big data and cognitive computing

MDPI

Article

Modelling Early Word Acquisition through Multiplex Lexical Networks and Machine Learning

Massimo Stella

Complex Science Consulting, Via Amilcare Foscarini 2, 73100 Lecce, Italy; massimo.stella@inbox.com

Received: 22 December 2018; Accepted: 19 January 2019; Published: 24 January 2019

Abstract: Early language acquisition is a complex cognitive task. Recent data-informed approaches showed that children do not learn words uniformly at random but rather follow specific strategies based on the associative representation of words in the mental lexicon, a conceptual system enabling human cognitive computing. Building on this evidence, the current investigation introduces a combination of machine learning techniques, psycholinguistic features (i.e., frequency, length, polysemy and class) and multiplex lexical networks, representing the semantics and phonology of the mental lexicon, with the aim of predicting normative acquisition of 529 English words by toddlers between 22 and 26 months. Classifications using logistic regression and based on four psycholinguistic features achieve the best baseline cross-validated accuracy of 61.7% when half of the words have been acquired. Adding network information through multiplex closeness centrality enhances accuracy (up to 67.7%) more than adding multiplex neighbourhood density/degree (62.4%) or multiplex PageRank versatility (63.0%) or the best single-layer network metric, i.e., free association degree (65.2%), instead. Multiplex closeness operationalises the structural relevance of words for semantic and phonological information flow. These results indicate that the whole, global, multi-level flow of information and structure of the mental lexicon influence word acquisition more than single-layer or local network features of words when considered in conjunction with language norms. The highlighted synergy of multiplex lexical structure and psycholinguistic norms opens new ways for understanding human cognition and language processing through powerful and data-parsimonious cognitive computing approaches.

Keywords: language acquisition; multiplex networks; cognitive network science; machine learning

1. Introduction

Language is a complex system mapping reality into ideas, concepts and words [1–5]. The mental representations of words in the human mind constitute the so-called mental lexicon, an idealised data system arising from the biological signalling of the language map in the human brain [6] and aimed at performing cognitive computing tasks [1–8].

Despite acting as a highly dynamical memory supporting language processes such as word acquisition [4,9,10], retrieval [11] and use [8], the mental lexicon cannot be directly reproduced in a lab setting because of its cognitive, ideal nature [1]. Although the structure of the mental lexicon cannot be directly quantified, decades of cognitive experiments have reconstructed it in several ways through psycholinguistic tasks [1,5,8,10,12]. Overwhelming evidence indicates that the mental lexicon is a highly organised data structure, where the layout of semantic [12–14] and phonological similarities [7,15] between concepts deeply influences cognitive processes such as word acquisition [4,9,10,16–21], language use [4,8,12,22], memory [12,23] and even creativity [24], stance and affect [25], expertise acquisition [26], writing styles [27] and performance under cognitive impairments [22,28] or ageing [29]. For this reason, understanding the cognitive mechanisms regulating the mental lexicon is key to shedding light on human cognition and, subsequently, achieving effective cognitive computing in automated tasks dealing with natural language processing [2].

Over the last decade, network science provided several important tools for exploring the mental lexicon and relating its organisation of word–word similarities with language-related processes [3–5,7,14]. In this context, the approach of multiplex lexical networks [19,30,31] allowed for combining semantic and phonological word–word similarities in a single, multi-layer network representation of the mental lexicon, with applications ranging from successfully predicting early word learning in English toddlers [19,20] to highlighting phonological priming effects [32] or predicting naming performance in clinical populations with cognitive disorders [28].

This manuscript focuses on over predicting early word learning with an innovative combination of psycholinguistic data, multiplex network features (exploiting the multi-layer structure of the mental lexicon) and automatic machine learning techniques.

Network models of language acquisition constitute a rather recent approach, enabled by the release of longitudinal datasets following language production in toddlers [10,18]. The most prominent dataset of this type is the CHILDES project (Child Language Data Exchange System), a multi-language corpus of the TalkBank system established by MacWhinney and Snow and storing data about language acquisition in both normative learners and clinical populations [33]. Based on CHILDES, the pioneering work by Hills and colleagues [16] modelled word learning through a synergy of psycholinguistic data and complex networks. That study focused on English speaking young children between age 16 and 30 months. It operationalised the learning environment of toddlers as a network of free associations between concepts, e.g., "dog" reminding a toddler of "kitty" in a free association task [34]. Hills and colleagues showed that the words that had more similarities in this associative learning environment were also more likely to be acquired at earlier ages, a phenomenon known also as *lure of the associates* (cf. also [10]). A subsequent study by Carlson and colleagues [35] found a similar effect also when the learning environment was modelled as a network of phonological similarities, where words are connected if differing by one phoneme [7]. Further investigations of co-occurrence and feature sharing networks by Beckage and colleagues reported that late learners did not display highly connected words in comparison with normative learners [4,9].

Stella and colleagues [19] combined together the semantic and phonological information represented by free associations, co-occurrences, semantic feature sharing and phonological similarities within a unique multiplex representation of the learning environment. The authors showed that the multiplex interplay between different aspects of semantics and phonology boosted the lure of associates, making the multiplex lexical representation of words in the mental lexicon a suitable tool for predicting early word acquisition. The best predictor of word learning was found to be closeness centrality, a network metric indicating how many associations are necessary on average to connect a concept with all others in a connected network component (cf. [19]).

However, the original approach by Stella and colleagues was limited in that, although it quantified the relative importance of individual layers in predicting word learning by *using a single predictor*, it did not provide a quantification of the relative influence of different network metrics and psycholinguistic features of words (e.g., frequency, length) *when combined together* for predicting word learning.

Therefore, the main aim of the current investigation is using prediction tools from machine learning in order to quantify the relative importance of psycholinguistic and both single- and multiplex network metrics for predicting early word learning in normative English learners.

The relevance of this approach is both in terms of its descriptive and quantitative potentials. From a descriptive point of view, this analysis can indicate the importance of integrating also network representations of the mental lexicon in automatic models of cognitive computing inspired by how humans perform in cognitive tasks such as language learning. From a quantitative point of view, the model implementing automatic prediction of word learning can be of relevance for potentially quantifying and grading early language skills in children.

2. Materials and Methods

This study combines machine learning techniques, multiplex network representations and psycholinguistic data, all aimed at achieving an accurate automatic prediction of early word acquisition. The following subsections report on the methodology used in the current investigation, providing quantitative definitions and cognitive interpretations. Notice that the second and the fourth sections contain also methodological discussions of past relevant psycholinguistic and network approaches modelling early word learning.

2.1. Predicting Normative Word Acquisition with Machine Learning

In this work, word acquisition was modelled as a learning trajectory r^*, where words were ranked according to their normative age of acquisition. This word ranking included 529 lexical items and was produced from the Child Language Data Exchange System (CHILDES), a longitudinal corpus accessible through the TalkBank system, based on transcripts of child-directed speech including hundreds of children and widely used for modelling early cognitive development and language acquisition [33].

For the current investigation, the focus was on production norms for normal learning English toddlers between 18 and 30 months of age, which quantify the percentage of normative learners producing and understanding a given word at a given age. For instance, 95% of the normative learners in the dataset produced (and understood) the word "mommy" at month 18, so that the production norm for "mommy" at month 18 was 0.95. In order to construct the normative ranking of word acquisition r^*, words were ranked in descending order of production norm, starting from month 18, proceeding in chronological order and assuming that words are known once at least 50% of the children in the dataset can produce them. Hence, the resulting ranking r^* represented a normative learning trajectory, where the words learned earlier by most children were ranked higher. Notice that this normative age of acquisition ranking was used also in a previous investigation [20] and it represents a proxy for the average normative learning of most toddlers.

It is well documented that individual normative learners might display deviations from the average behaviour, due to different learning environments and linguistic proficiency [16,18]. In order to account for variability in word learning, partially randomised learning rankings r_i^* were produced, where all words learned in a month were reshuffled at random using their production probabilities as weights. This random procedure attributed higher ranks to words with higher production probabilities only on average, introducing also small fluctuations in the normative learning trajectories. An ensemble of 20 partially randomised learning rankings were used for identifying the order with which normative learners acquired words during language learning. These rankings assume no language development dysfunction and refer only to normative word learning.

Once the rankings were built, developmental stages were quantified in terms of vocabulary size, namely how many words were acquired. At a given developmental stage n, when n words were learned in the first n positions of ranking r_i^*, a label attribute 1 was given to learned words and 0 to the remaining words. In this way, all 529 words were labelled as "learned" (1) or "not learned" (0) at a given developmental stage. A machine learning classification routine was then trained on a subset of the data (training set) and later used for predicting the learning labels of the remaining data (validation set). Among the set of most commonly used machine learning techniques, four were selected and tested for the current investigation, namely random forests, support vector machines, naive Bayes and logistic regression (cf. [36]). These classifiers were selected because they are the most popular and accurate classification methods for language investigations and also share similar parsimony in algorithm tuning compared to more delicate and data intensive approaches such as deep learning, as reported in the relevant literature, cf. [37]. At a general description level, all of the above four classifiers work by (i) learning the correlations in the features of training data provided in input, (ii) mapping these correlations over manifolds of measurable spaces, (iii) separating the space into different subcomponents, each one probilistically corresponding to a given label and (iv) quantifying the location of points into subcomponents, thus providing in output a classification of unobserved data. For the detailed explanation of each of the above classifiers, the interested reader is invited to see, for reference, [36]. All the classifiers were used as implemented in Mathematica 11.3 (Wolfram Research, Champaign, IL, USA), whose software implementation included also data standardization and automatic parameter tuning for enhanced classification performance.

