Employing distributional semantics to organize task-focused vocabulary learning

Haemanth Santhi Ponnusamy **Detmar Meurers**

University of Tübingen, Germany
{hsp,dm}@sfs.uni-tuebingen.de

Abstract

How can a learner systematically prepare for reading a book they are interested in? In this paper, we explore how computational linguistic methods such as distributional semantics, morphological clustering, and exercise generation can be combined with graph-based learner models to answer this question both conceptually and in practice. Based on highly structured learner models and concepts from network analysis, the learner is guided to efficiently explore the targeted lexical space. They practice using multi-gap learning activities generated from the book. In sum, the approach combines computational linguistic methods with concepts from network analysis and tutoring systems to support learners in pursuing their individual reading task goals.

1 Introduction

Learning vocabulary is a major component of foreign language learning. In the school context, initially, vocabulary learning is typically organized around the words introduced by the textbook. In addition to the incrementally growing vocabulary lists, some textbooks also provide thematically organized word banks. When other texts are read, the publisher or the teacher often provides annotations for new vocabulary items that appear in the text. A range of tools has been developed to support vocabulary learning, from digital versions of file cards to digital text editions offering annotations.

While such applications serve the needs of the formal learning setting in the initial foreign language learning phase, where the texts that are read are primarily chosen to systematically introduce the language, later the selection of texts to be read can in principle follow the individual interests of the student or adult, which boosts the motivation to engage with the book. Linking language learning to a functional goal that someone actually wants to achieve using language is in line with the idea of Task-Based Language Teaching (TBLT), a prominent strand in language teaching (Ellis, 2009).

Naturally, not all authentic texts are accessible to every learner, but linguistically-aware search engines, such as FLAIR (Chinkina and Meurers, 2016), make it possible to identify authentic texts that are at the right reading level and are rich in the language constructions next on the curriculum. Where the unknown vocabulary that the reader encounters in such a setting goes beyond around 2% of unknown words in a text that can be present without substantial loss of comprehension (Schmitt et al., 2011), many digital reading environments provide the option to look up a word in a dictionary. Yet, frequently looking up words in such a context is cumbersome and distracts the reader from the world of the book they are trying to engage with. Relatedly, one of the key criteria of TBLT is that learners should rely on their own resources to complete a task (Ellis, 2009). But this naturally can require pre-task activities preparing the learner to be able to successfully tackle the task (Willis and Willis, 2013). But how can a learner systematically prepare for reading a text or book they are interested in reading?

In this paper, we explore how computational linguistic methods such as distributional semantics, morphological clustering, and exercise generation can be combined with graph-based learner models to answer this question both conceptually and in practice. On the practical side, we developed an application that supports vocabulary learning as a pre-task activity for reading a self-selected book. The conceptual goal is to automatically organize the lexical-semantic space of any given English book in the form of a graph that makes it possible to sequence the vocabulary learning in a way efficiently exploring the space and to visualize this graph for the users as an open learner model (Bull

Proceedings of the 16th Workshop on Innovative Use of NLP for Building Educational Applications, pages 26–36

and Kay, 2010) showing their growing mastery of the book's lexical space. Lexical learning is fostered and monitored through automatically generated multi-gap activities (Zesch and Melamud, 2014) that support learning and revision of words in the contexts in which they occur in the book.

In section 2 we discuss how a book or other text chosen by the learner is turned in to a graph encoding the lexical space that the learner needs to engage with to read the book, and how words that are morphologically related as word families (Bauer and Nation, 1993) are automatically identified and compactly represented in the graph (2.1.1). In section 3 we then turn to the use of the graph representation of the lexical semantic space of the book to determine the reader's learning path and represent their growing lexical knowledge as spreading activation in the graph. In section 4, the conceptual ideas are realized in an application. We discuss how the new learner cold-start problem is avoided using a very quick word recognition task we implemented, before discussing the content selection and activity generation for practice and testing activities. Section 6 then provides a conceptual evaluation of the approach and compares it with related work, before concluding in section 7.

