This is a preprint of an article accepted for publication in Discourse Processes

Predicting word maturity from frequency and semantic diversity

A computational study

Guillermo Jorge-Botana

Universidad Nacional de Educación a Distancia / Semantia Lab

Spain

&

Ricardo Olmos

Universidad Autónoma de Madrid / Semantia Lab

Spain

&

Vicente Sanjosé

Universidad de Valencia

Spain

ABSTRACT

Semantic word representation changes over different ages of childhood until it reaches its adult form. One method to formally model this change is the word maturity paradigm (Kireyev & Landauer, 2011). This method uses a text sample for each age, including adult age, and transforms the samples into a semantic space by means of Latent Semantic Analysis. The representation of a word at every age is then compared with its adult representation via computational maturity indices. The present study used this paradigm to explore to the impact of word frequency and semantic diversity on maturation indices. To do this, word maturity indices were extracted from a Spanish incremental corpus and validated, using correlation scores with Age of Acquisition and Word Difficulty indices from previous studies. The results show that both frequency and semantic diversity predict word maturity, but that the predictive capacity of frequency decreases as exposure to language increases. The latter result is discussed in terms of inductive processes suggested in previous studies (Landauer & Dumais, 1997).

Introduction

Mental representation of word meaning is not static like a storage repository, or like a dictionary of pre-established contents that is not modified by use (Elman, 1995). The representation of word meaning in people's minds changes over time and with experience throughout the life cycle. During their development, children are exposed to different lexical entries, giving rise to a more or less rapid development in the representations of different words, until an adult representation is acquired. Some measures exist in the literature to account for the time when a word is acquired and its representation is matured. The main measure is Age of Acquisition (AoA), or the age at which a word is learned at the first time. AoA norms consist of subjective measures collected by asking participants to estimate in years the age they

learned the word (see Ghyselinck, Lewis, & Brysbaert, 2004, for a review). However, AoA does not refer to the global representation of a word, but to the acquisition or first use of one of the meanings of that word (the meaning participants are thinking of at that time or the meaning specified to them). Thus, AoA does not capture the continuous and parsimonious process of acquisition of new meanings (Biemiller, Rosenstein, Sparks, Landauer, & Foltz, 2014).

In recent years, the acquisition, representation, and development of language have been approached from the point of view of computational psycholinguistics. One of these approaches is Latent Semantic Analysis (LSA). LSA has traditionally focused on modeling cognitive processes pertaining to a single final or adult representation of the lexicon (Jorge-Botana, León, Olmos, & Hassan-Montero, 2010; Kintsch, 2008; Kintsch & Bowles, 2002; Kintsch & Mangalath, 2011; Kintsch, Patel & Ericsson, 1999; Landauer, Foltz, & Laham, 1998). However, some studies have also studied the gradual acquisition of lexical knowledge via the inductive processes that operate on controlled exposure to language (Denhière & Lemaire, 2004; Landauer & Dumais, 1997). These studies analyzed change of similarities between words, as a result of gradual introduction of new texts, as an indirect index of word change. By means of this method, Landauer and Dumais (1997) suggested how a functional architecture that uses induction processes could solve the problem known as "poverty of the stimulus" or "Plato's problem", that is, how lexical knowledge increases without a massive and explicit exposure to words.

But changes in word representation itself (the change in vectors that represent words in the semantic space) had not been investigated until recently, when Kireyev and Landauer (2011) proposed the word maturity paradigm. In this paradigm, the semantic representation of the same word is longitudinally examined at different ages taking its final adult representation as the reference (Biemiller et al., 2014; Kireyev & Landauer, 2011; Landauer, Kireyev, & Panaccione, 2011). A set of texts is used to build the corresponding corpus in each age, in such

a way that the corpus for each age includes the corpora for previous ages. Following the usual LSA procedure (Deerwester, Dumais, Furnas, Landauer, & Harshman, 1990), a semantic space is created for each incremental corpus, including the adult one. The evolution over time of a specific word is then determined by the changes in the vector representing this word in the different semantic spaces. However, the between-spaces comparison has to be done in terms of same-basis vectors. Thus, a specific technique called "Procrustes procedure" is used to align the different semantic spaces (see Ross, 2004). The most used measure for these comparisons is the cosine. Kireyev and Landauer (2011) proposed the index they called Word Maturity (WM) defined as the cosine between any particular word at any particular age and the same word at adult age (in the final, larger semantic space).

The process of word meaning maturation is then mathematically modeled as the evolution of WM over time. Given that WM changes over time until full maturation, a continuous function can be adjusted for available values at specific points in time (i.e., the age ranges), obtaining what is known as the Maturity Curve (see an example in Figure 1) for each word (Kireyev & Landauer, 2011). There is a different Maturity Curve for every word considered. Within this paradigm, a word is considered as having acquired sufficient maturity when its WM function value surpasses a certain threshold, usually .65 (Biemiller et al., 2014). The value for the time corresponding to this threshold value is known as "Time To Maturity" (TTM) for the particular word analyzed. TTM(i) thus expresses the time required by a word *i* to reach a maturity of .65. TTM-based indices are good predictors of text difficulty (Nelson, Perfetti, Liben & Liben, 2011). In other words, TTM corresponds to the age in which a particular word is used in an adult-like manner. TTM values have also provided good evidence of criterion validity in exhaustive studies (Biemiller et al., 2014; Landauer et al., 2011) using the Age of Acquisition (AoA) of words as the external criterion.

