etc. (López-Morales 1999).

The theoretical framework of prototype semantics seems an adequate way of representing how this cognitive process is carried out, as well as a fundamental key to building more solid foundations for lexical availability and, hence, for a better understanding of lexical breadth (Ávila-Muñoz and Sánchez-Sáez 2016). Combination of principles from prototype lexical-semantics and advances from Fuzzy set theory (Zadeh 1965; Zimmermann 2001) allow us to develop a satisfactory model for quantifying lexical richness (Ávila-Muñoz and Sánchez-Sáez 2014; Šifrar-Kalan 2017).

As developed below, when building on the concept of "availability" as it is used in our work, each stimulus is assumed to revolve around a prototype brought about by the concept the stimulus itself points to. Speakers gain access to their lexical network through the prototype that has been generated by a specific initial stimulus as entry gate. Hence, "accessibility" means the act of entering a particular lexical network nucleus as is represented by a stimulus. From this access gate on, each speaker will be able to follow his or her personal lexical network (Ávila-Muñoz and Sánchez-Sáez 2011). Once all the data provided by the informants during availability tests are gathered, quantification of accessibility determines each word's "centrality" for each stimulus. Ensuing association between words and their respective accessibility reveals the lexical structure of a given category. Within this structure, the closer are the words to the prototype's thematic nucleus, the higher their accessibility scores will be—in other words, the higher their Index of Centrality score (IC)—and, hence, they will form the common conceptual categorization.

Accordingly, different lines of conceptual categorization can be proposed. The first one would include the closest words to the prototypical nucleus—i.e., the most central ones to each studied category. As far as the IC decreases, lexical items start approaching what we can call "uncertainty" line, namely, a limit from which a given individual could doubt about the adequacy to the initial stimulus itself of the data he or she is providing during the availability test. To decide when a given word belongs to the prototype's core and, hence, is assimilable to the community prototype, we use the mathematical concept of "fuzzy set" (Zadeh 1965; Zimmermann 2001). It is a generalization of set theory where rather than on the elements' belonging to the group, we focus on their compatibility. Thus, different levels of compatibility can be established between a given set and their elements.

This notion of "compatibility" corresponds to our concept of "accessibility". In fact, determination of compatibility value or "Fuzzy Expected Value" FEV—or even its alternative

“Weighted Fuzzy Expected Value” WFEV—is one of the tools from the fuzzy set theory we have access to. By using these tools, we can set limits of characterization of belonging values, as well as parameters to identify “very characteristic” or “little characteristic” elements for a given sample. Thereby, the concept of centrality itself entails the foundation of a model of language representation where lexicon is a component that is built by the speakers themselves (Aitchinson 2003). Given that speakers belong to the same speech community, they share a great deal of lexicon from their language. Hence, the lexical core of any stimulus—i.e., a shared conceptual category—would be available for every speaker from the speech community. Moreover, this core could be estimated on the basis of a lexico-statistical analysis of a set of word lists provided by informants in the course of association test sessions like the ones developed within the frame of lexical availability research.

III

Given the foregoing, the objectives of our work are as follows:

1. The first step is to develop a theoretical framework capable of both underpinning the concept of “lexical breadth”—which is central to our work— and providing information that is easily interpreted.
2. The second stage aims at developing efficient mathematical parameters for generating measurements that will contribute to common determination of individual lexical breath and, then, lexical richness.
3. The third stage focuses on avoiding ad hoc restrictions observed in previous attempts to measure individual lexical availability—i.e. the data we use to determine each speaker's lexical breadth. Existing models undergo the disadvantages of the additivity principle of probability and, thus, their effects if uncontrolled. In our model, the additivity principle is *substituted* by the monotonicity principle through the incorporation of fuzzy

set theory, meaning that the higher the number of repetitions of a token within a lexical set—for example, the list of words produced by members of the same community—the higher the possibility that it is chosen.

4. The fourth and last step is to develop a flexible evaluation process that will allow for the continuous incorporation of fresh new data and let other researchers replicate the process with different data.

2. Fundamentals. From availability to accessibility: A necessary transformation

As shown so far, in our model of linguistic representation, lexicon is perceived as a quality component constructed by speakers themselves. We can therefore assume that all informants, as well as any other individual within the same speech community, will be capable of accessing the vocabulary belonging to a prototype's nucleus—in our case, the prototype is presented via the cognitive stimulus that triggers the production of a word list. It follows that the individual lexical breadth will primarily be determined by the speaker's most specific vocabulary that is not as available to the whole community as it is to him or herself. This approach allows us to consider the individual lists of available vocabulary as complementary to each other, resulting in new opportunities for quantitative analysis that have yet to be explored: the hypothesis of this study is that an individual's lexical breadth will be higher if they provide words with a lower availability index. This is logical, since such individuals should be able, by definition, to access the common vocabulary shared by the speech community more easily aside other less common words.

Our approach to lexical availability builds on the simple assumption that members of a speech community have specific vocabulary in common associated with certain frequently observed cognitive prototypes (Ávila-Muñoz and Villena-Ponsoda 2010). As commented on above, word association tests are used to access this available vocabulary. No specific

restrictions are given on the provided words' morphology, proximity of meaning to the CI's prototype core, further associations, etc. A sample of such word lists can be used to calculate an index that considers—as mentioned before—both the frequency of appearance of each element and its positions in the individual lists (Strassburger Frías and López-Chávez 2000).

The resulting lexical availability index (LAI) is useful for characterising the words themselves and building both overall and category-specific vocabularies. However, to be a part in our model, LAI should undergo a profound transformation, so that it may characterise not only the words themselves but also the speakers who underly the lists of available words. This transformation occurs via the following premises:

1. The stimulus that activates the association process is an access point to a network of lexical elements perceived by the subject as interrelated, as well as related to the concept referenced in the stimulus.
2. The structure of this network is subjective and inherent to the individual, meaning it cannot be strictly extrapolated to other subjects.
3. For various reasons, including both social and cultural factors, individuals within the same speech community develop similar lexical structures. We regard this shared sociocultural metastructure as being the vocabulary associated with a communal cognitive prototype; hence, the communal prototype does not exist in itself, but it is rather a creation of the members of a community. Most likely, a change in the group of speakers would therefore produce a change in the shared metastructure.
4. When an individual provides a list of available vocabulary, he or she first accesses lexical elements most closely related to the subject referred to in the stimulus. They then move through their lexical network, searching for elements less closely related to the original access point and giving way to a list of words in a decreasing degree of accessibility.

There is the possibility that, should an individual feel she or he has moved too far from the stimulus, they will re-enter their network via a new access point (Ávila-Muñoz and Sánchez-Sáez 2011). That is, they will restart the word-association process by returning to the subject proposed.

5. Both the amount of vocabulary provided and the speed at which it is accessed—i.e. the required speed to access all the words of a list—are relevant. It is self-evident why the number of words provided is not the sole determinant of lexical breadth. However, currently carried out tests just take account of words' rank order while time is not recorded.
6. The order of the terms on a single list is relevant to determine their accessibility, although the importance of order may be limited by the individual re-entering his or her network via a new access point.
7. The structure of a cognitive prototype is considered to consist of a nucleus and a periphery. The former is accessible to the entire community, whilst the latter is accessible to individuals depending on their degree of lexical breadth. The transition from nucleus to periphery is not abrupt but gradual.

The model underlying this study results from the last point. Due to the format and structure of the associative tests, however, this initial model does exhibit limitations. Although we assume that lexicon is structured as a multiconnected network, determining its layout would go beyond the scope of our work. Nevertheless, such network is inherent in every informant and changes over time. We therefore resort to a simplified version of the model, in which the structure of a communal prototype consists of a network of terms organised by degree of accessibility. In this way, the need for the provision of information about the nature of the

connections is removed and the focus of our study is guided towards determining the ease with which terms provided during a test can be accessed.

Based on the assumptions above, we have developed an innovative model that is capable of capturing reality and can be linked to a specific area of Mathematics due to the incorporation of fuzzy set theory (Zadeh 1965; Zimmermann 2001). As has already been mentioned, the latter is a generalisation of set theory in which the compatibility of elements with their set is considered, instead of their belonging. In this way, different levels of compatibility can be established between elements and their set.