In order to improve the robustness of results to statistical fluctuations in the assembly of training and validation sets, and also reduce overfitting, a 4-cross validation procedure was used. At the very same developmental stage n, the whole dataset of 529 words was split into a training set including 25% of words (i.e., 132 words) and a validation set made of the remaining 75% of words (i.e., 397 words). Both the training and validation sets were assembled in order to preserve the richness of learned words of the whole dataset, e.g., if at a given developmental stage 1/3 of the words were learned, both the training and the validation set included 1/3 of learned words, on average. A classification task was performed over two stages: (i) training the classifier over the training set and (ii) using the trained classifier for labelling the unseen data in the validation set. The performance of the classification of the validation set was measured in terms of *accuracy*, which is the fraction of correctly labelled words (either 1—"learned" or 0—"not learned") [36]. After accuracy was computed, at the same developmental stage, another training set of 132 words was built by using words from the previous validation set, while all words in the previous training set became part of the new validation set. Another classification task was performed and another accuracy measure was registered. Subsequently, words were reshuffled as in the previous round. A total of four classification tasks were performed for a given developmental stage over a given ranking r_i^*. For a given developmental stage n, the classification accuracy was then cross-validated (averaged) across the four runs and across the 20 normative rankings r_i^*.

Notice that, in every classification task, a word was represented as a vector of features. For the current investigation, this combination was considered: (i) psycholinguistic features of words as used in widely studied regression models of age of acquisition (e.g., features like word frequency, length, etc.) and (ii) network features of words derived from network representations of the mental lexicon and analysed in past approaches predicting word learning with complex networks. All the data used for machine learning training is available as Supplementary Material for this manuscript.

2.2. Psycholinguistic Data

As a baseline model for machine learning, only psycholinguistic features were used for representing words numerically. These features were selected based on the examination of previous approaches, relying on regression analysis, aimed at predicting the age of acquisition of words. The selected psycholinguistic features were:

- Word frequency, representing how commonly a given word occurs in language [23,38]. A network investigation of the mental lexicon by Steyvers et al. [39] showed that high frequency words tended not only to be acquired early on during cognitive development but also to reside in the core of semantic network representations of the mental lexicon. These results highlighted the interplay between frequency, word learning and mental lexicon structure. Furthermore, a recent large-scale study of Kuperman et al. [40] found that word frequency was the most important predictor for the age of acquisition of over 30,000 English words, spanning from early childhood up to adulthood. Although the current paper focuses mainly on early childhood, word frequency is still expected to play an important role for early word learning prediction, as also reported by previous investigations [16–18]. The large-scale Opensubtitles dataset [38], including 3×10^5 frequency norms from parsing over 2.7 million sentences from movie subtitles, was used as a proxy for quantifying word frequencies of English words. Importantly, the Opensubtitles dataset was found to be superior to other frequency norms from text corpora in explaining variance in reaction time from lexical decision tasks in English [41]. For this reason, the Opensubtitles dataset is suitable for further psycholinguistic investigations.
- Word length, counting the number of characters composing a given word. Kuperman et al. [40] found that word length strongly correlated (negatively) with the age of acquisition norms of

30,000 words, highlighting a tendency for shorter words to be acquired at earlier developmental stages. Empirical confirmations of the positive effect of short length over early word acquisition were found in previous investigations [16–18]. During learning, children also displayed a tendency to imitate shorter words with an increased likelihood [42], further underlining the importance of word length for the cognitive processes regulating early language development.

- Polysemy norms, counting the number of context-dependent definitions a word can have [23]. Polysemy represents a language ambiguity related to meaning, so that the same word (e.g., "character") can have different meanings according to its context of use (e.g., "character" can be related to "nature" but also to "font"). Recent investigations highlighted a tendency for children to learn words with low polysemy early on during language acquisition [43]. Polysemy was also related to the explosive emergence of language cores in a network representation of the mental lexicon [31], ultimately decreasing the semantic network distance of concepts in networks of semantic associations [13] and potentially impacting language processing [11]. Polysemy is of relevance to early word learning also because it positively correlates with the number of semantic associations a given concept can feature (cf. [31]). Recent studies [10,21] reported a tendency for children to learn novel words filling gaps in semantic networks. Therefore, polysemy is expected to encapsulate additional insights over word learning. In fact, polysemous words might better fill semantic gaps due to their richer contextual diversity. For a review about the relevance of context diversity in early word learning, the interested reader should refer to [17,44]. Polysemy here was quantified in terms of the number of different meanings attributed to a word by the curated WordData dictionary maintained by Wolfram Research and obtained by the intersection of several large-scale dictionaries. The documentation of WordData is available online [45] (last accessed: 01-23-2019). The same data was used also in a previous investigation [31].
- Being a noun, encoded as a binary variable. Previous studies showed that word category influenced early word acquisition in English [17] as well as in other languages [18]. By using networks of free associations, Hills et al. [17] showed that nouns with more associations/larger degrees tended to be learned earlier. The authors related this finding to a wider contextual diversity helping children in capturing the meaning of concepts. The same study also found the opposite effect for words not being nouns (e.g., verbs and adjectives), for which a reduced contextual diversity favoured early acquisition, instead. Word classes were computed from the CHILDES dataset.

Although there are additional psycholinguistic features affecting early word learning such as concept concreteness [18,40], these norms were not available for all of the 529 words in the current dataset. In order not to reduce the already small sample size of tested words, it was decided to limit the current investigation to the above word features only. This choice is further explained in the Discussion section.

2.3. Multiplex Lexical Network

Beyond the baseline model featuring only psycholinguistic features of words, the current investigation employed machine learning techniques based also on network features of words. These features were extracted from a multiplex lexical network representing semantic and phonological associations available to a child at the end of the considered time window for early word learning (i.e., 30 months of age). Such network representation of the mental lexicon of a child was suggested by Stella and colleagues in [19] for predicting early word learning. The powerfulness of this multiplex representation is that it considers multiple types of word–word similarities at once [30], differently from previous investigations of early word learning focusing only on either free associations [16], or co-occurrences [9,17], or semantic feature sharing [4,21], or phonological similarities [35] in isolation. The multiplex lexical network used here considers all of the above aspects of the mental lexicon as different layers of the same network. For a review focusing on multiplex networks, the interested reader is referred to [46]. As visualised in Figure 1, the adopted multiplex lexical network features words represented as network nodes and replicated across four network layers:

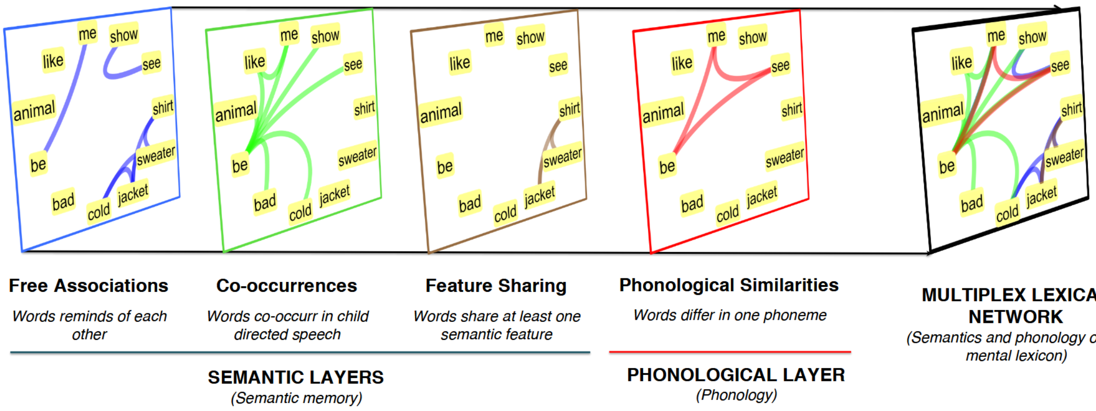

me
show
like
see
animal
shirt
be
sweater
bad
cold
jacket
Free Associations
Words reminds of each other
Co-occurrences
Words co-occurr in child directed speech
Feature Sharing
Words share at least one semantic feature
Phonological Similarities
Words differ in one phoneme
MULTIPLEX LEXICAL NETWORK
(Semantics and phonology of the mental lexicon)
SEMANTIC LAYERS
(Semantic memory)
PHONOLOGICAL LAYER
(Phonology)

Figure 1. Visualisation of a subgraph of the multiplex lexical network used for predicting early word learning. Words are represented as nodes and are replicated across all network layers. A network layer represents a given aspect of the mental lexicon, namely free associations, semantic feature sharing, co-occurrences and phonological similarities. The different layers are equivalent to projecting all links onto a single network while keeping links of different layer/colours as distinct, as highlighted in the right-most network.

1. Free associations built from the University of South Florida association norms [34]. In this layer, links between words indicate which words reminded of each other in a free association task (cf. [12,14,34]). Idiosyncratic associations were filtered out. Although free associations might contain both semantic and phonological associations, previous work on this layer has pointed out that its link overlaps with phonological similarities is negligible [19], so that it can be considered mainly a semantic layer in the current multiplex lexical network.
2. Co-occurrences built from child-directed speech from the CHILDES dataset [19]. In this layer, links between words indicate which words co-occurred in the same 5-g more than T times. The threshold T was chosen to be 45 in order for the co-occurrence layer to have analogous link density compared to the other semantic layers. Co-occurrences in this multiplex lexical network capture some information about the syntactic structure of speech and some semantic features of concepts [4,9].
3. Semantic feature sharing built from the McRae feature norms [47]. In this layer, words are linked if sharing at least one semantic feature (e.g., “being an animate object”).
4. Phonological similarities built from phonological International Phonetic Alphabet (IPA) transcriptions of words. In this layer, two words are connected if they differ by the addition/substitution/deletion of one phoneme. Phonological similarities capture some of the phonological information available in the mental lexicon [7,8,15,22,48].

For more details about the network structure of the above layers, the interested reader is referred to the original work by Stella and colleagues [19].