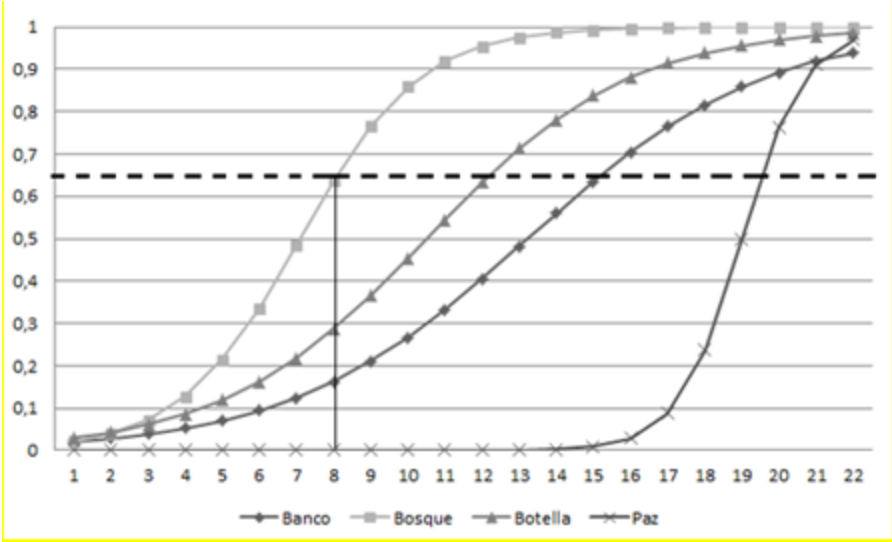

Figure 1. Maturity curves based on the WM point for four words from our study. The X axis represents ages and the Y represents the maturity measured as the cosine between the vector of the term in an intermediate space and the vector of the term in the adult space. "banco" in Spanish has several meanings, the most common ones being "bank", "bench", "pew", and "shoal"; "bosque" means "forest"; "botella" means "bottle"; "paz" means "peace". The TTM measure expresses the time required by a word to reach a maturity of .65, for instance, 8 years in "bosque".

In contrast to AoA, Word Maturity is a continuous and longitudinal measurement (the maturity of a word has a certain value at every age) which makes it possible to monitor changes in words over time and hence link them to two important variables involved in lexicon acquisition and maturity, namely frequency and semantic diversity. Frequency has been widely studied in experiments on lexical access (Ellis, 2002). Words that are very frequent in the child's linguistic environment are more likely to be understood and incorporated into the child's vocabulary. We expect that the evolution of word meaning in a person's mind depends upon that person's degree of exposure to that word. Higher-frequency words tend to get acquired earlier, which in turn protects those words from future changes (Brysbaert & Ghyselinck, 2006; Monaghan, 2014). Semantically diverse words are polysemic, i.e., they appear in widely different conceptual semantic fields. Such words are difficult to understand and use properly. In this study we use the semantic diversity index, a measure well suited to the

LSA environment and very sensitive to semantic representation (Hoffman, Ralph, & Rogers, 2013). We expect highly semantically diverse words to have a high difficulty index, a high AoA, and also to mature later by comparison to less semantically diverse words. This is shown by previous studies using analogous measures of semantic diversity over network analysis (Hills, Maouene, Maouene, Sheya, & Smith, 2009; Sailor, 2013; Steyvers & Tenenbaum, 2005).

Moreover, previous studies also suggest the AoA of a word is related to the number of connections of the old words in the network that are connected to it. This is called "preferential attachment" (Steyvers & Tenenbaum, 2005): the probability of acquiring the correct meaning of a new word will depend on the relationships of the already acquired words connected to the new word. Therefore, it can be expected that the power of frequency and semantic diversity to predict word maturity will be attenuated as a critical mass (Marchman & Bates, 1994) of already matured words exists, analogous to preferential attachment. For example, a new word would not need a high exposure to mature if the words related to it are already matured. This is the kind of mechanism that Landauer and Dumais (1997) suggested as a solution of the "poverty of stimulus" problem. Having the word maturity paradigm indices for each word, we can monitor the trajectory to maturity of each new word and observe whether frequency and semantic diversity account for it.

Goals and Hypotheses

This study has two main goals. First, we will design and implement a computational environment for generating the nested semantic spaces necessary to obtain maturation curves for a sample of words at different ages. The different maturation curves extracted will be validated. This validation will be performed bringing the TTM measure into relation with other concomitant and subjective lexical difficulty and acquisition variables available in Spanish (AoA indicators from various studies and difficulty indices based on expert criteria). Second, we will

investigate whether word frequency and word semantic diversity predict the maturation process of each new word in each age. The maturation index (WM) for each term that appears for the first time in each age was regarded as the dependent variable on a linear regressions whose predictor variables were the vector length for that terms, and their semantic diversity.

Given this second objective, our hypotheses are the following: (1) among words that appear for the first time at a specific age, more frequent words will have greater maturity, and (2) among words that appear for the first time at a specific age, those which are less semantically diverse will have greater maturity.

Given that the effect of words that are already mature on newly-acquired words increases over time, we propose two additional hypotheses: (3) The effect of frequency on maturity becomes attenuated as age increases, and (4) the effect of semantic diversity on maturity becomes attenuated as age increases.

Method

As was previously described, the first goal is to implement the word maturity paradigm on a longitudinal corpus in Spanish and validate the word maturity scores by means of convergent indices, like AoA and other difficulty indices. To this end, a set of text samples from different ages, including the adult space, was selected. A semantic space was "trained" for each of the age-specific set of texts, using LSA. The spaces for each intermediate age were then aligned with the adult space, using the Procrustes procedure, so that the terms in them could be compared. Then, the set of WM indices for each term (one for each age) were calculated, as well as one unique TTM for each term. The TTM was used for validation of the computational model, correlating it with the convergent indices described above. Finally, each WM was used as the dependent variable in a regression model whose predictor variables are vector length (the LSA measure of frequency) and semantic diversity.