3. A Proposal for the calculation of Individual Lexical Breadth

The measure of lexical breadth we propose accounts for both the quantity and specificity of the vocabulary provided by an individual and considers their respective degree of participation in the results of the studied sample. Our hypothesis is simple: The bigger the number of specific words provided by the speaker—i.e. words with lower LAI—the higher the speaker's lexical breadth. This is the result of considering that such individuals should be able to access the common vocabulary shared by the speech community easily in addition to other less general words. The precise order of the elements on a list should be inessential for the calculation of individual lexical breadth, due to the above-mentioned recurring possibility of re-entry.

Our measure establishes an index for each speaker. This index represents the individual speakers' lexical breadth. The notion of "lexical breadth" underlies the speakers' lexical capacity or competence—and, hence, their lexical richness—as based on the availability of each of the words provided during the test duration. For this purpose, we have employed approaches based on the probabilistic sum of values of not-availability from the vocabulary supplied by each speaker—i.e. a specific kind of operator which should not be confused with sum of

probabilities. It has been essential to develop an index capable of considering both the number of elements provided and their relevance to the concept—or cognitive category—they represent, since the overall vocabulary accessible to most of the community's members should not necessarily be given the same relevance as more specific vocabulary in fact would receive.

Elements belonging to the nucleus of a communal prototype's lexical spectrum should be of little relevance to the calculation of an individual's lexical breadth. However, terms that do not belong to the nucleus, also called non-prototypical vocabulary, should have greater weight in the calculation, the reason being that these terms indicate a higher degree of lexical breadth when provided by an individual, as they are not as widely used and are thus considered less accessible to the entire group of speakers. Consequently, we must establish a logical framework according to which non-prototypical vocabulary, i.e. vocabulary of restricted use is given greater importance than words more central to the lexical spectrum of a shared cognitive category, also called nuclear vocabulary.

Based on these logical assumptions, a proposal for measuring lexical capacity should embrace two different aspects in order to obtain a viable index of individual lexical breadth: (a) the total number of words provided on a list (ILA, Index of Lexical Amplitude), and (b) the specificity of the terms provided (ILD, Index of Lexical Decentralisation). These are not mutually exclusive, and individuals that provide more specific vocabulary are generally capable of accessing common vocabulary faster, corroborating the hypothesis that speakers who provide more specific vocabulary can also access the words belonging to the nucleus of the communal prototype.

3.1. Mathematical modelling of compatibility and decentralisation

The model we propose is capable of determining the spectrum of the lexicon associated with each stimulus for the population under study. As a result, we are in a position to establish

the degree of “compatibility”, between 0 and 1, of each term on a list with the concept it represents. When the “compatibility index” of each word in use by the sample of speakers is observed in relation to the lists of available vocabulary provided by an individual, we gain insight into the degree of “centralisation” or “decentralisation” of the provided lexical production, both to the lexical spectrum of each stimulus and to the entire lexical spectrum consisting of all the stimuli.

The decentralisation of a production with respect to the lexical spectrum of a communal prototype is the opposite of its compatibility with it. If a term is related to a prototype but shows a low degree of compatibility—that is, it is non-prototypical—it should be assumed it is not regularly used by the general population, meaning it has a high degree of decentralisation or specificity. In this manner, a term’s specificity is complementary to its compatibility, and a standardised operation within the theory of fuzzy sets can be employed to calculate it: the complement operator. In the fuzzy sets framework, the measure assigned to each element corresponds to the degree to which that element belongs to the set represented by the fuzzy set, and the complement represents the degree to which it does not.

So far, for the lexical availability index (LAI), the order of appearance of terms has been considered relevant. When subjects produce lists of available vocabulary, they presumably provide lexical elements more central and therefore more relevant to the stimulus prior to elements that are more distant from the prototype’s nucleus. Subjects may, however, re-enter their lexical network using a new access point, whenever they consider they have moved too far from the prototypical nucleus of the concept referenced in the stimulus. Nevertheless, when the data provided by all subjects from a particular sample are combined to construct a prototype’s lexical spectrum, this re-entry is blurred in the result. It is for this reason that we believe the order of elements should be disregarded when measuring individual lexical breadth,

with emphasis only on the vocabulary provided and the lexical spectrum of the communal prototype.

If a term's (x) compatibility with the shared lexical structure (c) is expressed by $m_c(x)$, as its degree of compatibility must be in the interval $[0,1]$, its decentrality could be determined as $1 - m_c(x)$. The result would express what the term is missing for it to be considered part of the nucleus of the shared structure. Irrespective of its position on a list, the degree of specificity is calculated for every term provided by a subject. We must therefore develop a single indicator that combines the information about each term's specificity. That is the way we used the aforementioned index of lexical decentralisation (ILD), as it is the expression of a term's distance to the nucleus of the communal prototype, accessible to all the subjects.

The concept underlying the model that we have developed to determine a speaker's degree of lexical decentralisation is the aggregating degree of decentralisation of the vocabulary they provide.

Two methods may be used to calculate this:

The first one involves obtaining the accumulated specificity by summing the not-availability—i.e. decentrality—of the individual elements. This approximation corresponds to the concept of cardinality as defined in the fuzzy set theory.

The second method involves establishing the compatibility of the speaker's realization with the concept of “wide lexical capacity”, by building a new fuzzy set over the set of speakers.

The first has a direct interpretation and requires, as pointed out above, only a simple calculation in the form of a sum. The second establishes an index with an upper and lower limit value, which can be interpreted in terms of fuzzy sets as the compatibility of the speaker realization with the notion of “wide lexical capacity”. Both approaches show advantages and drawbacks. In this work we opted for the second one in order to stay in the same representation model as fuzzy sets.

In this way, to calculate each subject's degree of lexical decentralisation, or their ability to access non-prototypical vocabulary, the index of lexical decentralisation (ILD) of each provided term is aggregated via the addition rule of probability:

$$a + b - a * b = 1 - (1 - b)(1 - a)$$

By generalising to the set all the words enunciated by the speaker in a particular cognitive category, it results in the following formula:

$$d = 1 - \prod_x (1 - (1 - m_c(x))) = 1 - \prod_x m_c(x)$$

where x is every term provided on the lists (c), and $m_c(x)$ is the degree of compatibility of a term with the shared lexical spectrum of each category.

However, this process is not trouble-free: given that the probability additive aggregation operator value shows a consistent tendency to increase, never to decrease, which means that should the degree of decentralisation of one single term be valued as 1, the entire calculation will adjust to suit. As this situation is uninformative, a control procedure is added by scaling the obtained decentralisation measures through a parameter k , which has to be determined in order to both control the above-mentioned increment and obtain an informative range. This way, the calculation of individual lexical decentralisation, the measure most closely related to the concept of lexical breadth, is given the following definite formula:

$$d = 1 - \prod_x (1 - k \cdot (1 - m_c(x)))$$

3.2. Validation of the Methodological Instrument

3.2.1. Participants, instruments and procedures

To demonstrate the validity of our proposal, we analysed lists of available vocabulary provided by a sample of individuals from the city of Malaga in southern Spain, pre-stratified by sex, education level and age ($N = 72$, Table 1) given that the goal of the investigation was

to build a dictionary of lexical availability for the whole city (Ávila-Muñoz and Villena-Ponsoda 2010).

Table 1. Distribution of the sample of speakers

Education	Male				Female			
	0	1	2	Total	0	1	2	Total
Age								
20-34	4	4	4	12	4	4	4	12
35-54	4	4	4	12	4	4	4	12
> 55	4	4	4	12	4	4	4	12
Total	12	12	12	36	12	12	12	36

Note. Education:

0 = no education or uncompleted primary education

1 = uncompleted secondary education

2 = completed secondary education

The 72 lists contain 24,047 words, or *tokens*, for a total amount of 5,860 types (distinct words). Table 2 breaks this information down into the cognitive stimuli used in the tests. Selection of stimuli was based on previous work on available lexicon in both Spain and several American Speaking countries (see López Morales 1989; Ávila and Villena 2010).