It is important to underline that the resulting multiplex network topology gives rise to a network where words are connected by different types of semantic or phonological links, as highlighted in the right-most network visualisation of Figure 1. Hence, the multiplex structure gives rise to denser neighbourhoods for words and alters significantly the way individual words are connected compared to considering individual layers. This difference was highlighted by Stella and colleagues, who found that the closeness of words on the whole multiplex network was a much more powerful predictor for early word learning than any other single-layer closeness centrality [19]. Multiplex closeness centrality was also found to outmatch other psycholinguistic variables such as frequency or word length in predicting early word learning.

The above results provide compelling evidence for the importance of multiplexity in predicting early word acquisition (cf. [19,20,49]), thus motivating the choice of coupling psycholinguistic features with network metrics in the current investigation.

2.4. Cognitive Interpretation of Network Measures

Three different kinds of network features of individual words were considered in the current analysis, each one relative to one scale of observation of the network structure:

- At the local scale of observation, word degree over the whole multiplex structure was chosen. This is the sum of the degrees of a word across all the multiplex layers [46]. For instance, with reference to Figure 1, the word "show" has multiplex degree equal to 2, since it has one free association link, one co-occurrence link, no feature sharing link and no phonological similarity. Degree is a local feature of words because it neglects the global network structure and it focuses on identifying the amount of lexical associations for a given word. On phonological networks, degree is also called neighbourhood density and it is informative of psycholinguistic data about lexical decision tasks, word confusability and memory retrieval [7,8,22]. On semantic networks, degree is similarly informative about memory retrieval from semantic memory in fluency tasks [12]. Multiplex degree (multidegree) combines phonological and semantic information [19].
- At the meso-scale level of observation, multiplex PageRank versatility was chosen. This variant of PageRank for multiplex networks was introduced by De Domenico and colleagues [50] in order to quantify the likelihood for a random walker navigating a multiplex network to visit a given node. The random walk explores a multiplex network by randomly crossing links within a given layer and teleporting from layer to layer. As a result, the exploration process is a good proxy of the community structure of a multiplex network, providing information on how central a given node is based on the centrality of its neighbours. On multiplex networks, versatility PageRank identifies those nodes acting as brokerage nodes for the flow of information across two layers (cf. [50]). An example in Figure 1 is the word "be", which is well connected in both the phonological and co-occurrence layers. Single-layer PageRank versatility was reported to predict fluency data from semantic network structure [51].
- At the global level of observation, where the whole network structure is measured, multiplex closeness centrality was chosen. Closeness centrality quantifies the average network distance of a node from all the other nodes in the same connected component of a network [52]. The multiplex variant exploits network distances combining links from different layers. No explicit cost for traversing semantic/phonological layers is considered. However, multiplex closeness still depends on inter-layer link–link correlations, so that this metric considers an implicit coupling between layers even when no explicit transition costs are considered. With reference to Figure 1, the word "be" is the one with the highest closeness centrality, since it is at lower network distance from most of the other network nodes. Notice that network distance indicates the smallest amount of associations connecting two nodes and it has been used for predicting data from both semantic [11] and phonological [48] relatedness tasks. More recently, network distance was used also as a predictor of knowledge acquisition [26], creativity [24,53] and picture naming performance in patients with aphasia [22,28].

The above network metrics operationalise in different ways the concept of network centrality and have been shown to capture different cognitive patterns. From the perspective of predicting early word learning, word centrality in the mental lexicon is relevant because of the well documented phenomenon of *lure of the associates*, i.e., the tendency for central words in the network of word–word associations to be acquired at earlier developmental stages [10,16,17].

It is important to underline that the above network metrics were reported to successfully predict early word acquisition significantly better than random word guessing by Stella and colleagues [19]. However, that past approach considered and evaluated each network metric in *isolation*, by measuring the overlap between the network-predicted learning trajectory and normative acquisition. While that study provided compelling evidence for the influential interplay between word acquisition and mental lexicon structure, it also neglected the impact that potential correlations or interaction effects between network metrics and psycholinguistic variables might have on predicting early word learning.

The current investigation aimed at overcoming the above limitation by training machine learning models combining psycholinguistic and network features of words. Comparisons were performed between the baseline model (with psycholinguistic features only) and the network-enhanced models (with psycholinguistic features and network features) in terms of accuracy for word learning prediction.

In order to test also the relevance of the multiplex approach in comparison with single-layer network approaches, beyond multidegree, multiplex PageRank versatility and multiplex closeness centrality, the analysis included also the degree of words in the free association layer only. This choice was motivated by such single-layer degree being the best network-based predictor of early word learning after multiplex closeness centrality in a previous study [19].

2.5. Interpreting the Adopted Research Methodology

To sum it up, the current approach modelled individual words as possessing specific features in relation to their learning environment. Such environment determined observable word features coming either from language in general [18,43] (e.g., psycholinguistic variables) or from the associative structure of the mental lexicon, which heavily influences word learning [2,10,16,21,35,44]. The mental lexicon represented here reflects how adults would semantically or phonologically associate those words known by most children early on during development, around 30 months of age, a representation adopted in previous successful investigations [19,20]. Differently from other embodied approaches mapping the physical reality in which a child is exposed to word instances [54], the network and language features represent two pivotal components modelling the learning environment of young toddlers in a cognitive way.

Overwhelming evidence from psycholinguistics has shown that the learning environment influences word learning in complex ways [4,16,21,42,55], so that its cognitive representation in terms of psycholinguistic and single-layer network features can be informative on the way most children acquire words [10,16–18,21,35,44]. The current approach builds on such evidence and investigates the interplay between psycholinguistic features of language and the multiplex structure of the mental lexicon [19,20,31] through machine learning techniques, accounting for individual variability and testing the mental lexicon from local to global scales of observation.

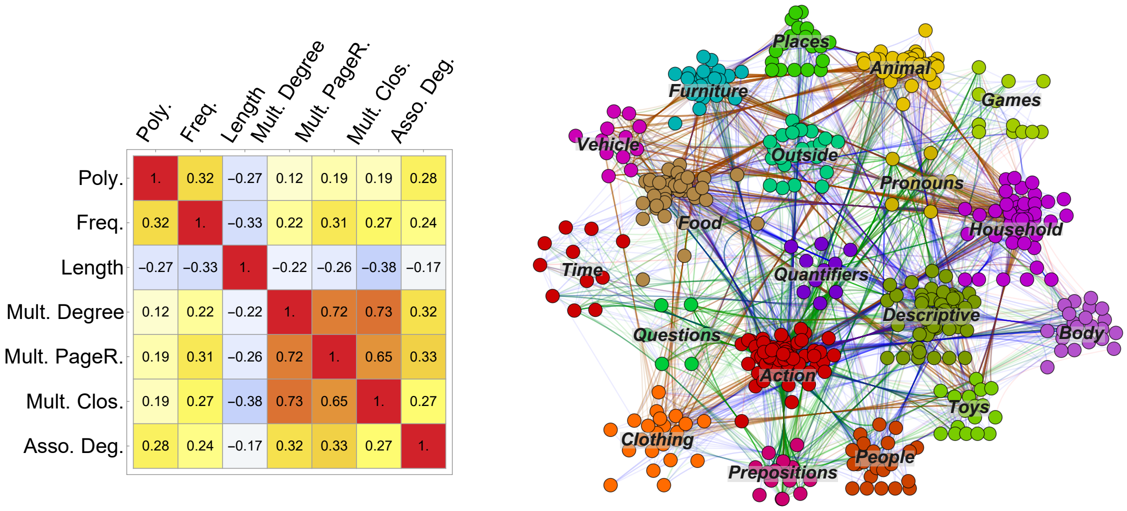
3. Results

3.1. Psycholinguistic and Network Features Are Correlated

Before proceeding with a classification task, it is important to quantify any potential correlation between input features, since these correlations can influence the accuracy of machine learning [36]. The left sub-panel of Figure 2 reports Kendall tau correlations between numerical psycholinguistic (frequency, length, polysemy) and network (multiplex closeness, multiplex degree, multiplex PageRank versatility and free association degree) features of words.

These correlations relate the topological features of words in the mental lexicon structure with the psycholinguistic characteristics of words as observed in language. Shorter words in the analysed dataset tend to be more frequent and with more contextual diversity (i.e., higher polysemy), in agreement with previous investigations [23]. Furthermore, shorter words also tend to be involved in more associations (higher degree) and occupy more central regions of the mental lexicon at both the meso-scale (higher PageRank) and global (closeness) levels of observation. Notice that, although all these features correlate with each other, no perfect correlation is reported. The lack of perfect correlations provides no a priori reasons for discarding redundant input features before training the machine learning classifiers. Nonetheless, it might still be that a given classifier is not able to distinguish between strongly correlated input measures, thus making necessary further checks. Multiplex network features are found to strongly correlate with each other, more than with other psycholinguistic features or with free association degree. This stronger correlation might indicate redundancy, i.e., these features might encode analogous information so that including one or more of them might not improve the accuracy of the machine learning classification.

Hence, different models including all psycholinguistic variables and only one multiplex feature should be tested first, before checking also the performances of a model adding all multiplex features at once.



	Poly.	Freq.	Length	Mult. Degree	Mult. PageR.	Mult. Clos.	Asso. Deg.
Poly.	1.	0.32	−0.27	0.12	0.19	0.19	0.28
Freq.	0.32	1.	−0.33	0.22	0.31	0.27	0.24
Length	−0.27	−0.33	1.	−0.22	−0.26	−0.38	−0.17
Mult. Degree	0.12	0.22	−0.22	1.	0.72	0.73	0.32
Mult. PageR.	0.19	0.31	−0.26	0.72	1.	0.65	0.33
Mult. Clos.	0.19	0.27	−0.38	0.73	0.65	1.	0.27
Asso. Deg.	0.28	0.24	−0.17	0.32	0.33	0.27	1.