The software Gallito 2.0 was used (Jorge-Botana, Olmos, & Barroso, 2013). Gallito 2.0 trained the different corpora under the following conditions: the window of process is the paragraph, a stop list of closed words was use, lemmatization was performed, frequencies were smoothed using log-entropy option, and dimensionality was reduced to 300 dimensions. We also deleted words that appeared in fewer than seven paragraphs in order to ensure a minimal representation of the terms analyzed. By means of this process, a matrix of terms and a matrix of paragraphs were obtained for each space.

Generation of the semantic spaces

Sampled ages and corpora. We sampled four age ranges in Spanish language development and selected a sample of texts for each one. The first sample was language from texts for children aged 0 to 9, the second one was from texts for children aged 10 to 12, the third one for children aged 13 to 16, and the final sample was a sample of adult language. On the basis of these samples, we compiled the cumulative corpora, i.e., the corpus for a specific age also contained the texts of the previous ages. Thus, we finally compiled four corpora that were labeled as 0-9, 0-12, 0-16 and adult.

Corpora composition. The corpora were taken from the Internet. For the 0-9 corpus, most of the texts were narratives like tales and stories. In intermediate ages, the proportion of academic texts increased in comparison to narrative texts. School books were also included (e.g., biology, chemical, history, physics). For the adult corpus, the Lexesp corpus (Sebastián-Gallés, Martí, Carreiras, & Cuetos, 2000) was used, which contains an extensive sample of texts from a wide range of topics, including literary texts, op-eds and press features, scientific papers, and essays. Using Gallito 2.0 (Jorge-Botana et al., 2013) the corresponding semantic spaces for each corpora were built. Space 0-9 included 7,286 terms and 31,722 paragraphs, Space 0-12 included 10,507 terms and 59,746 paragraphs,

Space 0-16 contained 13,209 terms and 75,434 paragraphs, and finally the adult space comprised 25,413 terms and 263,928 paragraphs.

Space alignment

The purpose of this step is to align any intermediate-age semantic space with the adult-LSA space so that word vectors from different spaces can be compared. This is done by the Procrustes procedure (see Kireyev & Landauer; 2011), a series of steps used to compare geometric figures in different scales and different spatial locations (Ross, 2004). The process comprises three main steps: (a) centering the figures to be compared (that is, putting one figure on the other so that both share a common point), (b) scaling the figures (equating the metrics of both shapes), and (c) rotating one shape onto the other one around the central point. If both figures are the same, they will match after these three steps. In the LSA case, the idea is to find a rotation of the term matrix for each intermediate age over the term matrix of the adult language so that the dimensions of both matrices are equated and all the term vectors of the intermediate age space can be compared to the term vectors of the adult age space. To find this rotation, initial geometric "figures" that are supposed to be the same in the both spaces, are crucial in this process. These two "figures" must be composed of vectors that in theory remain invariant between the intermediate space and the adult space (i.e. the vectors considered are the same, although represented in different vector basis). In other words, it has to be supposed that the distances between these vectors in the intermediate space (which forms the first figure) are proportionally preserved in the adult space (which forms the second figure). But the question is, are there any term-vectors that remain invariant until adult age? According to the definition of the model, all words change with age, so it is not a good idea to take terms as invariants. This problem is solved by means of a theoretical assumption. Kireyev and Landauer (2011) postulated that the semantic content of a paragraph (a context) is

much more stable than the contents of isolated terms because each term inside a paragraph is context-constrained. Therefore, they proposed that the set of invariant vectors in each pair of spaces (intermediate and adult) were constituted by the paragraph-vectors that are common to the intermediate space and to the adult space. The rotation angle is then calculated to match these paragraph-vectors in the first semantic space, into the adult LSA space. Once the rotation angle has been obtained, it is applied to the rest of vectors in the first LSA space (including the term matrix vectors). The resulting rotated space is expected to be aligned with the adult one, and then vector-vector comparisons, in our case term-term, are allowed (for instance, via cosine).

The alignment technique for semantic spaces was also performed by means of Gallito 2.0, which has a module specialized in aligning LSA spaces. Gallito 2.0 receives the whole two spaces as arguments. Then Procrustes is performed and generates an output: a list including each word in both spaces, and the cosine between its intermediate and adult representations (i.e. WM values).

WM, maturation curves, and TTM

Once the different alignments have been performed, we can calculate the cosine between the same term in each intermediate space and the adult space obtaining three WM for this term (WM for 0-9, WM for 0-12, WM for 0-16). With these three WM values plus zero and one (no maturation at all, and total maturation, respectively), we can obtain the maturation curve for every word. Logistics curves such as (1) were adjusted using Matlab R2014a.

$$f = \frac{1}{1 + e^{-(a+bx)}} \qquad (1)$$

In (1), *a* and *b* represent free parameters to be determined for every word; *x* is the age considered, and *f* is the maturation level of the term at that age. Some maturation curves are shown in Figure 1.

Logistic curves have properties that make them good candidates for this kind of variable: a) they start from a value of almost zero at their origin, but their maximum value is 1 (which is an asymptote for the function); b) the growth is not constant, it can be rapid for certain values. In fact, some authors regard the logistic curve as the most suitable function to simulate developmental mechanisms (Van Geert, 2014). These curves establish a relationship between age and the maturation level of each word on the basis of the set of five points provided for every word[1]. The analysis was restricted to 20,051 terms that had been assigned an Instituto Cervantes difficulty index (García Santa-Cecilia, 2000; Instituto Cervantes, 2006). Once the logistic functions for all words were obtained, the TTM index was calculated by applying the same criterion as for previous studies (Biemiller et al., 2014), that is, by finding the age value (*x* value in the function) corresponding to *f* = .65.