Table 2. Number of tokens and types by cognitive stimulus

Stimulus	Tokens	Types
01. The human body	1659	272
02. Clothing	1434	255
03. Parts of a house (not including furniture)	1217	128
04. Furniture in the home	1226	194
05. Food	1729	313
06. Objects laid on the table for a meal	1190	190
07. The kitchen and kitchenware	1297	281
08. The school: furniture and materials	1422	284
09. Heating, lighting, and means of ventilating a space	850	184
10. The city	1461	376
11. The countryside	1377	467
12. Means of transport	1278	177
13. Agricultural and garden tasks	919	254
14. Animals	1752	300
15. Games and entertainment	1306	350
16. Trades and professions	1519	354

17. Money and economy	1250	403
18. Internet and computers	1238	439
19. Colours	1430	187
20. The sea	1493	452
TOTAL	27047	5860

We created two separate measures, as mentioned above, to determine individual lexical breadth: (a) quantity (index of lexical amplitude, ILA), and (b) degree of decentralisation (index of lexical decentralisation, ILD). The first is a measure of the *total number* of words (*types*) provided by a subject throughout their test—i.e. based either on all the included CI (ILA) or separate for each stimulus (Stimulus Lexical Amplitude, SLA). The second measures the accumulated relative decentralisation of his or her vocabulary; that is, the number of words provided that are not common to the community, both throughout their test (ILD) and for each stimulus (Stimulus Index of Lexical Decentralisation, SILD).

The preliminary idea is that the more words a subject provides, the higher the probability that their lexical production will become decentralised, meaning they will provide less frequently used terms, that are more technical, specialised or cultivated (taken from literature, science, etc.). A strong and positive correlation is thus expected between lexical decentralisation (ILD) and lexical amplitude (LA).

To prove this correlation, we developed two lexical variables (ILA and ILD) corresponding to the two measures mentioned and used them to analyse the data collected throughout the investigation. For this study, we considered LA as the total number of words provided for the 20 stimuli. As for the ILD—i.e. the degree of decentralisation of a production—we considered this a quality of the vocabulary provided by a subject, describing its degree of compatibility by means of the one proposed by other subjects for the same stimulus—in fact, the SILD, as explained above. The index for the latter, taking values in the interval [0.1], allows each individual to be described according to the synthesis of the degrees of non-prototypicality of all the terms he or she provided during their test. This index can also be calculated in an easy

way using a computer tool that we have created *ad hoc* (*DispoCen*, vid. Ávila-Muñoz, Sánchez-Sáez and Odishelidze 2021). This tool allows to obtain the ILD of a given sample. The SILD moves closer to 1 if a speaker mentions a high number of words for a stimulus that do not coincide with those provided by the rest of the group. Should the opposite occur, meaning a certain individual delivers less new vocabulary and more terms that coincide with those of the group, the SILD will move closer to 0.

The ILA and the ILD are related and, as expected, the more words an individual provides, the higher the probability that a subset of these will not coincide with those produced by the other subjects. It is logical that this has to do also with the availability of each term to the individual, as more accessible terms will be provided first, leading to these terms to occur most frequently throughout the sample. (Notice that the ILD reflects decentrality of the speaker's vocabulary, independently of its size, whilst LA is strongly size-dependent and it only shows the number of words each speaker mentions for each conceptual category).

As shown in Figure 1, a positive, strong, and significant correlation was found between the two variables ($Rho = 0.951$, $R^2 = 0.904$, sig. = 0.000, level of sig. = 0.01, $N = 72$). In fact, this figure is a statistical reflection of a well-known qualitative constant in lexicology: cognitive structures consist of a central nucleus of lexical elements that share many attributes with the prototype of the category in question (such as *arm* and *mouth* in connection with the human body, or *eagle* and *sparrow* in connection with birds). As we distance ourselves from the centre of a prototype's lexical spectrum and move away from frequently used, very characteristic elements that have much in common, the words become less frequent in use and therefore less prototypical, whilst becoming increasingly decentralised (such as *steatopygia* and *lipodystrophy* in connection with the human body, or *ostrich* and *gannet* in connection with birds).

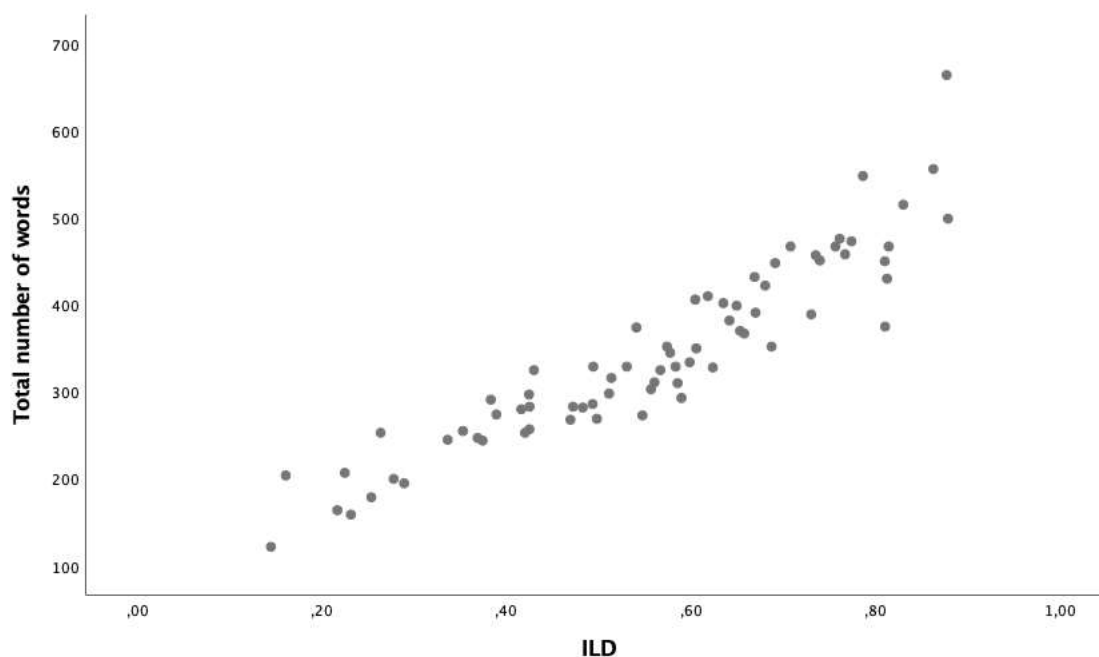


Figure 1. Correlation between the number of words provided and their decentralisation ($R^2 = 0.904$)

3.3. Corrected Lexical Values: The Gradation of Nuclearity

The calculation of the two lexical variables considered thus far is adequate. It does not, however, account for the internal structure of the vocabulary provided by a speaker for each stimulus but just a quantitative measure of his or her lexical capacity. Considering each individual list of words, it is quite reasonable to suppose that, when a particular prototype is proposed to a certain informant during the test, she or he will go through their lexical network, picking up first nuclear words and progressively getting away from the prototype centre through paths produced by diverse associations. This would make the speaker move away from the entry point explicitly proposed by the prototype label.

An underlying problem emerges then regarding distance to the prototype. On the one hand, it is conceivable that a particular word could be deeply related to the prototype but not generally accessible—maybe due to its technical specificity—and it could indicate, nevertheless, a high lexical capacity. On the other hand, it is very likely that associations would leave out of the lists words directly connected to what we consider the prototype centre,

providing then lexical entries only indirectly related to this prototype nucleus and hence, irrelevant for assessing the informant's lexical capacity related with that prototype centre.

Though it is true that this problem is irrelevant when quantifying the lexical availability of a particular word, it is vital to clearly identify and distinguish both situations to be able to determine the speakers' lexical capacity through their degree of decentralisation (ILD). So far, we have not found a way to address this problem without adding more information to the system, i.e. by considering the researcher's assessment on the compatibility of each word with the proposed prototype.

Beyond the constants deduced from the individual's cognitive structures and the above-mentioned fact that the index of lexical decentralisation (ILD) builds on the concept of gradation of compatibility, the vocabulary given by each speaker is quite variable, consisting of nuclear elements (N-1) as well as the following three groups, each distanced progressively further from the nucleus: denotational associations (N-5), derivatives, simply representing repetitions of already mentioned elements (N-25), and individual associations (N-0).