Figure 2. Right: visualisation of the multiplex lexical network of normative learners at age 30 months. Links are coloured according to their type (free associations in blue, co-occurrences in green, feature sharing links in brown and phonological similarities in red). Nodes represent words and are clustered in categories according to the Communicative Development Inventory classification. **Left:** heat map of Kendall tau correlations between psycholinguistic and network features of words. Red (blue) indicates fully positive (negative) correlation. All correlations with a Kendall tau $|\tau| > 0.05$ were found to be statistically significant in a Kendall tau test at a 0.01 significance level.

3.2. Multiplex Lexical Structure Reflects Conceptual Categories

The set of words tested in the current investigation is organised in several subsets, each one representing a given word category. These sets of semantically related concepts are called *Communicative Developmental Inventories* (CDI) [33]. The number of words acquired by toddlers in every CDI has been shown to be informative about the future language ability of children [33,56]. Figure 2 (left) visualises the multiplex lexical network where words are clustered in CDI classes, e.g., toys, clothing, animals, etc..

When phonological and semantic links from the multiplex lexical network are combined and aggregated together, a modularity analysis (as implemented in IGraph [57]) reveals that words in CDI classes are more clustered than random expectation. In fact, the modularity [58] of words in CDI classes as defined by the empirical structure of semantic and phonological similarities is $\Phi = 0.161$. When rewiring at random all semantic and phonological links in the multiplex lexical network, the modularity of words in CDI classes vanishes. This indicates that the above modular structure is a distinct feature of the empirical multiplex lexical network. This modular structure indicates that the layout of semantic and phonological links reflects a tendency for words to cluster into specific lexical categories, as approximated by CDI classes. For instance, words related to time tend to connect mainly to each other across the whole multiplex network rather than to other words, in comparison with random expectation. The above modularity does not consider edge overlap between layers, but it still represents a reasonable approximation since edge overlap was found to be negligible in the investigated multiplex lexical network [19].

Hence, even though the multiplex lexical structure does not feature an explicit categorical layer, it implicitly provides additional information over the organisation of words and word–word similarities into specific semantic/phonological categories. Notice that these categories can display different connectivity patterns on different layers. For instance, the categories of animals and vehicles tend to cluster their links mainly in the feature sharing layer (i.e., brown links in Figure 2 left). Since categorical information has been shown to influence early word acquisition in pre-schoolers [59], it is important to underline that the current multiplex lexical network implicitly accounts for the presence of conceptual categories, mainly in terms of words clustering their links within CDI classes across multiplex layers.

Having highlighted the above correlations between word centralities and the presence of additional correlations in link assortment across CDIs in the multiplex lexical structure, the next step is to examine and interpret the results of the machine learning classification of early word acquisition.

3.3. Machine Learning Highlights the Influence of the Global Mental Lexicon Structure on Word Learning

Figure 3 reports the cross-validated accuracy of four different classifiers (naive Bayes, logistic regression, random forest and support vector machine) for different developmental stages. According to the regression analysis performed in [19], the age at which most normative learners reach a vocabulary size of 150 is around month 22, while a vocabulary size of 400 words corresponds to month 26. Hence, the predicted word acquisition is relative to the end of the second year of toddlers.

A baseline model is trained over the four psycholinguistic variables reported and described in the Methods, namely word frequency, word length, polysemy and is-a-noun flags. Accuracy is cross-validated according to a 4-fold cross-validation and averaged over 20 different partially randomised normative rankings r_i^*.

At a given developmental stage n, n words have been acquired. In cases when roughly half of the words in the dataset are acquired ($n = 264$) and the other half has not been acquired yet, a random classifier would have a theoretical accuracy of 0.5 in correctly guessing word learning. Remember that, by construction, the training and the validation sets have the same richness of learned words of the whole set of tested words, so that 50% of acquired words in the whole dataset corresponds to both the validation and training sets being made for 50% by learned words.

As reported in Figure 3, at the developmental stage, $n = 264$ the baseline model provides accuracy that is almost always higher than random expectation even when error bounds are taken into account. The most accurate baseline model is the one based on the *logistic regression* classifier, which provides an accuracy of $61.7 \pm 0.2\%$. At the same developmental stage, the baseline model performs worse with other classifiers, with the worst accuracy being provided by the random forest classifier (accuracy of $0.54 \pm 0.5\%$).

Adding network features to the baseline model almost always corresponds to improvements in the overall prediction accuracy across all the tested classifiers. However, it is important to underline that different network features impact accuracy in dramatically different ways.

Across all four classifiers, word centrality as encapsulated within multiplex closeness provides the highest increases in accuracy for word learning prediction, followed by free association degree.

The highest accuracy for $n = 264$ is reached with the model including all psycholinguistic features and closeness centrality in the logistic regression classifier (Figure 3, top right). When half of the words have been acquired, this combination of model and classifier achieves an accuracy of $67.7 \pm 0.2\%$. The same classifier with the model employing the degree of words in the free association layer, rather than multiplex closeness, achieves a lower accuracy of $65.2 \pm 0.1\%$. Both of these values are higher than the accuracy of the baseline model at the same developmental stage $61.7 \pm 0.2\%$. Hence, in the most accurate model predicting word learning, multiplex closeness centrality of words proves to be an important network feature for achieving accurate prediction results compared to mainstream psycholinguistic features of words. The degree of words in the free association layer provides worse prediction accuracy. Adding multiplex PageRank or multidegree to psycholinguistic features produces results close to the baseline model.

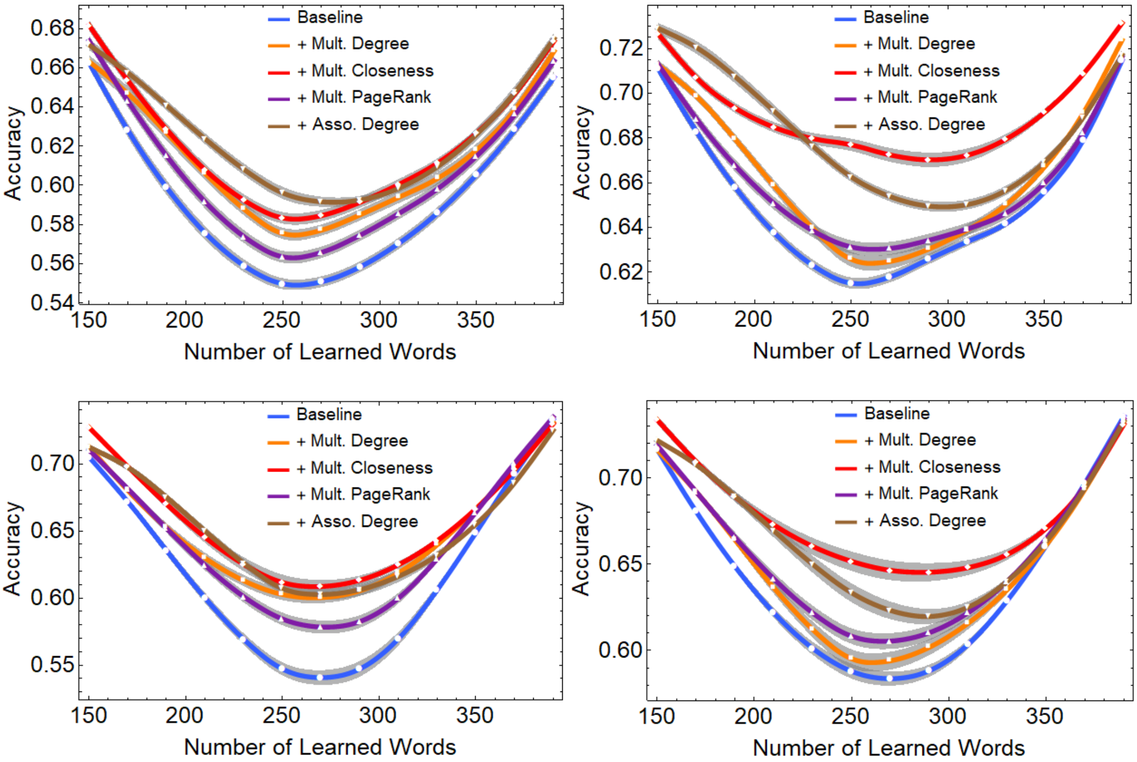


Figure 3. Accuracy for predicting early word learning with different models and different classifiers, namely naive Bayes (**top left**), logistic regression (**top right**), random forest (**bottom left**) and support vector machine (**bottom right**). The baseline model includes word frequency, length, polysemy scores and is-a-noun flags. Other models include also one of the following: multiplex degree, multiplex closeness centrality, multiplex PageRank and free association degree. Different developmental stages are considered spanning vocabulary sizes from 150 to 400 words, corresponding to months 22 and 26 of age. Error margins are reported in grey and represent standard deviations over the ensemble of partially randomised normative rankings r_i^* and cross-validation iterations (see Methods).

Analogous findings about the relevance of closeness centrality are found also in the support vector machine (SVM) classifier, where the model encompassing psycholinguistic features and multiplex closeness achieves an accuracy of $64.7 \pm 0.4\%$. The model adding free association degree achieves a slightly lower accuracy of $62.4 \pm 0.4\%$. As a reference, the baseline model provides an accuracy of $58.3 \pm 0.3\%$. The models based on psycholinguistic features and either multiplex degree or multiplex PageRank perform in between the baseline and the model with free association degree.

The random forest classifier consistently achieves the overall least accurate baseline model. In this classifier, single-layer free association degree, multiplex closeness centrality and even multiplex degree all provide analogous accuracy around $60.8 \pm 0.5\%$. This might indicate an inability for the random forest classifier to distinguish the different information encoded in the above network features at the different local and global, single-layer and multiplex scales. These poorer performances might be due to the well documented difficulty for random forest classifiers to perform well in small-sized datasets with a small number of predictor variables [36], like in the current investigation.