Lexical variables

Age of Acquisition, Instituto Cervantes difficulty index, and LEXIN recommendation index. To validate the text samples and provide criterion validity with the TTM measure, three indices related to lexicon maturation were used. We expected

[1] The ages on the X axis used for the adjustments were 0, 9, 12, 16 and 21. It was assumed that adult age would be represented by the value 21 to have an additional point in the adjustments. Even though this is an arbitrary value, the technique employed and the kind of function used (logistic functions) make it possible to minimize the impact of the specific value. Changing the value 21 to 25 or 30 would minimally change the adjusted functions, and thus the TTM values determined by them. In addition, the sample corpus for adult age contains novels, newspapers, and technical texts which correspond to an age of approximately 21.

these indices to be related to the TTM values. First, AoA values from various studies were used. This research included AoA data from six different studies and databases (Alonso, Fernandez & Díez, 2015; Álvarez & Cuetos, 2007; Cuetos, Samartino & Ellis, 2012; Izura, Hernández-Muñoz & Ellis, 2005; Moreno-Martínez , Montoro, & Rodríguez-Rojo, 2014; Davis & Perea, 2005). Second, the Instituto Cervantes difficulty index, which ranges from 1 (easy) to 4 (difficult), was used. Finally, the educational level recommendation indices from the LEXIN database were used (Corral, Ferrero, & Goikoetxea, 2009). The LEXIN database provides a difficulty rating for 13,184 words from a corpus of samples for beginning readers. As it only classifies those words into two levels (kindergarten and primary), we assume that the words in our database which are not included in those 13,184 are words are recommended for older readers. Thus, we added a new level, establishing a range of 1, 2, and 3 (kindergarten, primary and adult words respectively).

Semantic diversity and frequency. Both frequency and semantic diversity were calculated for every intermediate space and for the adult space. Both metrics are extracted from each of the semantic spaces. Frequency was computed by means of the vector length for each word. Vector length is a classic measure for LSA models. It indexes how much information LSA has about a word, and correlates strongly with frequency (Kintsch, 2001). The correlation between vector length in the adult space and frequency is .72 (N = 19,803) in the Espal database (Duchon, Perea, Sebastián-Gallés, Martí, & Carreiras, 2013) and .71 (N = 8,064) in the BuscaPalabras Database (Davis & Perea, 2005). In this study we used the semantic diversity index for each word, a measure well suited to the LSA environment and sensitive to semantic representations (Hoffman, et al., 2013). To obtain this index, the average cosine between every paragraph-vector pair in which a word appears is computed. The logarithm is then

applied to this scalar and the sign is changed to turn the resulting value from negative to positive. High values of this index represent high semantic diversity and are associated with an average cosine near 0.

Results

Validity

Text samples. If the text samples are well selected for each age, the lower age corpora should have a lower average AoA than the corpora for higher ages. The occurrence of a term for the first time is an ordinal variable: Each term appears for the first time either in the adult corpus (value = 4), or in the 0-16 corpus (value = 3), or in the 0-12 corpus (value = 2), or in the 0-9 corpus (value = 1). We therefore computed Spearman's correlation between this ordinal variable and the AoA values of each of the six studies from which they were obtained. The results (Table 1) show positive correlations between both variables, all of which are significant ($p < .01$). The text samples therefore represent relevant words for later analyses.

Table 1. Spearman correlations between the first time of occurrence in the semantic space and different metrics of age of acquisition

	AoA Alonso	AoA Moreno	AoA Izura	AoA Davis	AoA Álvarez	AoA Cuetos
Semantic space	.437 (5,145)	.503 (593)	.480 (384)	.387 (136)	.292 (295)	.360 (480)

Note: AoA Alonso = Alonso, Fernandez & Díez (2014); AoA Álvarez = Álvarez & Cuetos (2007); AoA Cuetos = Cuetos, Samartino & Ellis (2012); AoA Izura= Izura , Hernández-Muñoz & Ellis (2005); AoA Moreno = Moreno-Martínez , Montoro , Rodríguez-Rojo, (2014); AoA Davis = extracted from BuscaPalabras, Davis & Perea, (2005)

TTM. Correlations were computed between TTM and AoA values, Instituto Cervantes and LEXIN difficulty indices. Results appear in Table 2. Generally speaking,

correlations between TTM and the AoA values were strong and stable, exceeding r = .50 at times. The lowest correlation was obtained for the Alvarez and Cuetos (2007) study, but AoA from this study also was the least correlated with the other AoAs. The highest correlations were found for the most recent studies (Moreno-Martínez et al. (2014), Davis and Perea (2005), and Alonso et al. (2015). Correlations were lower in general, but remained stable and moderately strong. .361 was obtained using the Instituto Cervantes indices and .389 was obtained using the recommendations in the LEXIN database. In global terms, TTM displayed the same profile of correlations with the other variables as the Instituto Cervantes difficulty index.

Table 2. Pearson correlations between TTM values, age of acquisition and lexical difficulty

	1	2	3	4	5	6	7	8	9
1. TTM									
2. AoA Alonso	.459 (5,161)								
3. AoA Moreno	.502 (595)	.843 (490)							
4. AoA Izura	.449 (387)	.786 (465)	.825 (201)						
5. AoA Davis	.467 (139)	.798 (139)	.787 (84)	.625 (53)					
6. AoA Álvarez	.316 (305)	.617 (315)	.648 (157)	.162 (96)	.500 (137)				
7. AoA Cuetos	377 (498)	.738 (444)	.580 (149)	.700 (104)	.597 (131)	.317 (264)			
8. Cervantes	.361 (20,051)	.399 (5,162)	.564 (595)	.485 (387)	.396 (139)	.358 (306)	.312 (498)		
9. LEXIN	.389 (20,051)	.640 (7,039)	.489 (779)	.585 (497)	.242 (139)	.259 (318)	.526 (498)	.354 (20,178)	
Mean	15.63	6.91	3.92	7.03	4.22	84.15	5.04	2.67	2.32
SD	3.00	2.09	1.13	2.42	.71	44.88	1.22	.86	.83

Note. All the correlations were significant ($p < .001$). Sample size between brackets. Equivalences data: AoA Alonso = Alonso, Fernandez & Díez (2014); AoA Álvarez = Álvarez & Cuetos (2007); AoA Cuetos = Cuetos, Samartino & Ellis (2012); AoA Izura= Izura , Hernández-Muñoz & Ellis (2005); AoA Moreno = Moreno-Martínez , Montoro , Rodríguez-Rojo, (2014); AoA Davis = extraídos de BuscaPalabras, Davis & Perea, (2005); Cervantes = Índices de dificultad del Instituto Cervantes; Lexin = Difficulty índices from the Lexin database in Corral, Ferrero & Goikoetxea (2009).