2 Constructing a structured domain model for the lexical space of a book

Going beyond the benefits of interactivity and adaptivity of individualized digital learning tools, supporting learner autonomy is known to be important for boosting motivation and self-regulation skills (Godwin-Jones, 2019). This includes the choice of reading material a learner wants to engage with, where the texts prepared by a teacher or publisher cannot reflect the interests of individual students, the topics and genres they want to explore in the foreign language. The freedom of choosing a text that the learner wants to engage with also identifying a clear functional goal for learning vocabulary – learning new words to enable us to read a text of interest, so that the interest in the content coincides with the interest in further developing the language skills. In that sense learning vocabulary becomes a pre-task activity in the spirit of task-based language learning. Organizing vocabulary learning in this way also helps turn the otherwise open-ended challenge of learning the lexical space of a new language to the clearly delineated task of mastering a sub-space. This functionally guided approach contrasts with the approach of other vocabulary learning tools selecting random infrequent lexical items from the language to be learned, which given their rare and often highly specialized nature are likely to only be useful for impressing friends when playing foreign language scrabble.

To make text-driven vocabulary learning work, we need to map the text selected by the learner into a structured domain to support systematic and efficient learning of the lexical space as used in the book. We distinguish the process of structuring the vocabulary used in the book, independent of the learner's background, from the representation of the individual learner's knowledge. The former is tackled in this section and can be regarded as our domain model, while the latter is a learner model that essentially is an overlay over the domain model, and will be discussed in section 3.

Since vocabulary learning is about establishing form-meaning connections, in principle the basic unit best suited for this would be word senses. At the same time, full automatic word sense disambiguation is complex, error prone, and often domain-specific – and in the context of a given book, a given word will often occur with the same meaning. We, therefore, limit ourselves to only disambiguating homographs in terms of their part-of-speech (POS), following Wilks and Stevenson (1998). Throughout our approach, we therefore use <word, POS> pairs as basic units. To POS annotate the book selected by the user, we use the Spacy NLP tools (http://spacy.io). Given our focus on learning the characteristic vocabulary of the book, we eliminate stop words as well as word-POS pairs appearing less than five times in the given book.

2.1 Semantic and thematic relations

To structure the lexical space in terms of meaning, there are two related options. Words can be semantically related, e.g., *tiger*, *elephant*, and *crocodile* all have the property of being wild animals; from the perspective of a WordNet, they are hyponyms of *wild animal*. On the other hand, words can also be thematically related, such as *blackboard*, *teacher*, and *chalk* all belonging to a school theme. Gholami and Khezrlou (2014) highlights the benefits of the semantic approach over the thematic approach from the perspective of a tutor. As we are building a system that acts as a tutor tracking and fostering the learner's vocabulary knowledge, we decided to focus on semantic relatedness.

2.1.1 Word families

Complementing the lexical semantic relationships, words are also related to each other through derivational and inflectional morphology. Many of these morphological processes are semantically transparent. Bauer and Nation (1993) proposed the idea of grouping words into so-called *word families* stating that "once the base word or even a derived word is known, the recognition of other members of the family requires little or no extra effort". The creation of word families is based on criteria involving frequency, regularity, productivity, and predictability of all the English affixes. Bauer and Nation (1993) arranged the inflectional affixes and common derivational affixes into the graded levels, as exemplified on the left-hand side of Figure 1.

2	develop develops developed developing	wood wood's woods wooded	bright brighter brightest
3	developable undevelopable developers(s) undeveloped	woody woodiest woodier woodiness	brightly brightish brightness
4	development(s) developmental developmentally		
5	developmentwise semideveloped antidevelopment	wooden	brighten
6	redeveloped predevelopment	anti-wooden	

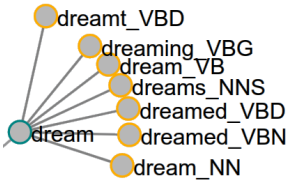

Figure 1: A word family example (Bauer and Nation, 1993) and an expanded family node in our graph

We adopt the idea of word families to compactly represent morphologically related words. The graph on the right side of Figure 1 exemplifies the word family that becomes visible when selecting the lemma *dream* in our graph representation (where word families normally are shown in collapsed form and represented by their underlying lemma). We currently put words up to level three, which generally will be transparently related, into one family – though in the future one could make this a parameter, which could also depend on the level of the learner.