A similar issue with the scarcity of data is shown also by the naive Bayes classifier, whose accuracy in the developmental stage $n = 264$ always remains below 60%. Within this classifier, adding the free association degree to the baseline is marginally better (59.2 ± 0.3%) than closeness centrality (58.4 ± 0.3%). Multiplex PageRank and multidegree are performing slightly worse than closeness or degree but are still better than the baseline. Interestingly, the naive Bayes classifier is overall performing slightly better than random forests, even though the naive Bayes method assumes statistical independence of input features [36]. This assumption is not met by the strongly correlated data used here (see also Figure 2 left).

Interestingly, multiplex closeness provides the best results in the best classifier. The second best network feature to add to the baseline model is the degree in the free association layer. As discussed in the Methods, free association degree is a local feature capturing different local, single-layer information compared to the global, multiplex scope of closeness centrality. Why is there this interplay between multiplex closeness centrality and a single-layer network feature? The network visualisations reported in Figure 4 highlight how multiplex closeness and the degree of the free association layer are capturing specific centrality patterns of words, with some differences and some commonalities.

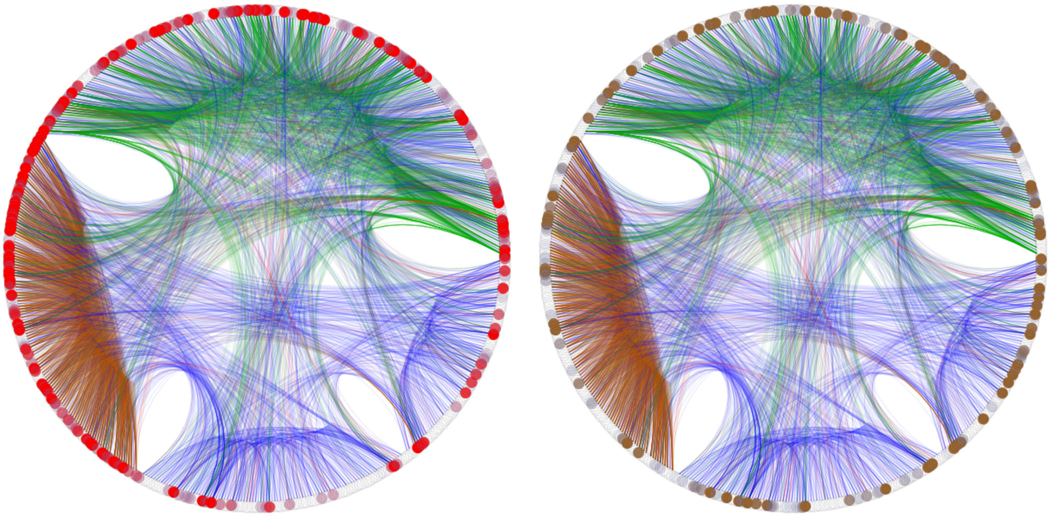


Figure 4. Visualisations of the multiplex lexical network of normative learners at age 30 months. Nodes are reported in a circular embedding and links are bundled according to a hierarchical edge bundling algorithm as implemented in Mathematica 11.3. Blue links indicate free associations, green is for co-occurrences, brown for semantic feature sharing and red links denote phonological similarities. **Left**: nodes in the 90th percentile of the multiplex closeness distribution are highlighted in red. **Right**: nodes in the 90th percentile of the free association degree distribution are highlighted in brown.

On the one hand, multiplex closeness considers as "central" a subset of nodes highly connected in the feature sharing layer, which are involved in fewer free associations (cf. Figure 4, left). Feature sharing links tend to cluster together in a network core of tightly connected nodes (cf. the k-cluster analysis from Stella et al. [19]). By definition, a core is densely connected and link density increases the amount of network paths connecting words. As a consequence, words in the core tend to be more central than words outside when metrics based on network distance such as closeness centrality are computed [60]. The short-cuts encoded in the network core provide also access to other network layers, since words in the feature sharing core are involved also in co-occurrences and free associations (cf. Figure 4). This combination of increased connectivity in the feature sharing layer and access also to other multiplex layers attributes high closeness centrality to most of the words in the feature sharing core. Hence, multiplex closeness heavily captures explicit aspects of semantic feature sharing, differently from the free association degree. This difference might be one of the factors leading to multiplex closeness predicting more accurately early word acquisition in the logistic regression classifier, compared to free association degree. This is compatible with several studies highlighting the importance of feature sharing over early word learning [10,21,55].

On the other hand, notice that, outside the small core of tightly connected words in the feature sharing layer, multiplex closeness and degree of the association layer capture similarly central words across the whole multiplex network. This commonality can motivate why free association degree performs analogously to multiplex closeness for some classifiers.

Figure 4 highlights an additional element of similarity between multiplex closeness and free associations. Closeness crucially depends on the presence of short-cuts, i.e., paths of shortest network distance between nodes [60]. The edge bundling algorithm in the figure highlights clusters of densely connected words across the whole multiplex network [60], which roughly correspond to the brown core of word–word similarities in the feature sharing layer, the green core of co-occurrences and a small red core of phonological similarities (overlapping with features sharing links in the figures). All these links tend to cluster together. Instead, free associations tend to spread across different clusters, joining together groups of words mainly connected according to feature sharing links, co-occurrences and phonological similarities. Hence, the free association layer acts as a *distributed network* connecting together and providing short range connection for access to the other network layers. As a consequence, multiplex closeness centrality strongly depends on the short-cuts provided by the free association layer. Differently put, multiplex closeness centrality captures analogous information compared to the degree in the association layer, as both these metrics strongly depend on the allocation of free association links. This visual pattern is quantitatively supported by previous results from Stella and colleagues [19], who found that free associations corresponded to the most influential layer, out of the four tested, in predicting word acquisition. Additional numerical experiments with edge betweenness, as implemented in IGraph, confirm quantitatively that free associations appear on average on more multiplex short-cuts between words than co-occurrences and feature sharing links at a significance level of 0.05. This difference vanished in randomly rewired multiplex networks preserving word degree. Since closeness is based on short-cuts, then free associations have the strongest impact on the multiplex closeness centrality of words in comparison with other semantic layers. Given this interplay between free associations and multiplex closeness centrality, it is expected for multiplex closeness and free association degree to display analogous, although not equivalent, patterns. Indeed, multiplex closeness includes more information about the global multiplex network than free association degree, an example being capturing signals from the feature sharing layer, as reported above. These patterns all support the above findings of higher accuracy for multiplex closeness within the best performing classifier.

It is important to underline that the best performing classifier is the one that captures the most information about the empirical data despite small sample size issues and the small number of features. In this study, the most accurate classifier is *logistic regression*, which consistently outperforms all other classifiers in terms of achieving the highest accuracy of prediction at the intermediate developmental stage $n = 264$ (when roughly half of the words in the dataset have been learned by most normative learners). Within this best performing classifier, the results reported in Figure 3 indicate that closeness centrality is the most important network feature, among those tested here, for predicting early word acquisition in combination with psycholinguistic word features. Hence, achieving an accuracy up to 68% in predicting early word learning is possible even with few predictor variables when the global multiplex network structure of the mental lexicon and mainstream psycholinguistic features of words are combined. Where previous studies indicated that the individual multiplex closeness centrality of words was already a powerful predictor of early word learning [19,20], the current results indicate that the correlations between psycholinguistic variables and multiplex closeness centrality allows for achieving the highest accuracy in predicting early word acquisition.

This finding is further corroborated by a feature importance analysis, where the same classifier is trained with the same set of features but with all entries in a feature reshuffled at random. The random reshuffling is expected to introduce noise in the classifier, which destroys correlations between one feature and all the other ones in the model. Hence, this reshuffling can potentially lower accuracy in the prediction task, even though, in the current case all features are highly correlated and reshuffling one of them is not expected to have drastic repercussions. Table 1 reports the relative decreases in accuracy from the original model when reshuffling individual model features. The largest decrease in accuracy was relative to reshuffling multiplex closeness centrality, thus making multiplex closeness the most important feature in the prediction model. Interestingly, among the psycholinguistic features considered here, the most important one is the is-a-noun flag. This is in agreement with previous studies based on the CHILDES corpus and highlighting different learning strategies for nouns and not nouns [17,18]. According to logistic regression, the noun flag is even more important than other psycholinguistic features such as polysemy and word length, which were both shown to influence early word acquisition [16–18,43]. Interestingly, reshuffling word frequency does not impact the accuracy of the prediction. This indicates that frequency correlates with all the other features in a way that reshuffling only it, while leaving the other features untouched, does not cripple classification. Instead, the fact that reshuffling multiplex closeness centrality impacts the most the model performance indicates that multiplex closeness, and hence the information about the global structure of the mental lexicon captured by it, provides *additional information* in relation with other psycholinguistic features. Hence, exploiting the synergy between multiplex closeness and psycholinguistic features is important for obtaining more accurate predictions of early word learning.

Table 1. Feature importance analysis for the logistic regression classifier and the model with frequency, length, polysemy, is-a-noun flag and multiplex closeness centrality trained for the developmental stage $n = 264$. Only one feature at a time is randomly reshuffled while the others are kept fixed. The average relative decrease in accuracy from the original model performance (accuracy 67.7 ± 0.2%) is computed. A location equivalence test(Kruskal-Wallis) between the accuracy of 50 reshuffles and original accuracy provides the reported p-values. Features are ranked in decreasing order of relative accuracy decreases.