Structural equation modeling. Validity was also assessed using structural equation models (SEMs). As six AoA metrics were available, a model with a single AoA (from now on, C-AoA) was tested, accounted for by the six available AoA indicators or metrics. The advantage of SEMs is that it makes it unnecessary to perform six regressions (one for every AoA measure) and, more importantly, measurement error associated with the particularities of each study is controlled for. The TTM variable was added to the AoA metrics to account for C-AoA in a first model.

In a second model, vector length in the adult space was also added as a predictor of C-AoA, as some studies have shown frequency to be directly associated with lexical maturation (Kireyev & Landauer, 2011). Because three AoA averages had an excessive percentage of missing values, the SEM models with six metrics did not converge. Therefore only the other three AoA metrics were used (Izura et al. 2005; Alonso et al., 2015; Moreno-Martínez et al., 2014). The two models appear in Figure 2.

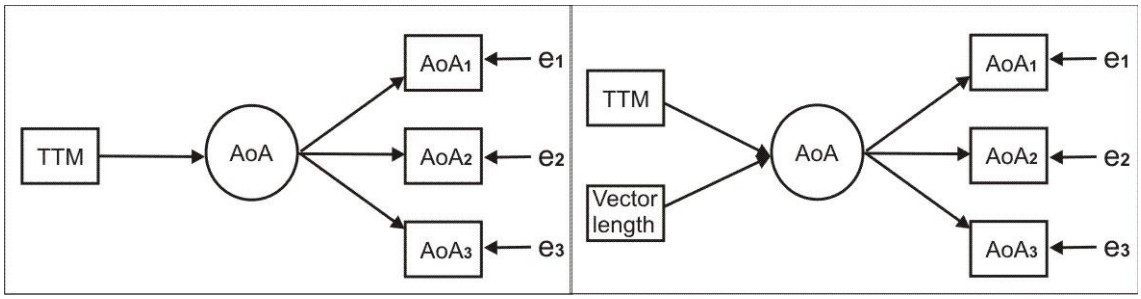

Figure 2. Two Structural Equation Models representation with and without vector length

The goodness of fit indices were adequate (RMSEA < .06, CFI and TLI > .95, SRMR < .08; Bentler, 1990), which supports the hypothesis that C-AoA is a single construct, thus simplifying results and controlling measurement error. Regarding the interpretation of the models, the TTM standardized factorial loading was high, positive, and statistically

significant (the higher the TTM, the higher the AoA of the term: .493 in Model 1 and .334 in Model 2. The factor loading for vector length was negative (the higher the vector length, the lower the age of acquisition), and displayed high values (higher in terms of absolute value than TTM scores). Table 3 displays the results of the models.

Table 3. Structural Equation Models to account for C-AoA

	Model 1			**Model 2**		
	Est.	*S.E.*	*p*	Est.	*S.E.*	*p*
TTM	.309 (.493)	.018---	<.001	.209 (.334)	.022	<.001
Module	---		---	-.397 (-.261)	.052	<.001
$\chi^2(df)$	7.679 (2) p = .022			8.367 (4) p = .079		
RMSEA (90% CI)	.047 (.015-.084)			.029 (.000-.057)		
CFI	.996			.997		
TLI	.987			.993		
SRMR	.013			.010		
R^2	.243			.294		

Note: The estimation method was Maximum Likelihood; N = 1,275. *Est.* = Unstandardized loading, (Standardized loading); *S.E.* = Standard Error; *p = p value*.

The proportion of variance accounted for (R^2) was .243 in Model 1 and .294 in Model 2. Therefore, adding vector length to the model improves the proportion of variance explained. When vector length is added to the model, AoA variance accounted for increases considerably, up to practically 30%. Moreover, incorporating vector length does not decrease the contribution of TTM (both predictors were in fact significant in Model 2).

Frequency and semantic diversity as predictors of WM

Having validated the measures, the main objective of this study was to investigate the effects of frequency and semantic diversity on WM at each age interval. Given that frequency and semantic diversity are values proper to each of the three semantic spaces generated (one for each age), the study was performed for each of them. But we tested only words that appear for the first time in each space. Thus, we tried to predict WM for

new words via linear regression with vector length and semantic diversity as predictor variables.

As there were three age corpora (0-9, 0-12 and 0-16 years), two dummy predictor variables were also created, the first one representing 0-9 years (code = 1) vs. others (code = 0), and the second one representing 0-12 years (code = 1) vs. others (code = 0). This left the third corpus (0-16 years) as the reference group.

Further, we investigated the extent to which age moderates the potential effects of vector length and semantic diversity on WM by computing the interaction effects between these variables and the age group dummy variables. Finally, we also included the interaction between vector length and semantic diversity as a predictor variable (previously, both predictors were mean-centered).