The nucleus of elements most available to the speakers in the sample—that is, the terms appearing most frequently throughout and most often in the first positions of the lists—consists of elements that are highly likely to belong to the communal prototype from a semantic point of view, which means that it is highly probable they will be mentioned by the subjects (parts of the body: *head, leg, hand*). In this sense, they may be defined as nuclear elements (N-1), independent of their individual degree of centrality to the nucleus.

Although associations (N-5) do form part of the cognitive prototype, they are not considered nuclear and stem rather from connotations (evocations, associations, etc.) of the denotations of nuclear elements. As connotations of denotations, these are communal and not individual (parts of the body: *taste, touch, amylase, lipase*).

Derivatives (N-25) imply a quantitative or qualitative increase in already mentioned vocabulary (*spoon, wooden spoon, dessert spoon, baby spoon, etc.*). Whilst these do belong to the prototype, they are an effect, or rather an artefact of the experiment and contribute little in terms of lexical breadth.

Finally, individual associations (N-0) are elements that have neither a direct nor communal link to the prototype. They are the product of evocations or associations stemming from references of nuclear elements and not their denotations (connotations of references; see Cruse 1986). The relevance of these terms in the study of a single conceptual category's available lexicon and lexical breadth is highly debatable, though they may well be of great interest in the analysis of psycholinguistic processes.

In this manner, the vocabulary provided by each speaker can be divided into four groups corresponding to those above: nuclear elements (N-1), communal associations (N-5), derivatives (N-25) and individual associations (N-0). Each denomination corresponds to its importance in the constitution of the individual lists and thus to the value that should be given to the lexical elements in each group; 1, 0.5, 0.25 and 0.0, respectively.

The calculation of the indices described in the previous section should therefore be modified—based on the assessment of each word by the researchers themselves on the four scores classification: 0.0-0.25-0.5-1.0—to account for the effects of these differences both on the quantity (ILA) and the decentralisation (ILD) of the provided vocabulary. The measure of the number of words on each list will now consider the lexical elements as classified into the four groups (corrected lexical amplitude, COILA) and the same system of classification will be used to produce the corrected index of lexical decentralisation (CoILD). The COILA can be calculated via the following addition:

$$\text{COILA} = \text{N1} + (\text{N5} * 0.5) + (\text{N25} * 0.25) + (\text{N0} * 0)$$

More exactly, given that $n(x)$ is the value of compatibility of a certain word x with the proposed prototype (1, 0.5, 0.25, 0), as assigned by the researcher, the corrected specificity of this word can be calculated as the intersection of the word's decentralisation and the assigned relevancy. In fuzzy sets theory this intersection can be computed as a compatibility value of

$$(1 - m_c(x)) \cdot n(x)$$

Therefore, a given word would suggest high specificity when showing non nuclearity (i.e., whether $m_c(x)$ is near 0) but being closely related to the prototype (i.e., $n(x)$ is near 1). Through a process alike the one described above, specificity can easily be integrated for each speaker

$$CoILD = 1 - \prod (1 - k \cdot n(x) \cdot (1 - m_c(x)))$$

and then normalised with a suitable value of k .

Table 3 compares these variables in the studied speakers. Those whose $ILD > CoILD$ are highlighted in grey. It is to be recalled, firstly, that the indices of decentralisation (ILD , $CoILD$) are calculated on the basis of the aggregation—for each informant—of each word's values of and, secondly, that communal associations (N-5), derivatives (N-25) and individual associations (N-0) are simply repetitions or variants of elements already included in the group of nuclear vocabulary (N-1). After the normalization process, the ILD should therefore generally be lower than the $CoILD$; the former being based on all the terms mentioned by a speaker and assigning each the same value (that is, $N = 1$), whilst the latter considers only the nuclear elements in their entirety (N-1), and assigns their respective lower values (N5 to N0) to the other elements. In other words, on the one hand, the ILD includes the vocabulary from the groups N 5, N 25, and N 0 assuming that these terms are singular elements with the same weight and relevance as the nuclear words (N 1). The $CoILD$, on the other hand, does not generally consider these non-nuclear elements (N 0, N 25, N 5) as singular units equivalent in weight and relevance to the nuclear ones (N 1), since they are often simply variants of previously mentioned terms.

As a consequence, ILD is likely to oversize the speaker's lexical capacity due to the impossibility of ascertaining the relative centralisation of the vocabulary provided—i.e. its relative closeness to the experience field's centre since, in fact, speakers who do mention words marked as N0 would obtain wrong high scores. Conversely, COILD is able to deal with this issue through an adjustment process that underestimates evaluation of these sort of speakers' realisation and, as a side effect, the respective evaluation of the rest of the speakers increases in the normalisation process.

As explained, the information derived from the N1, N5, N25 and N0 sets results in the corrected index of lexical decentralisation (COILD) by modifying the decentralisation measures, scaling them by factors 1, 0.5, 0.25 and 0, as a previous step to the aggregation. In this way, researchers are able to discern between, on the one hand, terms closely related to the proposed prototype and terms enunciated randomly, on the other. Thanks to this simple procedure, the lexical spectrum is eventually adjusted.

For individuals whose $ILD > COILD$, it is assumed that they will have provided truly distinct elements in the non-nuclear N-5, N-25 and N-0 groups from those included in the nuclear group (N-1). This may serve as an indication of a special ability in the individual to provide original lexical associations for the proposed stimuli, although it should be noted here that—as explained above—the index itself (CoILD) also assigns elements different values depending on whether they did well in a set.

Table 3 classifies speakers according to their scores in the lexical measures developed so far. As shown, two groups are separated based on the speakers' values in decentralisation—group 0 ($ILD > COILD$) and group 1 ($ILD < COILD$). Inversely, as expected, ILA is always higher than COILA.

Table 3. Initial and corrected lexical variables organised in increasing order of COILD

Speaker	N-0	N-25	N-5	N-1	ILA	COILA	ILD	COILD
64	5.00	5.00	1.00	111.00	122	112.750	.14	.21
72	0.00	6.00	12.00	149.00	164	156.500	.22	.26
66	12.00	4.00	16.00	127.00	159	136.000	.23	.28
56	7.00	2.00	2.00	193.00	204	194.500	.16	.28
61	20.00	11.00	6.00	163.00	200	168.750	.28	.31
58	5.00	5.00	12.00	185.00	207	192.250	.22	.32
48	14.00	1.00	15.00	149.00	179	156.750	.25	.32
98	8.00	12.00	17.00	158.00	195	169.500	.29	.33
35	3.00	3.00	5.00	242.00	253	245.250	.26	.40
50	20.00	7.00	21.00	197.00	245	209.250	.34	.41
69	15.00	1.00	6.00	269.00	291	272.250	.38	.43
60	117.00	.00	40.00	172.00	329	192.000	.49	.44
2	6.00	16.00	16.00	209.00	247	221.000	.37	.45
65	3.00	3.00	21.00	217.00	244	228.250	.37	.46
97	17.00	11.00	21.00	234.00	283	247.250	.43	.48
51	11.00	.00	20.00	224.00	255	234.000	.35	.49
1	25.00	3.00	20.00	232.00	280	242.750	.42	.49
39	6.00	11.00	7.00	229.00	253	230.250	.42	.50
26	8.00	5.00	7.00	254.00	274	258.750	.39	.51
3	12.00	.00	18.00	227.00	257	236.000	.42	.51
59	13.00	23.00	13.00	219.00	268	231.250	.47	.51
30	58.00	10.00	30.00	205.00	303	222.500	.56	.53
29	4.00	8.00	17.00	268.00	297	278.500	.42	.53
4	.00	7.00	15.00	303.00	325	312.250	.43	.55
36	14.00	5.00	16.00	247.00	282	256.250	.48	.56
37	15.00	3.00	29.00	222.00	269	237.250	.50	.56
75	14.00	11.00	18.00	273.00	316	284.750	.51	.57
40	8.00	2.00	15.00	258.00	283	266.000	.47	.57
9	45.00	.00	22.00	206.00	273	217.000	.55	.57
42	7.00	8.00	9.00	262.00	286	268.500	.49	.58
53	13.00	5.00	37.00	243.00	298	262.750	.51	.59
47	10.00	7.00	31.00	263.00	311	280.250	.56	.60
63	51.00	8.00	28.00	263.00	350	279.000	.61	.60
8	20.00	23.00	20.00	266.00	329	281.750	.53	.61
32	37.00	12.00	19.00	257.00	325	288.500	.57	.61
46	14.00	31.00	20.00	245.00	310	262.750	.59	.61
45	19.00	11.00	31.00	267.00	328	285.250	.62	.63
67	11.00	2.00	40.00	321.00	374	341.500	.54	.64
71	14.00	2.00	32.00	245.00	293	261.500	.59	.64
74	3.00	35.00	12.00	279.00	329	293.750	.58	.64
41	18.00	17.00	19.00	280.00	334	293.750	.60	.65
28	12.00	13.00	22.00	298.00	345	312.250	.58	.66
55	12.00	4.00	26.00	310.00	352	324.000	.57	.67
70	32.00	13.00	14.00	351.00	410	361.250	.62	.67
52	26.00	15.00	38.00	291.00	370	313.750	.65	.68
24	10.00	22.00	28.00	307.00	367	326.500	.66	.69
77	26.00	6.00	21.00	349.00	402	361.000	.63	.69
33	38.00	26.00	29.00	296.00	389	317.000	.73	.70
76	20.00	109.00	27.00	219.00	375	259.750	.81	.70
49	20.00	4.00	52.00	306.00	382	333.000	.64	.70
22	16.00	19.00	17.00	347.00	399	360.250	.65	.70
34	9.00	1.00	19.00	377.00	406	386.750	.60	.72
99	21.00	23.00	20.00	288.00	352	303.750	.69	.72