Reshuffled Feature	Relative Decrease in Accuracy	*p*-Value
Closeness Centrality	−9%	$p < 10^{-5}$
Is-a-noun Flag	−6%	$p < 10^{-5}$
Word Length	−2%	$p = 3 \times 10^{-4}$
Polysemy Score	−2%	$p < 10^{-5}$
Frequency	−1%	$p = 0.4798$

The logistic regression classifier, as reported in Figure 3, highlights the superior performance of multiplex closeness centrality not only at the developmental stage $n = 264$ but consistently across the whole time window between $n = 230$ (23rd month of age) and $n = 350$ (25th month of age). Over this period, adding multiplex closeness centrality to the baseline model corresponds to the highest increases in accuracy, with an average absolute increase of +5–6%. All the other network features like free association degree, multiplex degree and multiplex PageRank consistently lead to worse accuracy across the above developmental phase. Although the absolute improvement on the baseline might look modest, it has to be underlined that such enhancement is relative to predicting a process, i.e., early word learning, which was considered as being *completely random* up until a few years ago (cf. [4,10]). In this way, the above patterns represent important quantitative evidence for the presence of consistent word learning strategies captured by multiplex closeness centrality and influencing a whole period of early language development.

Notice that all the tested classifiers and models display U-shapes in their accuracy versus the number of acquired words (cf. Figure 3). This is not a genuine psycholinguistic pattern but rather a consequence of the fact that with lower (higher) n, the number of "not-learned" ("learned") words is higher and this sample size inequality artificially increases the number of true negatives (positives) that even a random classifier would achieve, ultimately increasing the accuracy. It is for this reason that the above analysis mainly focused around the balanced developmental stage $n = 264$, where roughly half of the words are "learned" and half are "not-learned" and where a random predictor would achieve an accuracy of 50%.

Including *all* the above network and psycholinguistic features in a model with the logistic regression classifier did not lead to further improvements in accuracy. Hence, already the global metric of multiplex closeness centrality encapsulates most of the information necessary for predicting early word learning in comparison with other local statistics such as degree or meso-scale measures such as PageRank. From a computational perspective, this finding indicates that a model with only five features can already achieve the same accuracy of more redundant models with more (but not equally important) features.

4. Discussion

The above results open several insightful points for understanding the cognitive processes driving language learning and early assembly of the mental lexicon.

Firstly, the above accuracy is considerably higher than random expectation (accuracy of 50%) at developmental stage $n = 264$. This represents additional evidence that normative early word learning *is not* a random process where every word has the same chance of being learned during any stage of cognitive development, as considered until recently [10]. Instead, the machine learning models built in this investigation provide compelling evidence that already simple psycholinguistic word features bear enough cognitive information for predicting early word acquisition. This is in full agreement with rather recent independent studies investigating cognitive development in toddlers and reporting the existence of specific strategies of word acquisition [4,10,16,17,21,35]. However, the findings of this study do not represent only additional confirmation for the presence of word learning strategies but rather importantly indicate that already simple machine learning techniques and network approaches can capture such learning strategies. Importantly, some classifiers achieve higher accuracy than others. Although this crucially depends on the data at hand (e.g., classifying linearly separable variables, statistical independence, etc.), it is expected [36,37] for logistic regression to perform well in small-sized data samples like the current one, whereas support vector machines are known to be inefficient to train and, in fact, perform ineffectively in here.

Secondly, where previous mainstream studies focused mainly on features of words in language for predicting word learning [40], this investigation took inspiration from other approaches combining single-layer networks and psycholinguistic features [16,21] and explored the influence on word learning exerted by the phonological and semantic structures of word similarities in the human mental lexicon. This was possible thanks to a multiplex lexical network representation of the mental lexicon. Previous studies highlighted the relevance of multiplex lexical networks in understanding how combined semantic and phonological aspects of the mental lexicon influenced explosive cognitive development [31] and lexical restructuring [30], highlighted phonological priming effects and language iconicity [32] and even predicted production performance in picture naming tasks for clinical populations [28]. The multiplex lexical network used here constitutes the intersection of several independent datasets and it includes the semantic and phonological associations between words known to normal language learners of age 30 months. In agreement with the empirical findings of lure of associates [10,16,17,35], the final associative structure of the mental lexicon is considered to be informative about the way it was built during cognitive development and thus to provide insights about language learning [19,20,30]. This assumption motivated the choice of using the structural features of words in the multiplex lexical network as input data for the trained machine learning classifiers.

Thirdly, the mental lexicon structure was considered here at multiple levels of observation: the local level focused on how many semantic and phonological associations each word was involved into (multiplex degree [46]), the meso-scale level identified words acting as brokerage points facilitating communication between network layers (multiplex PageRank versatility [50]), the global level identified words of relevance for the flow of information on the whole multiplex network, on all layers simultaneously (multiplex closeness [19]). The above machine learning classifiers achieved the highest accuracy when the global structural feature of multiplex closeness centrality was added as model feature. PageRank versatility and degree produced more modest improvements. This comparison represents compelling evidence that *the global* structure of the mental lexicon influences word acquisition *in addition* to other features of language such as frequency, length or polysemy. This is an important advancement compared to previous network approaches adopting multiplex lexical networks but testing network and psycholinguistic features in isolation [19,20].

Then why closeness centrality? Ample evidence indicates that language processing works in an associative way [1,8,11,12,14,24,26,48,53]. For instance, semantic associations can influence the time it takes for understanding or producing a given concept when stimulated by another one, a phenomenon known as semantic priming [1]. Representing conceptual associations in the mental lexicon as a network naturally leads to the definition of a network distance between concepts, i.e., the smallest amount of associations that need to be traversed for connecting any two nodes [60]. Distance itself leads to a definition of the ease of access of a given node under a certain process of network exploration. Averaging all the network distances from a node to all its neighbours gives closeness centrality, which quantifies the ease of access of individual nodes under uniform diffusion processes [52,60]. In addition, PageRank can be interpreted as the ease of access of a single node while exploring a network. However, closeness and PageRank differ in the dynamical process representing network exploration. Closeness is relative to a uniform diffusion process where information has to flow along all shortest paths and from all nodes in a network. PageRank is relative to a random walker that can either traverse local connections or rather teleport from node to node. The teleportation is a key element that makes PageRank and closeness distinct. In a network of word–word similarities [39], words with higher closeness centrality can be reached by *all other concepts* traversing fewer word–word associations, on average, compared to words of lower closeness centrality [52]. Even though the exact cognitive mechanisms of mental exploration remain an open challenge [3,23,24], decades of psycholinguistic research have reported that the exploration dynamics of the mental lexicon related to language recall and use can be captured by theoretical models focusing on the layout of conceptual associations, e.g., spreading activation models (cf. [1]). Recent studies highlighted that network distance and closeness can be alternative theoretical descriptions to spreading activation models, capturing important cognitive patterns like predicting semantic similarity from networks of free associations [11], sound similarity from phonological networks [8,48] or picture naming errors in people with aphasia [22,28]. Hence, closeness centrality relates the global layout of network distances between concepts with spreading activation patterns about language processing and use. Since the main finding of the current study indicates that closeness is highly predictive of early word learning *in addition* to other psycholinguistic variables, then the link between closeness and spreading activation provides an indication that the way the mental lexicon is built during early language acquisition is related to the global organisation of concepts, in agreement with the lure of associates phenomenon [10].

It is important to underline that the current investigation has limitations. One key limit of machine learning techniques is that they do not directly provide evidence for the directionality of correlations between predictor and dependent variables [36]. In the current study, this is relative to the inability of determining potential inhibitory or rather facilitative effects of the individual predictor variables on word acquisition. For instance, the logistic regression classifier highlights that closeness centrality is *the most relevant tested feature* for predicting word learning at a macroscopic level. However, the feature importance analysis of the classifier does not provide information about potential trends in which words of higher (or lower) closeness are learned earlier (or later). It could be argued that this limitation is due to the main strength of machine learning. Classifiers learn statistical distributions from the data and do not assume internal models like regression analysis (e.g., a dependent variable scaling linearly with a predictor variable). This flexibility allows for achieving higher accuracy at the cost of not quantifying explicitly testable correlation effects as encapsulated in a regression model [36]. However, this issue can be resolved by cross checking machine learning results with additional approaches. For instance, by using overlap measures, Stella and colleagues [19,20] found that high multiplex closeness centrality in isolation predicted a large fraction of early acquired words, confirming a facilitative effect captured by closeness for the lure of associates in early word acquisition [10,16,17].

Hence, understanding the impact of individual predictor variables with machine learning requires a crucial step of cross-checking with additional experiments. In order to achieve a better understanding of early word acquisition, machine learning has to be used and interpreted not by itself but rather in comparison with other psycholinguistic experiments and data.

Another limitation of the current analysis is that it refers only to the average word learning strategies of a population of healthy children. Hence, the above results and interpretations should address cognitive patterns at a population level, with individual children potentially displaying deviations from the average behaviour reported here. Variability across children is expected. However, notice that the current study tried to account for individual variability by considering partially random reshuffled word learning trajectories (see Methods). The observed results reported here should therefore be robust to individual differences in normative language learners.

It is important to underline that the current investigation is descriptive, as it focuses on the acquisition of words directly from a given network representation rather than from the underlying statistical mechanisms leading to the attribution of meaning to concepts. Concept-meaning mapping is extensively studied in other frameworks, such as statistical learning [55] and Bayesian inference [59]. Instead, the approach of this investigation is inspired by previous studies successfully modelling word acquisition with complex networks without statistical learning [4,10,16,17,35]. These network approaches are related with statistical learning, since a recent study by Karuza and colleagues showed that statistical regularities of language (e.g., co-occurrence frequencies) are strongly related to network representations [55]. In this way, the current investigation can be considered a simpler yet relevant approach to modelling word learning in relation to statistical learning.

The results reported here rely on child-directed speech and consider as "learned" those words that children understand and produce. Both these requirements might limit to some extent the amount of words effectively assimilated and parsed by children, who might either understand a concept but ignore its phonetic structure or imitate the sound of words without understanding their meaning, i.e., language imitation (cf. [42]). Notice that the requirements for toddlers to understand and produce words in speech is in agreement with the adopted multiplex lexical network, where associative knowledge about word–word similarities is assumed to be present both on the phonological and semantic layers. Hence, the adopted definition of word learning is consistent with the used network representation. Notice also that such definition of word learning is atomistic, as it neglects other aspects of learning such as grammar acquisition or the development of syntactic relationships. For detecting these patterns, other frameworks measuring language proficiency should be implemented.