Table 4. Linear regression predicting word maturity by vector length, semantic diversity, age-group dummy variables and interaction terms

	Model 1		
Variable	*B*	*S.E.*	*β*
Constant term	.661***	.005	
Vector length	.110***	.009	.430***
Semantic diversity	-.122***	.018	-.224***
0-9 dummy age	-.161***	.005	-.421***
0-12 dummy age	-.136***	.005	-.308***
0-9 X vector length	.021*	.009	.075*
0-12 X vector length	.050***	.011	.065***
Semantic diversity X vector length	.057***	.006	.078***
0-9 X Semantic diversity	-.016	.019	-.023
0-12 X Semantic diversity	-.037	.020	-.034
R^2	.445		
Adjusted R^2	.444		

Note: * $p < .05$; ** $p < .01$; *** $p < .001$; *S.E.* = Standard Error

Table 4 shows the regression model. The model contains all the main effects and all the interaction terms (significant and non-significant). Results show a positive relationship between vector length and WM. However, the strength of this relationship varies by age, as evidenced by a significant interaction between 0-9 age and vector

length, as well as between 0-12 age and vector length. The standardized coefficients (β) show that the effect of vector length decreases with increasing age, but this decrease is substantive in the 0-16 space (the reference group in the analysis). The predictive power of vector length over WM particularly decreases at this age. This result suggests a kind of preferential attachment where the gradual emergence of a critical mass of mature words makes newly-appearing words in each space less conditional on their own frequency. The mere co-occurrence of these new words with those which are already stabilized in meaning can encourage their maturation, and even make them appear as already mature. To investigate the emergence of this gradual critical mass, the percentage of words appearing in previous spaces which appear already mature in the 0-12 and 0-16 space or which only mature in those spaces were calculated. Indeed, we found that 41% of the words that appear for the first time in the 0-9 space mature or arrive as mature in the 0-12 space. In addition, 93% of words that appear for the first time in the 0-9 or 0-12 space mature or arrive as mature in the 0-16 space. We also found that words with high semantic diversity tend to be less mature than words with low semantic diversity. But in contrast to the effect of vector length which decreases with age, the effect of semantic diversity remains significantly negative over age (no significant interaction term was found). However, a significant positive interaction between semantic diversity and vector length was found, indicating that vector length attenuates the negative effect of semantic diversity. In other words, if a word has high semantic diversity but also has occurred many times, the negative effect that semantic diversity has on maturation is limited. This suggests that such word has occurred enough times in the majority of the contexts that allow an adult representation.

Discussion

Model validation

The goal of this study was to explore the predictive power regarding maturity of two variables: vector length (analogous to frequency in LSA) and semantic diversity, both extracted from a set of age-incremental semantic spaces. But to do that, we had to previously replicate the word maturity paradigm (Biemiller et al, 2014; Kireyev & Landauer, 2011; Landauer et al., 2011) using our own incremental spaces, this time with Spanish text samples. To this end, we followed the entire process for extracting the three Word Maturation indices (WM) for each term. The WM points (plus absolute zero and a maximum represented by one) were used to adjust Maturation = *f*(age) logistic functions, and they were used to obtain the values of the time required for the maturation of each term, the TTM.

To validate the measures extracted, several analyses were performed. To begin with, texts that were collected to be part of the different spaces were studied in order to test if they reproduced the word acquisition order as adequately as possible. The correlational analyses show a rather high correspondence between the age ranges of the texts collected in this study and the average AoA, indicating that the samples are sufficiently representative. Secondly, once the suitability of the Spanish text samples was verified, our goal was to find evidence of criterion validity of TTM by means of external indicators from studies on people also linked to lexicon acquisition and maturation (AoA and difficulty index). Even taking into account that these indicators rate are somewhat different from the TTM indices, significant relations were found, both in the correlations and in the structural equation model obtained. In turn, we found significant relationships between TTM and AoA in addition to what is accounted by vector length. These results fit well with the results obtained by Kireyev and Landauer (2011) in English. However, it should be pointed out that this study has limitations with respect to the original study in English, because it has a narrower age range. The lack of a wider

age range (i.e. the lack of variability) may be why the correlations between TTM and these external indicators were not as high as in other studies (Biemiller et al., 2014). In any case, the correlations obtained in this study ranged between .30 and .50 corresponding to medium to high effect sizes.

From massive exposure to induction processes

The validation of our model ensures that the indices extracted from it are sufficiently reliable to be used as dependent variables towards the main goal of the study. We used vector length and semantic diversity for each age as predictors of the TTM of the new words appearing in each age. The results showed that vector length (frequency) and semantic diversity both contributed to account for TTM. The most frequent words at a specific age which have a lower semantic diversity reached higher maturation values, and thus had less time to full maturation. Otherwise put, words which have occurred a sufficient number of times at a given age and which have occurred in a focalized and repeated manner in the same semantic contexts are closer to their final representation than other words that are less frequent and/or less strongly linked to fixed semantic contexts. Following some previous studies, we can say that a word that has often occurred in the same contexts at a specific age could be a word for which the predominant meaning representation has already been internalized, and new information about it will elaborate on the same meaning already internalized, or else will not generate different meanings (Brysbaert & Ghyselinck, 2006; Monaghan, 2014). But since we found an interaction between semantic diversity and vector length, we also argue that the effect of vector length attenuates the negative effect that semantic diversity has over maturation. If a word is diverse but sufficiently frequent, the risk produced by such diversity decreases. One reason could be that such a word had

occurred in all its potential contexts. This could be the case for words that have no predominant meanings.