6	20.00	6.00	18.00	388.00	432	398.500	.67	.74
27	10.00	2.00	17.00	362.00	391	371.000	.67	.75
31	13.00	8.00	28.00	399.00	448	415.000	.69	.75
96	60.00	7.00	60.00	349.00	476	380.750	.76	.76
38	10.00	4.00	19.00	389.00	422	399.500	.68	.76
5	18.00	18.00	31.00	400.00	467	420.000	.71	.77
68	38.00	49.00	50.00	313.00	450	350.250	.81	.77
73	56.00	41.00	69.00	301.00	467	345.750	.81	.79
25	29.00	27.00	41.00	333.00	430	360.250	.81	.79
21	11.00	17.00	25.00	398.00	451	414.750	.74	.79
23	12.00	21.00	31.00	394.00	458	414.750	.77	.80
13	24.00	34.00	36.00	379.00	473	405.500	.77	.81
43	13.00	2.00	20.00	422.00	457	432.500	.74	.82
62	6.00	12.00	27.00	422.00	467	438.500	.76	.83
15	20.00	20.00	28.00	480.00	548	499.000	.79	.84
14	57.00	20.00	52.00	427.00	556	458.000	.86	.85
10	70.00	36.00	49.00	344.00	499	377.500	.88	.85
19	27.00	5.00	36.00	447.00	515	466.250	.83	.87
12	44.00	5.00	49.00	566.00	664	591.750	.88	.91
Average	20.6	12.8	24.38	285.9	343.7	301.5	0.56	0.61
Group 0	54.0	30.3	43.2	292.2	420.4	322.1	0.74	0.71
Group 1	14.6	9.7	21.0	284.6	330.0	298.0	0.52	0.59
Sig.	0.000	0.000	0.000	No sig.	0.007	No sig.	0.000	0.032

Note. Mean differences between both groups for each variable are shown. Speakers whose ILD > COILD are highlighted in grey

3.4. Nuclearity and displacement

The lexical variables discussed so far show a positive correlation (Table 4).

Table 4. Spearman correlation between the initial and corrected lexical variables (N = 72)

	ILA	N-1	N-5	N-25	N-0	ILD	COILD
N-1	.926						
N-5	.655	.457					
N-25	.404	.278	.242				
N-0	.484	.240	.619	.266			
ILD	.951	.831	.710	.494	.544		
COILD	.965	.910	.640	.424	.419	.973	
COILA	.952	.993	.511	.324	.289	.874	.940

Note. All correlations are significant. In bold: $p = 0.01$, rest: $p = 0.05$.

Table 5 shows Euclidean similarity and distance, giving an idea of the extent of the distance between the variables: the higher the bilateral coefficient is in the cells, the less distance there is between them and the stronger is their relation and viceversa. The structure of

the variables examined explains the connection between the COILA and N-1 as well as the advantages of using the corrected indices.

Table 5. Euclidean similarity between the lexical variables

	ILD	COILD	COILA	ILA	N-1	N-5	N-25
ILD							
COILD	.974						
COILA	.844	.919					
ILA	.926	.942	.966				
N-1	.801	.891	1.000	.940			
N-5	.684	.599	.473	.639	.399		
N-25	.407	.272	.065	.220	.000	.207	
N-0	.363	.204	.094	.335	.039	.584	.061

In line with expectations, the original (ILD) and corrected (COILD) indices of decentralisation are strongly related (0.974), yet their relationship varies with those measures describing the size of the provided lists (ILA, COILA). As stated above, it would be logical for the COILD to be in general higher than the ILD, due to correction of the centrality of each word as specificity and subsequent new normalisation. As shown in Figure 2, however, some speakers record higher values in the original index (ILD) than in its corrected version (COILD). This means that these speakers ($n = 11$; highlighted in grey in Table 3; also refer to Table 6) constitute the most productive group of individuals, since they also do the least to limit their capacities of lexical association, derivation and variation.

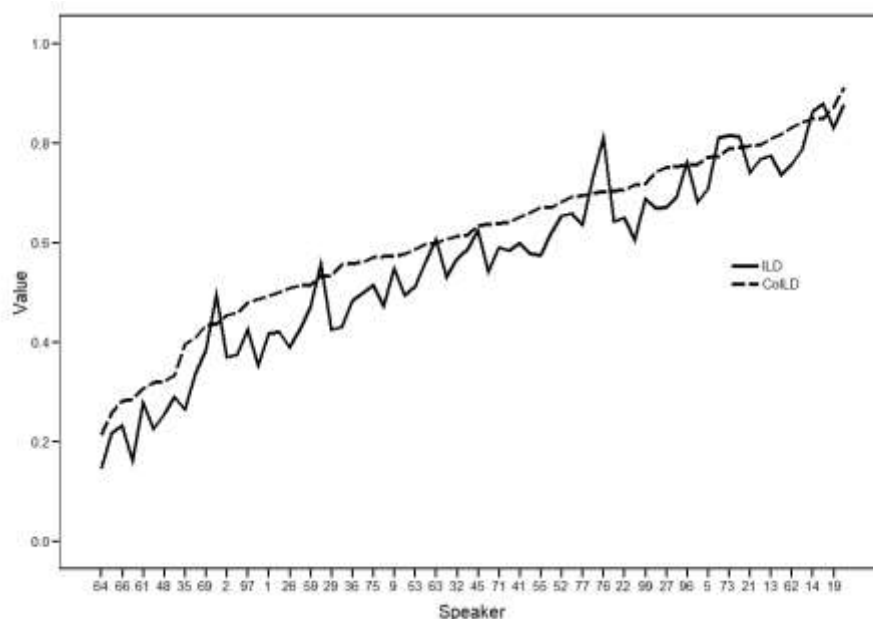


Figure 2. A comparison of the values of decentralisation recorded for each speaker, showing both the corrected (COILD) and original (ILD) indices

The speakers who have proved most productive have been isolated from the rest and are shown in Table 6. Though they mentioned a high quantity of terms per stimulus, many of these were associations or derivatives, leading to a lower COILD than expected. Interesting is, however, that these speakers do not only expand their lists by providing associations or derivatives, as they also belong to the group of speakers mentioning the most nuclear elements; only four individuals recorded values below the average (highlighted in grey in Table 6). These mention a relatively low number of nuclear elements and compensate by providing a higher number of non-nuclear elements (see Figures 3 and 4).