Another limitation of the current approach is that it does not explicitly keep into account the subordinate and superordinate categorical structure of language, e.g., a word being an animal or a vehicle or a vegetable, etc.. Extensive research in psycholinguistics has shown that the hierarchical organisation of concepts across levels of generality and categories influences both language processing and the acquisition of meanings [59]. However, the multiplex lexical network used here implicitly includes also information on word categories. As reported in the Results section, there is a tendency for concepts from the same category to cluster in the same communicative inventory over the aggregated structure of the multiplex lexical network of toddlers. Since both multiplex PageRank and closeness are expected to capture meso-scale or global patterns such as clustering in communities [61], the current machine learning predictions implicitly take into account also the organisation of words in cohesive subgroups representing semantic categories from the communicative development inventories (e.g., toys, house, animals, etc.).

The choice of potentially adding another layer of categorical word relationships leads to another issue of the current approach: the difficulty of selecting and merging together network layers of different nature, i.e., semantic and phonological. It has to be underlined that all the different types of word–word similarities reported in this study have been investigated separately in previous studies about word learning [4,9,16,17,21,35]. For the sake of comparing machine learning on multiplex lexical networks with previous approaches [19,20], no additional layer was included compared to the above-mentioned pioneering works.

Furthermore, it has to be underlined that adding data is not straightforward when working with machine learning approaches, as correlations and sample size issues have to be accounted for [36]. This is mentioned also in the next section, where potential future directions including additional data are outlined.

5. Conclusions

This investigation adopted machine learning techniques for quantifying the importance of psycholinguistic and network features in predicting early word acquisition. The best performing model, employing four psycholinguistic features and the global multiplex structure of the mental lexicon (modelled through multiplex closeness), reached an accuracy of 67.7 ± 0.2% in predicting early word learning around month 24. Excluding the network information of the mental lexicon and focusing only on word frequency, length, polysemy and word class led to a worse accuracy of 61.7 ± 0.2%.

An interesting direction for future work is represented by including recurrence for expanding the model. Recurrence represents the frequency of words as they appear not in extensive language corpora but in short windows (usually of the order of magnitude of minutes for speech production). Vosoughi et al. [62] found that the recurrence of words in child-directed speech was a valid predictor of early word learning. Their data was gathered for a single child and included statistically filtered recurrence estimates for 420 words. Recurrence norms were not adopted here because of the scope of the current investigation, dealing with the normative acquisition of a population of hundreds of children and requiring large sample sizes for the machine learning training. However, in the presence of potentially larger datasets (addressing the challenges of recording several children for years during early cognitive development), recurrence and other features of the learning environment such as spatial and temporal distinctiveness (cf. [54]) would be important features to consider for achieving more accurate automatic predictions of word learning.

Given the above balance between including more information while preserving enhanced accuracy of machine learning classifiers, two roads could mainly be pursued from here in future investigations. For the sake of reducing the complexity of word learning models, the current analysis with hierarchical edge bundling [60] shows that free associations are the backbone of the whole multiplex network, with free association degree providing worse results than multiplex closeness centrality but also requiring less data. Ample evidence indicates that networks of free associations are valid descriptors of psycholinguistic data concerning semantic memory [11,12,14,24,53]. However, free associations change their networked layout over cognitive development [29]. Hence, a simple approach for future work could be building specific datasets tailored for toddlers and more focused towards modelling early word acquisition through free associations only. The opposite road would be to pursue more realistic models including additional psycholinguistic features and network metrics for improving the accuracy of word prediction. This data-informed approach would potentially retain the multiplex network representation of the mental lexicon, since its closeness centrality provides the most accurate results. Furthermore, this approach would exploit the synergy of additional datasets, including features related to semantic gap filling, beyond the currently used polysemy, and impacting early word learning in toddlers [10]. The main problem with this approach is that intersecting additional datasets with the multiplex lexical network might add only redundant variables (e.g., features whose inclusion does not improve accuracy) and also reduce the amount of words for which all psycholinguistic and network features are fully available. For instance, for this investigation, considering concreteness norms from [63] led to a smaller sample size of 513 words, since for 14 of the original words tested above there was no concreteness entry available. Adding concreteness led to no statistically significant increase in accuracy. When adding features, further reducing the sample size of tested words could be detrimental for the training of the machine learning classifiers and potentially result in even worse accuracy. Hence, this road has to be pursued conditionally on several additional checks given by data availability and machine learning performance.

Further generalisation of the model could be achieved by considering additional network layers. In fact, the framework of multiplex lexical network is quite flexible in accommodating additional network layers. Importantly, the current investigation should be seen as a stepping stone rather than a definitive model, since it provides a general prototypical framework able to predict normative early word learning with an accuracy of up to 68% through a modest number of input features and a simplistic, descriptive assumption of word acquisition being driven only by the exposition of children to a learning environment represented as a multiplex network. The addition of more network layers and psycholinguistic features for explicitly modelling also other phenomena of early word learning such as gap filling [10,21], beyond the polysemy feature tested here, represents an exciting direction for future research.

Beyond word learning, also word imitation (without understanding) could be modelled with an analogous approach to the one reported here. A recent data-informed approach indicated that the structure of the mental lexicon also influences language imitation [42] in a similar fashion to the lure of associates detected for normative word learning [16]. However, considering that language imitation is mainly driven by phonological representations, the features of words in the phonological layer are expected to be better predictors of imitation compared to the multiplex structure, in contrast with what was found in the current study on word learning.

In addition, clinical applications would be interesting to pursue with this model. A study by Beckage and colleagues [9] reported structural differences in the organisation of co-occurrences between normative and late learners. In case these differences translated into different accuracy for a machine learning prediction of word learning, then the methodology described here could be a suitable approach for diagnosing late language emergence. The relevance of such approach would be to include the global structure of the mental lexicon (through closeness centrality) for such a classification.

All in all, the above results open new important methodological horizons for reaching a deeper understanding of language acquisition and cognitive development by combining together psycholinguistic norms, cognitive computing automatic techniques and the multi-layer structure of the mental lexicon represented as a multiplex lexical network.

Supplementary Materials: The following are available online at http://www.mdpi.com/2504-2289/3/1/10/s1, Dataset S1: Dataset of psycholinguistic and network features for reproducing the main results of this manuscript.

Funding: This research received no external funding.

Acknowledgments: The author acknowledges Nicole Beckage and Marco Cristoforetti for stimulating discussion about word learning and machine learning, respectively.

Conflicts of Interest: The author declares no conflict of interest.

References

1. Aitchison, J. *Words in the Mind: An Introduction to the Mental Lexicon*; John Wiley & Sons: New York, NY, USA, 2012.
2. Thomas, M.S.; Laurillard, D. *Computational Modeling of Learning and Teaching*; Handbook of Educational Neuroscience; Wiley-Blackwell: Oxford, UK, 2013.
3. Baronchelli, A.; Ferrer-i Cancho, R.; Pastor-Satorras, R.; Chater, N.; Christiansen, M.H. Networks in cognitive science. *Trends Cognit. Sci.* **2013**, *17*, 348–360. [CrossRef] [PubMed]
4. Beckage, N.M.; Colunga, E. Language networks as models of cognition: Understanding cognition through language. In *Towards a Theoretical Framework for Analyzing Complex Linguistic Networks*; Springer: New York, NY, USA, 2016; pp. 3–28.
5. Siew, C.S.; Wulff, D.U.; Beckage, N.; Kenett, Y. Cognitive Network Science: A review of research on cognition through the lens of network representations, processes, and dynamics. *PsyArXiv* **2018**, *9*. [CrossRef]
6. Thomas, M.S.; McLelland, J. *Connectionist Models of Cognition*; Cambridge University Press: Cambridge, UK, 2008.