2.2 Generating a lexical graph of word families and their semantic relations

To structure the lexical space of the user selected book in terms of a semantically related word graph, we start with a distributional semantic vector representation of each word, which we obtain from the pre-trained model of GloVe (Pennington et al., 2014) based on the co-occurrence statistics of the words form a large Common Crawl data-set (http://commoncrawl.org). Such word embeddings capture the distributional semantic properties of words (Goldberg and Levy, 2014).

On this basis, the relationship score between the families is computed to be the maximum pair-wise cosine similarity score of all its members. Let F_1 be a family with m members and F_2 be a family with n members. The relationship score between two families F_1 and F_2 is the maximum of cosine similarity score of all $m \times n$ pairs.

$$w_{12} = \max_{i \in F_1; j \in F_2} \frac{V_i * V_j}{\|V_i\| \|V_j\|}$$

where, w_{12} is the cosine similarity between the families F_1 and F_2

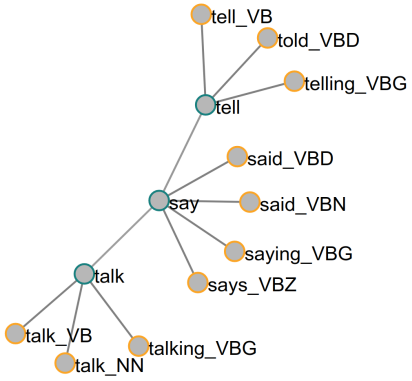

Figure 2: Formula for computing relationship between families and an example illustrating the result

The result is a network of word families, where families with members closer in the semantic vector space are connected with higher weights.

Following D'Angelo and West (1997), the number of edges in the graph can be computed as $e = \frac{n \times (n-1)}{2}$, where e is the total number of edges and n is the number of nodes (families) in the graph. The number of edges in the graph thus grows exponentially as the number of nodes increases.

When inspecting graphs derived for sample texts, we observe the majority of the connections are weak. To obtain a graph of semantic relationships that meaningfully structure the vocabulary used in

a book, we focus on the stronger relationship and eliminate edges with weights less than 0.3.

We also observed that the node families of very frequently occurring verbs tend to be very densely connected, and this impact of frequency on distributional semantic measures has been discussed in the literature (Patel et al., 1998; Weeds et al., 2004). In order to control for this kind of over-sensitivity of distributional semantic measures for highly frequent words, we restrict the node degree to a maximum threshold. Based on experiments with sample data, only the five edges with the highest weight are retained for each node.

As a result of the method described in this section, we obtain a lexical graph for the user-provided text that structures and compactly represents the lexical space of the text in a graph-based domain model. This is the lexical space that the user wants to explore and master enough to be able to read the book. In terms of computational linguistic methods, on the one hand, distributional semantics creates the overall structure of a *meaning*-connected lexical space, on the other hand, word families organize and collapse *forms* that are related by morphological processes in the linguistic system.

2.3 Example generation of graphs for books

To test the graph construction, we chose three books as a sample to study the characteristics of the vocabulary space created by our application: (a) Twenty Thousand Leagues Under the Sea by Jules Verne, (b) Harry Potter and the Sorcerer's Stone by J. K. Rowling and (c) A Game of Thrones: A Song of Ice and Fire by George R. R. Martin.

	Unique words	Learning targets	Graph nodes	Graph edges
a)	10k	1.7k	1.3k	3.3k
b)	6.5k	1.2k	1k	2.4k
c)	14k	3.7k	2.5k	5.6k

Table 1: Example graphs derived for three books

Table 1 shows the size of the text and the graph created for each book. Selecting the <word, POS> pairs occurring at least five times in the entire text, we find that 15–25% of the words from the text qualifies as lexical learning targets. These targets are grouped into families as discussed in 2.1.1, with each family being represented as a node in the graph. The resulting set of graph nodes representing word families is 20-30% smaller than the initial set of learning targets. The families then are linked as explained in section 2.2. The average number of links a family has with other families is around 2.5.