Table 6. Speakers whose $ILD > COILD$

Speaker	N-0	N-25	N-5	NON-NUCLEAR	N-1	ILA	COILA	ILD	COILD
60	117.00	0.00	40.00	157	172.00	329	192.000	.49	.44
30	58.00	10.00	30.00	98	205.00	303	222.500	.56	.53
63	51.00	8.00	28.00	87	263.00	350	279.000	.61	.60
76	20.00	109.00	27.00	156	219.00	375	259.750	.81	.70
33	38.00	26.00	29.00	93	296.00	389	317.000	.73	.70
96	60.00	7.00	60.00	127	349.00	476	380.750	.76	.76
68	38.00	49.00	50.00	137	313.00	450	350.250	.81	.77
25	29.00	27.00	41.00	97	333.00	430	360.250	.81	.79
73	56.00	41.00	69.00	166	301.00	467	345.750	.81	.79
10	70.00	36.00	49.00	155	344.00	499	377.500	.88	.85
14	57.00	20.00	52.00	129	427.00	556	458.000	.86	.85

Note. *Non-Nuclearity* is the sum of all the non-nuclear word sets (N-5, N-25, N-0) mentioned by an informant

In this way, the COILD discerns non-nuclear speakers, who produce a high number of non-nuclear elements N5, N25, N0 (Figure 3, *x* axis), independent of their ILD (Figure 3, *y* axis), although their lists do indeed become more repetitive in line with the decrease in their COILD.

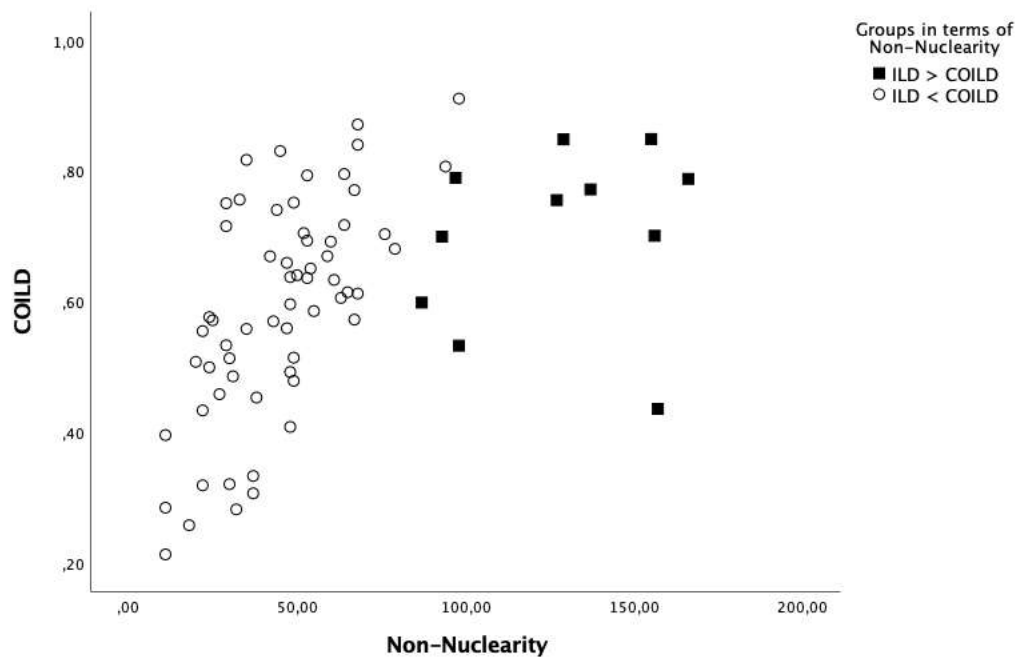


Figure 3. Dispersion of the subjects in terms of non-nuclearity (N-0 + N-25 + N-5)

This corroborates the hypothesis about the existence of non-nuclear speakers (group 0, marked with black squares in Figure 3). They are characterised by the provision of derivatives and associations (N-5, N-25, N-0) but differ in terms of the total number of provided words. Nuclear speakers, on the other hand, (group 1, marked with white dots in Figure 3) limit their associative and/or derivative capacities and focus on providing nuclear elements (N-1).

As a matter of fact, the relationship between the COILD and non-nuclearity is not linear but curved. At first it is positive; the COILD rises in line with the provision of non-nuclear elements up to a score of around 0.75. However, the increase stabilises at this point, remaining

constant with a tendency to fall (S-curve). Table 7 is a comparison of the linear and the alternative curved models.

Table 7. A curve fit estimate of the correlation between the COILD and Non-Nuclearity. Summary of the model and estimations of the parameters

Equation	Summary of the model					Estimations of the parameters			
	R^2	F	$df1$	$df2$	Sig.	Constant	b1	b2	b3
Linear	.27 2	26.175	1	70	.000	.470	.002		
Cubic	.43 6	17.552	3	68	.000	.209	.013	-9.63E- 005	2.28E- 007
S	.45 6	58.792	1	70	.000	-.237	-12.158		

The S-curve model accounts for 45.6% of the relationship between both variables, as opposed to the linear model that only explains 27.2%. Figure 5 compares the fit of the three curves.

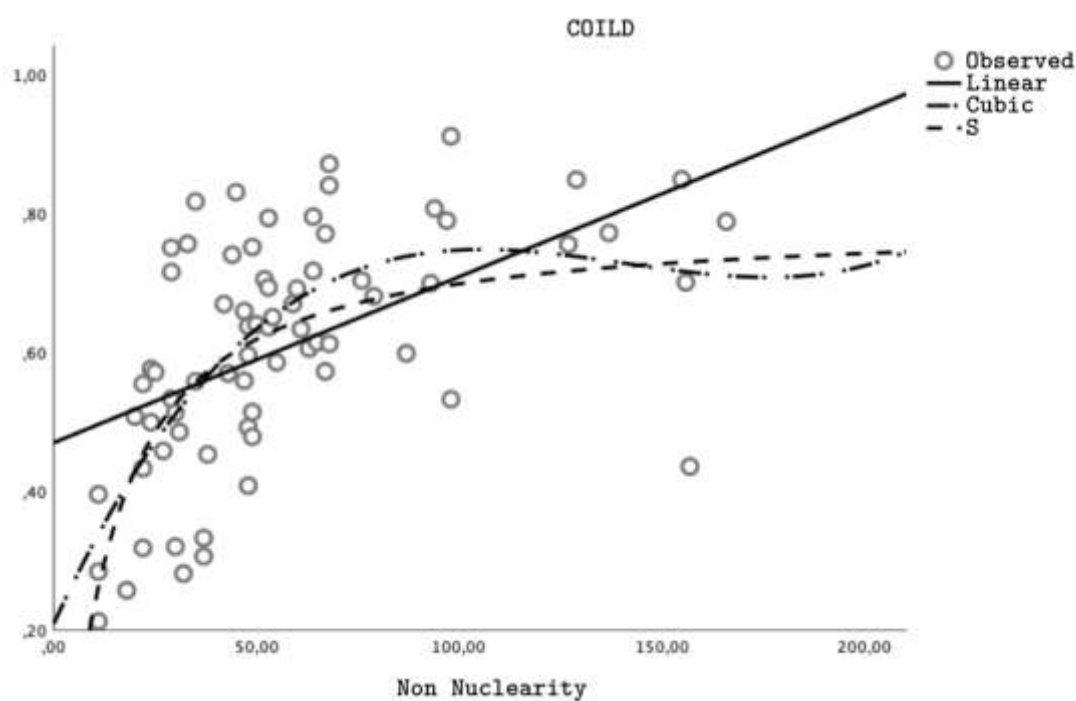


Figure 5. A comparison of the linear, cubic, and S-shaped fits between the COILD and Non-Nuclearity

The indices used in this study to measure speakers' lexical breadth provide an easily understandable picture that is interpreted in two complementary ways: both as a methodological model—i.e. a model of the test—and as a real model—i.e. a model of the speech community.

The totality of lexical elements provided by a speaker for a given stimulus during an availability test increases exponentially over the time spent on that stimulus. The speaker initially provides only nuclear elements for a stimulus. However, at a certain point, this process is discontinued and displacement ($N-5 + N-25 + N-0$) or deviation ($N-25 + N-0$) can be observed, as depicted in Figure 6.