Another result of interest obtained in this study is that the explanatory capacity of vector length gradually decreases as age increases, with an abrupt downturn at space 0-16. This fact would suggest that words that have already matured have an effect on those which appear later. These later-appearing words do not need to occur massively in several contexts. They just have to occur a few times coinciding in texts with already mature words and the system will infer their meanings from them. This was suggested in a classical LSA paper modeling induction learning and inferences (Landauer & Dumais, 1997): A functional architecture like LSA makes it possible to solve the problem known as "poverty of the stimulus" or "Plato's problem": how people have more lexical knowledge than they could reasonably extract from the information that they are exposed to. The key is that the architecture enables inductions from the micro-relations between words. Moreover, this study concludes that in order to acquire knowledge about a word, the texts in which that word does not appear are also important. All of this is also in line with studies that measure the capacity of *n*-order relations to induce knowledge (Kontostathis & Pottenger, 2006; Lemaire & Denhière, 2006). Given our results, we could say that the induction processes are more powerful when a considerable proportion of words are already mature, allowing new terms to be understood even with low exposure. In other words, micro-inferences arise when there is a critical mass of words that are already mature to enable them. In fact, the concept of critical mass has been coined to refer to the number of mature words at a certain age which are required to produce a change in the speed of acquisition of linguistic skills (Marchman & Bates, 1994). Taking also Siegler`s overlapping waves model as a reference (Siegler, 1996), our simulation seems to support the idea that there could be a

moment in which older strategies to exploit data (in our case, exposure to different contexts) cease to be used while new ones increase and become stable (in our case, inferring meaning by means of coincidence with other words, without massive exposure). Both the effect of frequency and semantic diversity in this study are also corroborated by studies on lexicon acquisition using network analysis methodology (Steyvers & Tenenbaum, 2005; Hills et al., 2009; Sailor, 2013). Studies based on association networks find that the words with a high association index (which means low diversity in LSA, see Jorge-Botana & Olmos, 2014) have associated lower AoA values, that is to say, they are acquired earlier. In addition, some of those studies find that the likelihood of a new word being acquired earlier is proportional to the degree of the already existing words with which they have relations; or, more suggestively, that it is proportional to the stability of the representations that already exist and appear together with it. This has been given the classic term of preferential attachment in network studies (Steyvers & Tenenbaum, 2005). Following the same logic as these studies, in our LSA model, words acquire relations on the basis of co-occurrences in the paragraphs (the contextual unit selected in this study and in many other studies in the LSA environment) of the corpora at each age. In earlier ages, the words that appear most frequently are usually those that earlier acquire a mature representation. As the words in these early ages appear in texts with words that are not yet mature, and the number of words already mature is low, the frequency of the newly incorporated words is important for their maturation. Otherwise put, there is no critical mass of words that are already mature and stable occurring in the contexts of new words to intensify their maturation. Thus frequency accounts for a greater part of maturation in these initial stages. However, in later stages, the words that appear for the first time are accompanied by words that have already appeared in texts from previous ages and a

large part of them arrive as stable and mature at the current age, and thus the potential maturing action of frequency is minimized by an effect similar to preferential attachment (many already mature words allow new words to achieve an stable representation within a short time). Studies like Sailor (2013) also found a decrease in the likelihood of establishing new relations between words at later ages. This may be because the meaning of newly-appearing words is strongly constrained by the meaning of the "mass" of words to which they are attached, especially if those words are surrounded by a large network with stable links. That is, there is less plasticity in the system for learning later-acquired words (Monaghan & Ellis, 2010). Indeed, this seems to be, as in Landauer and Dumais classic study, an induction mechanism enabling rapid learning of word meaning without having to experience each word in each of its possible contexts, which would be costly and even unfeasible.

Conclusion

In conclusion, a working hypothesis for future research can be suggested: the maturation of a word does not depend only on its frequency, but also on its semantic diversity and the maturation level of the words that co-occur with it in the contexts in which it appears. In early ages, words with little diversity and high frequency are predicted to achieve stable meanings. When a "critical mass" of these first words is reached, new words will not need to occur massively in all of their potential contexts to achieve a stable meaning. They will be attached to the meanings of the pre-existing matured words with which they co-occur.

Authors' note

The authors also wish to thank members of Semantia Lab (http://www.semantialab.es) for freely providing the software Gallito Studio and gallitoAPI for this research and also

to Molino de Ideas for allowing us to use the API to automatically retrieve the Instituto Cervantes difficulty index for each word (https://store.apicultur.com/).

References

Adelman, J. S., Brown, G. D., & Quesada, J. F. (2006). Contextual diversity, not word frequency, determines word-naming and lexical decision times. *Psychological Science, 17(9),* 814–823.

Alonso, M. A., Fernández, A., & Díez, E. (2015). Subjective age-of-acquisition norms for 7,039 Spanish words. *Behavior Research Methods, 47,* 268–274.

Álvarez, B., & Cuetos, F. (2007). Objective age of acquisition norms for a set of 328 words in Spanish. *Behavior Research Methods, 39(3),* 377–383.

Bentler, P. M. (1990). Comparative fit indices in structural models. *Psychological Bulletin, 107,* 238-246.

Biemiller, A., Rosenstein, M., Sparks, R., Landauer, T.K., & Foltz P.W. (2014). Models of vocabulary acquisition: direct tests and text-derived simulations of vocabulary growth. *Scientific Studies of Reading, 18(2),* 130–154.

Brysbaert, M., & Ghyselinck, M. (2006). The effect of age of acquisition: Partly frequency related, partly frequency independent. *Visual Cognition*, *13(7-8)*, 992–1011.

Corral, S., Ferrero, M., & Goikoetxea, E. (2009). LEXIN: A lexical database from Spanish kindergarten and first-grade readers. *Behavior Research Methods, 41*, 1009–1017.

Cuetos, F., Samartino, T., & Ellis, A. W. (2012). Age acquisition norms from elderly spanish people: characteristics and the prediction of word recognition performance in Alzheimer's disease. *Psicologica: International Journal of Methodology and Experimental Psychology, 33(1)*, 59–76.

Davis, C.J., & Perea, M. (2005). BuscaPalabras: A program for deriving orthographic and phonological neighborhood statistics and other psycholinguistic indices in Spanish. *Behavior Research Methods, 37,* 665–671.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science, 41,* 391–407.

Denhière G., & Lemaire, B. (2004). A Computational Model of a Child Semantic Memory, in *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (CogSci'2004), 297–302. Chicago, Illinois. USA.