7. Vitevitch, M.S. What can graph theory tell us about word learning and lexical retrieval? *J. Speech Lang. Hear. Res.* **2008**, *51*, 408–422. [CrossRef]
8. Vitevitch, M.S.; Siew, C.S.; Castro, N. *Spoken Word Recognition*; The Oxford Handbook of Psycholinguistics; MIT Press: Cambridge, MA, USA, 2018; p. 31.
9. Beckage, N.; Smith, L.; Hills, T. Small worlds and semantic network growth in typical and late talkers. *PLoS ONE* **2011**, *6*, e19348. [CrossRef] [PubMed]
10. Hills, T.T.; Siew, C.S. Filling gaps in early word learning. *Nat. Hum. Behav.* **2018**, *2*, 622. [CrossRef]
11. Kenett, Y.N. Going the extra creative mile: The role of semantic distance in creativity–Theory, research, and measurement. In *The Cambridge Handbook of the Neuroscience of Creativity*; Cambridge University Press: Cambridge, UK, 2018.
12. De Deyne, S.; Navarro, D.J.; Storms, G. Better explanations of lexical and semantic cognition using networks derived from continued rather than single-word associations. *Behav. Res. Methods* **2013**, *45*, 480–498. [CrossRef] [PubMed]
13. Sigman, M.; Cecchi, G.A. Global organization of the Wordnet lexicon. *Proc. Natl. Acad. Sci. USA* **2002**, *99*, 1742–1747. [CrossRef]
14. De Deyne, S.; Kenett, Y.N.; Anaki, D.; Faust, M.; Navarro, D. Large-scale network representations of semantics in the mental lexicon. In *Big Data in Cognitive Science: From Methods to Insights*; CRC Press: Boca Raton, FL, USA, p. 174–202.
15. Stella, M.; Brede, M. Patterns in the English language: phonological networks, percolation and assembly models. *J. Stat. Mech. Theory Exp.* **2015**, *2015*, P05006. [CrossRef]
16. Hills, T.T.; Maouene, M.; Maouene, J.; Sheya, A.; Smith, L. Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychol. Sci.* **2009**, *20*, 729–739. [CrossRef]
17. Hills, T.T.; Maouene, J.; Riordan, B.; Smith, L.B. The associative structure of language: Contextual diversity in early word learning. *J. Mem. Lang.* **2010**, *63*, 259–273. [CrossRef]
18. Braginsky, M.; Yurovsky, D.; Marchman, V.A.; Frank, M.C. From uh-oh to tomorrow: Predicting age of acquisition for early words across languages. In Proceedings of the 38th annual conference of the Cognitive Science Society, Philadelphia, PA, USA, 10–13 August 2016; pp. 1691–1696.
19. Stella, M.; Beckage, N.M.; Brede, M. Multiplex lexical networks reveal patterns in early word acquisition in children. *Sci. Rep.* **2017**, *7*, 46730. [CrossRef] [PubMed]
20. Stella, M.; De Domenico, M. Distance entropy cartography characterises centrality in complex networks. *Entropy* **2018**, *20*, 268. [CrossRef]
21. Sizemore, A.E.; Karuza, E.A.; Giusti, C.; Bassett, D.S. Knowledge gaps in the early growth of semantic feature networks. *Nat. Hum. Behav.* **2018**, *2*, 682. [CrossRef] [PubMed]
22. Vitevitch, M.S.; Castro, N. Using network science in the language sciences and clinic. *Int. J. Speech-Lang. Pathol.* **2015**, *17*, 13–25. [CrossRef] [PubMed]
23. Ferrer-i Cancho, R.; Vitevitch, M.S. The origins of Zipf's meaning-frequency law. *J. Assoc. Inf. Sci. Technol.* **2018**, *69*, 1369–1379. doi:10.1002/asi.24057. [CrossRef]
24. Kenett, Y.N.; Levy, O.; Kenett, D.Y.; Stanley, H.E.; Faust, M.; Havlin, S. Flexibility of thought in high creative individuals represented by percolation analysis. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 867–872. [CrossRef] [PubMed]
25. Stella, M.; Ferrara, E.; De Domenico, M. Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl. Acad. Sci. USA* **2018**, *115*, 12435–12440. [CrossRef]
26. Siew, C.S. Using network science to analyze concept maps of psychology undergraduates. *Appl. Cognit. Psychol.* **2018**. doi:10.1002/acp.3484. [CrossRef]
27. Amancio, D.R. Authorship recognition via fluctuation analysis of network topology and word intermittency. *J. Stat. Mech. Theory Exp.* **2015**, *2015*, P03005. [CrossRef]
28. Castro, N.; Stella, M. The multiplex structure of the mental lexicon influences picture naming in people with aphasia. *PsyArXiv* **2018**. [CrossRef]
29. Wulff, D.U.; Hills, T.; Mata, R. Structural differences in the semantic networks of younger and older adults. *PsyArXiv* **2018**. [CrossRef]
30. Stella, M.; Brede, M. Mental lexicon growth modelling reveals the multiplexity of the English language. In *Complex Networks VII*; Springer: New York, NY, USA, 2016; pp. 267–279.

31. Stella, M.; Beckage, N.M.; Brede, M.; De Domenico, M. Multiplex model of mental lexicon reveals explosive learning in humans. *Sci. Rep.* **2018**, *8*, 2259. [CrossRef] [PubMed]
32. Stella, M. Cohort And Rhyme Priming Emerge From The Multiplex Network Structure Of The Mental Lexicon. *Complexity* **2018**, *2018*, 6438702. [CrossRef]
33. MacWhinney, B. *The CHILDES Project: Tools for Analyzing Talk, Volume II: The Database*; Psychology Press: London, UK, 2014.
34. Nelson, D.L.; McEvoy, C.L.; Schreiber, T.A. The University of South Florida free association, rhyme, and word fragment norms. *Behav. Res. Methods Instrum. Comput.* **2004**, *36*, 402–407. [CrossRef] [PubMed]
35. Carlson, M.T.; Sonderegger, M.; Bane, M. How children explore the phonological network in child-directed speech: A survival analysis of children's first word productions. *J. Mem. Lang.* **2014**, *75*, 159–180. [CrossRef] [PubMed]
36. Murphy, K.P. *Machine Learning: A Probabilistic Perspective*; The MIT Press: Cambridge, MA, USA, 2012.
37. Pranckevičius, T.; Marcinkevičius, V. Comparison of naive bayes, random forest, decision tree, support vector machines, and logistic regression classifiers for text reviews classification. *Baltic J. Mod. Comput.* **2017**, *5*, 221. [CrossRef]
38. Barbaresi, A. Language-Classified Open Subtitles (LACLOS): Download, Extraction, and Quality Assessment. Ph.D. Thesis, BBAW, Berlin, Germany, 2013.
39. Steyvers, M.; Tenenbaum, J.B. The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognit. Sci.* **2005**, *29*, 41–78. [CrossRef] [PubMed]
40. Kuperman, V.; Stadthagen-Gonzalez, H.; Brysbaert, M. Age-of-acquisition ratings for 30,000 English words. *Behav. Res. Methods* **2012**, *44*, 978–990. [CrossRef]
41. Brysbaert, M.; New, B. Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behav. Res. Methods* **2009**, *41*, 977–990. [CrossRef]
42. Zamuner, T.S.; Thiessen, A. A phonological, lexical, and phonetic analysis of the new words that young children imitate. *Can. J. Linguist./Rev. Can. Linguist.* **2018**, 1–24. [CrossRef]
43. Casas, B.; Català, N.; Ferrer-i Cancho, R.; Hernández-Fernández, A.; Baixeries, J. The polysemy of the words that children learn over time. *arXiv* **2016**, arXiv:1611.08807.
44. Engelthaler, T.; Hills, T.T. Feature biases in early word learning: network distinctiveness predicts age of acquisition. *Cognit. Sci.* **2017**, *41*, 120–140. [CrossRef] [PubMed]
45. Available online: https://reference.wolfram.com/language/note/WordDataSourceInformation.html (accessed on 24 January 2019).
46. Battiston, F.; Nicosia, V.; Latora, V. The new challenges of multiplex networks: Measures and models. *Eur. Phys. J. Spec. Top.* **2017**, *226*, 401–416. [CrossRef]
47. McRae, K.; Cree, G.S.; Seidenberg, M.S.; McNorgan, C. Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* **2005**, *37*, 547–559. [CrossRef] [PubMed]
48. Goldstein, R.; Vitevitch, M.S. The influence of closeness centrality on lexical processing. *Front. Psychol.* **2017**, *8*, 1683. [CrossRef] [PubMed]
49. Stella, M.; Brede, M. Investigating the phonetic organisation of the English language via phonological networks, percolation and Markov models. In *Proceedings of ECCS 2014*; Springer: New York, NY, USA, 2016; pp. 219–229.
50. De Domenico, M.; Solé-Ribalta, A.; Omodei, E.; Gómez, S.; Arenas, A. Ranking in interconnected multilayer networks reveals versatile nodes. *Nat. Commun.* **2015**, *6*, 6868. [CrossRef]
51. Griffiths, T.L.; Steyvers, M.; Firl, A. Google and the mind: Predicting fluency with PageRank. *Psychol. Sci.* **2007**, *18*, 1069–1076. [CrossRef]
52. Borgatti, S.P. Centrality and network flow. *Soc. Netw.* **2005**, *27*, 55–71. [CrossRef]
53. Kenett, Y.N. What can quantitative measures of semantic distance tell us about creativity? *Curr. Opin. Behav. Sci.* **2019**, *27*, 11–16. [CrossRef]
54. Roy, B.C.; Frank, M.C.; DeCamp, P.; Miller, M.; Roy, D. Predicting the birth of a spoken word. *Proc. Natl. Acad. Sci. USA* **2015**, *112*, 12663–12668. [CrossRef]
55. Karuza, E.A.; Thompson-Schill, S.L.; Bassett, D.S. Local patterns to global architectures: influences of network topology on human learning. *Trends Cognit. Sci.* **2016**, *20*, 629–640. [CrossRef] [PubMed]

56. Fenson, L.; Dale, P.S.; Reznick, J.S.; Bates, E.; Thal, D.J.; Pethick, S.J.; Tomasello, M.; Mervis, C.B.; Stiles, J. Variability in early communicative development. *Monogr. Soc. Res. Child. Dev.* **1994**, *59*, 1–173. [CrossRef] [PubMed]
57. Csardi, G.; Nepusz, T. The igraph software package for complex network research. *InterJournal Complex Syst.* **2006**, *1695*, 1–9.
58. Newman, M.E. Modularity and community structure in networks. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 8577–8582. [CrossRef]
59. Xu, F.; Tenenbaum, J.B. Word learning as Bayesian inference. *Psychol. Rev.* **2007**, *114*, 245. [CrossRef] [PubMed]
60. Newman, M. *Networks*; Oxford University Press: Oxford, UK, 2018.
61. Fortunato, S.; Latora, V.; Marchiori, M. Method to find community structures based on information centrality. *Phys. Rev. E* **2004**, *70*, 056104. [CrossRef] [PubMed]
62. Vosoughi, S.; Roy, B.; Frank, M.; Roy, D. Contributions of prosodic and distributional features of caregivers' speech in early word learning. Proceedings of the Annual Meeting of the Cognitive Science Society, Portland, OR, USA, 11–14 August 2010; Volume 32.
63. Brysbaert, M.; Warriner, A.B.; Kuperman, V. Concreteness ratings for 40 thousand generally known English word lemmas. *Behav. Res. Methods* **2014**, *46*, 904–911. [CrossRef]