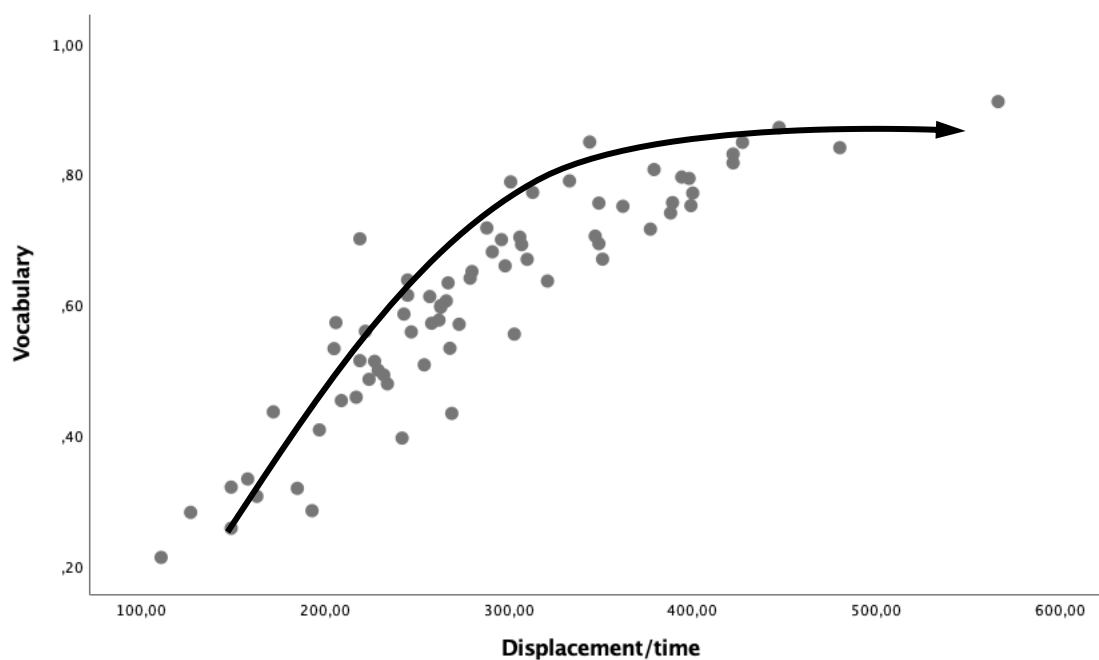


Figure 6. A model of the displacement of nuclear vocabulary during a test

This model produces results that were considered from different perspectives. By way of example, the relationship between the COILD—the most representative variable among those analysed—and the degree of non-nuclearity of the lexical productions exhibits a curvilinear relationship, the reasons for which undoubtedly lie in this model of displacement.

The notion is that certain speakers recording a notable degree of lexical breadth (in both general and specialised terms) develop the parallel ability to associate elements to the nuclear vocabulary of each category. The extent of the relationship of these elements with the nucleus varies, however (see Figure 6, above); some do belong and are linked through connotations or associations of the denotations that are prototypical of the category ($N-5$), whilst others are

mere digressions or derivatives (N-25) stemming from nuclear elements, or are simply individual evocations (N-0) that have no denotational relation to the stimulus. During their test, some speakers do develop the above-mentioned tendency towards providing non-nuclear elements (displacement). A number of these speakers persist in this way and record high values on the y axis as a result (displacement through deviation; see Figure 6). Others correct this displacement at some point and return to providing nuclear elements (looped displacement). The hypothesis is that, during the displacement, (a) the first group of speakers provides derivatives or individual associations (N-25, N-0) whilst the second provides communal associations (N-5); and (b) the second group puts a stop to the displacement, independent of its content, and returns to providing nuclear elements.

Figure 7 considers this difference. Deviation (F2) measures only derivatives and individual associations (N-25, N-0), whilst nuclearity (F1) is amplified to include denotational associations (N-5).

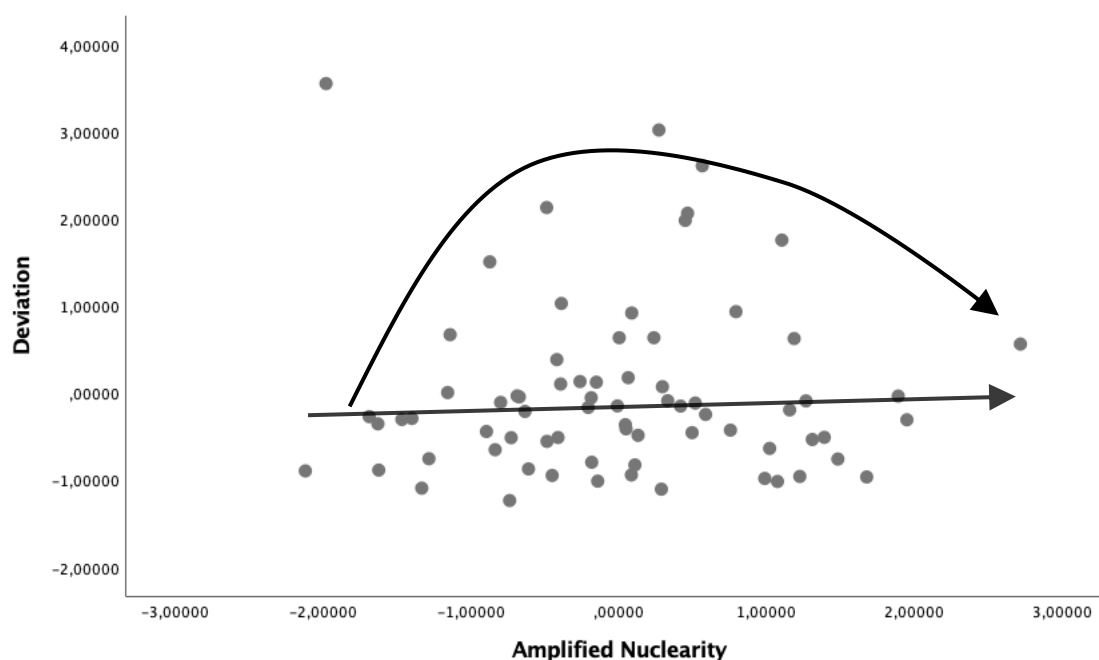


Figure 7. A model of deviation and amplified nuclearity of the variation in the speech community's available lexicon

The hypothesis is corroborated (Table 8). Speakers with the highest deviation (F2) provided the most deviating elements (N-25, N-0) relative to the quantity of nuclear vocabulary (N-1) and the total number of words mentioned (ILA).

Table 8. A comparison of speakers showing displacement and deviation

Speaker	ILA	N-1	N-5	N-25	N-0	Nucl.	Displ.	Dev.	F2
25	430	333	41	27	29	374	97	56	.97076
63	350	263	28	8	51	291	87	59	.75594
33	389	296	29	26	38	325	93	64	1.19530
96	476	349	60	7	60	409	127	67	.70239
30	303	205	30	10	58	235	98	68	1.23573
14	556	427	52	20	57	479	129	77	.98101
68	450	313	50	49	38	363	137	87	2.20044
73	467	301	69	41	56	370	166	97	2.34446
10	499	344	49	36	70	393	155	106	2.46100
60	329	172	40	0	117	212	157	117	2.34428
76	375	219	27	109	20	246	156	129	4.71905

From now on, we have available a new model capable of generating mathematical indexes that could be used—jointly with those existing already—to determine individual lexical richness in a more fitting and precise way.

4. Discussion and conclusions

In view of the obtained results, we can conclude that availability tests are an efficient method of accessing individual lexical breadth and although limitations were identified, simple solutions do exist. The instructions given at the beginning of an availability test should be altered in future, so that informants provide fewer individual associations (N-0), as well as fewer derivatives that simply constitute repetitions of already mentioned elements (N-25). However, specific restriction of the informants' freedom of writing anything it occurs to them will alter and, even, make impossible comparison with results from previous work on lexical availability. In any case, methodologically limiting the provision of nuclear vocabulary and denotational associations seems desirable if the aim of a study is to measure a speech

community's lexical availability and to analyse the differences between various groups of speakers—by determining the fields of experience observed most frequently in their daily lives, rather than representing conceptual networks through lexical connections. By avoiding such deviation, an associative test would result in a more precise measurement of the concept we call decentralisation, meaning the models developed for this study would be more likely to provide accurate predictions of individual lexical breadth.