Duchon, A., Perea, M., Sebastián-Gallés, N., Martí, A., & Carreiras, M. (2013). EsPal: One-stop shopping for Spanish word properties. *Behavior Research Methods, 45,* 1246–1258.

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

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port & T. van Gelder (Eds.), *Mind as motion: Explorations in the dynamics of cognition*, (pp. 195–223). MIT Press, Cambridge, MA.

García Santa-Cecilia, Á. (2000). *El currículo de español como lengua extranjera*. Madrid: Edelsa.

Ghyselinck, M., Lewis, M. B., & Brysbaert, M. (2004). Age of acquisition and the cumulative-frequency hypothesis: A review of the literature and a new multi-task investigation. *Acta Psychologica, 115,* 43–67.

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009). Longitudinal analysis of early semantic networks preferential attachment or preferential acquisition? *Psychological Science, 20(6),* 729–739.

Hoffman, P., Ralph, M. A. L., & Rogers, T. T. (2013). Semantic diversity: a measure of semantic ambiguity based on variability in the contextual usage of words. *Behavior Research Methods, 45(3),* 718–730.

Instituto Cervantes (2006). *Plan curricular del Instituto Cervantes. Niveles de referencia para el español.* Madrid: Biblioteca nueva.

Izura, C., Hernández-Muñoz, N., & Ellis, A. W. (2005). Category norms for 500 Spanish words in five semantic categories. *Behavior Research Methods, 37(3),* 385–397.

Jorge-Botana, G. & Olmos, R. (2014). How lexical ambiguity distributes activation to semantic neighbors: Some possible consequences within a computational framework. *The Mental Lexicon, 9(1),* 67–106.

Jorge-Botana, G., León, J.A., Olmos, R., & Hassan-Montero, Y. (2010). Visualizing polysemy using LSA and the predication algorithm. *Journal of the American Society for Information Science and Technology, 61(8),* 1706–1724.

Jorge-Botana, G., Olmos, R., & Barroso, A. (2013). Gallito 2.0: a Natural Language Processing tool to support Research on Discourse. In *Proceeding of the Twenty-third Annual Meeting of the Society for Text and Discourse,* Valencia. Spain.

Kintsch, W. (2001) Predication. *Cognitive Science, 25,* 173–202.

Kintsch, W. (2008). Symbol systems and perceptual representations. In M. de Vega, A.M. Glenberg and A.C. Graesser (Eds.): *Symbols and Embodiment: Debates on Meaning and Cognition*, pp.145–164, Oxford University Press, Oxford, UK.

Kintsch, W., & Bowles, A. (2002). Metaphor comprehension: what makes a metaphor difficult to understand? *Metaphor and Symbol, 17,* 249–262.

Kintsch, W., & Mangalath, P. (2011). The construction of meaning. *Topics in Cognitive Science, 3,* 346–370.

Kintsch, W., Patel, V., & Ericsson, K.A. (1999). The role of Long-term working memory in text comprehension' *Psychologia, 42*, 186–198

Kireyev, K., & Landauer, T. K. (2011). Word maturity: computational modeling of word knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA.

Kontostathis, A., & Pottenger, W. M. (2006). A framework for understanding LSI performance. *Information Processing and Management, 42 (1),* 56–73.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of the acquisition, induction, and representation of knowledge. *Psychological Review, 104*, 211–240.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25(2-3),* 259–284.

Landauer, T. K., Kireyev, K., & Panaccione, C. (2011). Word Maturity: A new metric for word knowledge. *Scientific Studies of Reading, 15(1),* 92–108.

Lemaire, B., & Denhière, G. (2006). Effects of high-order co-occurrences on word semantic similarity. *Current Psychology Letters, 18(1),* 1–11.

Marchman, V., & Bates, E. (1994). Continuity in lexical and morphological development: A test of the critical mass hypothesis. *Journal of Child Language, 21(2),* 339–366.

McDonald, S., & Shillcock, R. (2001). Rethinking the word frequency effect: The neglected role of distributional information in lexical processing. *Language and Speech, 44,* 295–323.

Monaghan, P. (2014). Age of acquisition predicts rate of lexical evolution. *Cognition, 133(3),* 530–534.

Monaghan, P., & Ellis, A. W. (2010). Modeling reading development: Cumulative, incremental learning in a computational model of word naming. *Journal of Memory and Language, 63,* 506–525.

Moreno-Martínez, F. J., Montoro, P. R., & Rodríguez-Rojo, I. C. (2014). Spanish norms for age of acquisition, concept familiarity, lexical frequency, manipulability, typicality, and other variables for 820 words from 14 living/nonliving concepts. *Behavior Research Methods, 46(4),* 1088–97.

Nelson, J., Perfetti, C., Liben, D., & Liben, M. (2011). *Measures of text difficulty: Testing their predictive value for grade levels and student performance*. Technical report to the Bill and Melinda Gates Foundation, Seattle, WA.

Ross, A. (2004). *Procrustes analysis*, Technical Report, Department of Computer Science and Engineering, University of South Carolina, SC 29208. Retrieved from www.cse.sc.edu/~songwang/CourseProj/proj2004/ross/ross.pdf

Sailor, K. M. (2013). Is vocabulary growth influenced by the relations among words in a language learner's vocabulary? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39(5),* 16–57.

Sebastián-Gallés, N., Martí, M.A., Carreiras, M., & Cuetos, F. (2000). *LEXESP: Una base de datos informatizada del español. Barcelona.* Universitat de Barcelona.

Siegler, R. S. (1996). *Emerging minds*. New York: Oxford University Press.

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: statistical analyses and a model of semantic growth. *Cognitive Science, 29(1),* 41–78.

Van Geert, P. (2014). Dynamic modeling for development and education: from concepts to numbers. *Mind, Brain, and Education, 8(2),* 57–73.