The notion of "lexical richness" seems now to be something more than a common-sense idea about how many words a certain speaker passively knows or is actively capable of using in a specific context while he or she talks on a particular theme. The concept of "lexical breadth" is meant to capture the idea of a vocabulary with weighted size or extension. To such vocabularies the relevance of each word does not depend strictly either on its frequency or on its position in a rank list but it would be determined through its relative distance to the prototype of the lexical centre of interest (CI) this particular word belongs to.

To build this concept of "lexical breadth", lexical availability measures (ILA, COILA) are used. These measures are based on the individual average of words used during test and allow us to calculate the degree of coincidence between the words provided by an informant and by the rest of the subjects of a sample. This way, indices of decentralisation are available (ILD, COILD) thanks to concepts and tools from Fuzzy set theory.

A model computing the individual speakers' lexical decentralisation is able to show the amplitude of their available vocabulary—i.e. the one ready to be used when a particular theme turns up during conversation—since, as proved in this paper, the concepts of decentralisation, specificity and breadth are deeply related.

With reference to the fuzzy operators proposed in this paper (complement to one, probabilistic sum and product) leading to model complement, union and intersection operations and enable further calculation of both decentrality and specificity, we have chosen the simplest

and most usual ones. To definitely establish which are the most adequate operators, intensive research on the models underlying these experiments will have to be carried out in the future.

REFERENCES

- Aitchinson, Jean. 2012. *Words in the mind. An introduction to the mental lexicon*. Oxford: Wiley and Blackwell, 4th edition.
- Ávila, Raúl. 1988. Lengua hablada y estrato social: un acercamiento lexicoestadístico. *Nueva Revista de Filología Hispánica* 36 (1). 131-148. doi:10.24201/nrfh.v36i1.668.
- Ávila-Muñoz, Antonio-Manuel. 2016. Can speakers' virtual lexical richness be calculated? Individual and social determining factors. *Spanish in Context* 13 (2). 285-307. doi:10.1075/sic.13.2.06avi.
- Ávila-Muñoz, Antonio-Manuel & Juan-Andrés Villena-Ponsoda (eds.) 2010. *Variación social del léxico disponible en la ciudad de Málaga*. Málaga: Sarriá.
- Ávila-Muñoz, Antonio-Manuel & José-María Sánchez-Sáez. 2011. La posición de los vocablos en el cálculo del índice de disponibilidad léxica: procesos de reentrada en las listas del léxico disponible de la ciudad de Málaga. *Estudios de Lingüística. Universidad de Alicante (ELUA)* 25. 45-74. doi:10.14198/ELUA2011.25.02.
- Ávila-Muñoz, Antonio-Manuel & José-María Sánchez-Sáez. 2014. Fuzzy sets and prototype theory. Representational of cognitive community structures based on lexical availability tests. *Review of Cognitive Linguistics* 12/1. 133-159. doi: <https://doi.org/10.1075/rcl.12.1.05avi>
- Ávila-Muñoz, Antonio Manuel, Inmaculada Clotilde Santos-Díaz & Ester Trigo-Ibáñez, Ester. 2020. Análisis léxico-cognitivo de la influencia de los medios de comunicación en las percepciones de universitarios españoles ante la COVID-19. *Círculo de Lingüística Aplicada a la Comunicación* 84. 85-95. <http://dx.doi.org/10.5209/clac.70701>.
- Ávila-Muñoz, Antonio-Manuel, José-María Sánchez-Sáez & Nina Odishelidze. 2021. *DispoCen*. Mucho más que un programa para el cálculo de la disponibilidad léxica. *Estudios de Lingüística de la Universidad de Alicante (ELUA)* 35. 9-36. doi: 10.14198/ELUA2021.35.1.
- Ávila-Muñoz, Antonio-Manuel, José-María Sánchez-Sáez & Nina Odishelidze. 2021. *DispoCen*. <https://m.youtube.com/watch?v=IU5VfUvG4Ag>.
- Castañeda-Jiménez, Gabriela & Scott Jarvis. 2013. Exploring lexical diversity in second language Spanish. In Kimberly L. Geeslin (ed.) *The handbook of Spanish second language acquisition*, 498-513. Hoboken, NJ: Wiley-Blackwell.
- Cruse, D. A. 1986. *Lexical semantics*. Cambridge: Cambridge University Press.
- Davies, Mark. 2005. *A Frequency Dictionary of Spanish: Core Vocabulary for Learners*. London: Routledge.
- Gougenheim, Georges, René Michéa, Pierre Rivenc & Aurélien Sauvageot. 1954. *L'élaboration du français élémentaire. Étude sur l'établissement d'un vocabulaire et d'une grammaire de base*. Paris: Didier.
- Hernández-Muñoz, Natividad, Cristina Izura & Andrew W. Ellis. 2006. Cognitive aspects of lexical availability. *European Journal of Cognitive Psychology* 18 (5). 730-755, DOI: [10.1080/09541440500339119](https://doi.org/10.1080/09541440500339119).
- Jarvis, Scott. 2002. Short texts, -fitting curves and new measures of lexical diversity. *Language Testing* 19. 1-15. doi:10.1191/0265532202lt220oa.
- Jarvis, Scott. 2013. Defining and measuring lexical diversity. *Vocabulary Knowledge. Human ratings and automated measures*. In Scott Jarvis & Michael Daller (eds.), 13-44. Amsterdam: John Benjamins.
- de Kock, Josse. 1983. *Elementos para una estilística computacional*. Madrid: Coloquio.
- Lonsdale, Deryle & Yvon Le Bras. 2009. *A Frequency Dictionary of French: Core Vocabulary for Learners*. London: Routledge.
- López-Chávez & Strassburger-Frías. 1987
- López-Morales, Humberto. 1999. *Léxico disponible de Puerto Rico*. Madrid: Arco Libros.
- Malvern, David & Bryan James Richards. 2000. Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing* 19. 85-104. doi:10.1191/0265532202lt221oa.
- Malvern, David, Bryan James Richards, Ngoni Chipere & Pilar Durán. 2004. *Lexical Diversity and Language Development: Quantification and Assessment*. Hampshire, United Kingdom: Palgrave Macmillan.
- McCarthy, Philip & Scott Jarvis. 2010. MTL, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior research methods* 42 (2). 381-392.
- Samper Padilla, Juan Antonio. 1998. Criterios de edición del léxico disponible. *Lingüística* 10. 311-333.
- Šifrar-Kalan, Marjana. 2017. La universalidad de los prototipos semánticos en el léxico disponible de español. *Verba Hispanica* 24. 147-165. DOI: 10.4312/vh.24.1.147-165.

- Strassburger Frías, Carlos & Juan López-Chávez. 2000. El diseño de una fórmula matemática para obtener un índice de disponibilidad léxica confiable. *Anuario de Letras: Lingüística y filología* 38. 227-251.
- Verhoeven, Ludo, Jan van Leeuwe, J. & Anne Vermeer. 2011. Vocabulary Growth and Reading Development across the Elementary School Years. *Scientific Studies of Reading* 15 (1). 8-25. doi:10.1080/10888438.2011.536125.
- Vermeer, Anne. 2004. The Relation between Lexical Richness and Vocabulary Size in Dutch L1 and L2 Children. In Paul Bogaards & Batia Laufer (eds.) *Vocabulary in a Second Language: Selection, Acquisition, and Testing* 173-189. Amsterdam: John Benjamins.
- Zadeh, Lotfi. 1965. Fuzzy sets. *Information and Control* 8. 338-353. doi:10.1016/S0019-9958(65)90241-X.
- Zimmermann, Hans Jürgen. 2001. *Fuzzy Set Theory and its Applications*. Boston: Kluwer Academic Publishers.