Some example word family clusters formed for these books at a threshold of similarity scores greater than 0.7 are shown in Figure 3. Only the root nodes of each family are shown. The examples illustrate that the semantically close families form meaningfully interpretable clusters.

3 Representing the lexical knowledge of a learner: an open learner model

With a structured domain model established for the vocabulary space to be explored by the user, we want to make use of it to efficiently guide the learner to cover the space and track learning in a learner model. The learner model is an overlay on the domain model that helps us track the learner's vocabulary knowledge in terms of a mastery score associated with each word family. On the basis of the learner model, we then can propose the next set of words to be practiced in a way that reduces the number of interactions required to cover the vocabulary space. It also serves as an open learner model (Bull and Kay, 2010) by allowing the user to view and explore the lexical space of the book as a graph, with each node being colored according to the current mastery score. In this section, we discuss how this is achieved.

3.1 Central node selection for efficient exploration of the vocabulary graph

Identifying the nodes that are more central than others is one of the vital tasks in network analysis (Freeman, 1978; Bonacich, 1987; Borgatti, 2005; Borgatti and Everett, 2006). Freeman (1978) formulated three major centrality measures for a node in a network: (a) *degree centrality:* a measure of strength of ties of each nodes in the network, (b) *closeness centrality:* a measure of closeness of a node to all other nodes in the network, and (c) *betweenness centrality*: a measure of the number of elements of a set S, the set of shortest paths of other node pairs in the network passing through a node.

The degree centrality measure is a greedy approach looking only at the immediate neighbours to decide the central node, whereas the closeness centrality measure accounts for the bigger picture of the entire network. So closeness centrality seems best suited for our goal of efficient coverage of the network, in our case: the graph representing the vocabulary of the given book. While the basic

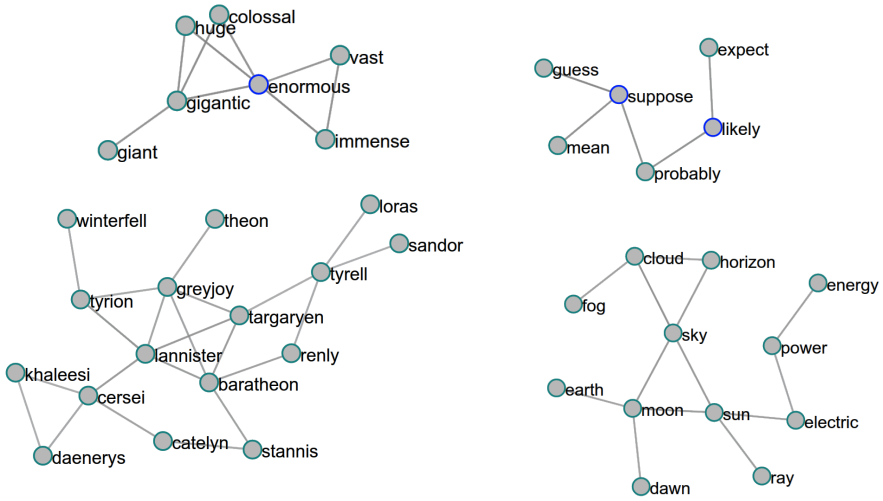

Figure 3: Example family clusters from the graphs resulting for the sample books

closeness centrality notion is only defined for fully connected networks, Wasserman et al. (1994) successfully extends it to apply to any graph. Based on this metric, we choose the top 20 (word family) nodes for a learning session and chose a word from each of those 20 families.[1]

Selecting the next words to be learned based on closeness centrality brings up the problem that neighbors that are tightly bound to the central node are likely to have a similar closeness centrality score. So when selecting the words to be practiced only based on closeness centrality, we would risk practicing closely related lexical items rather than systematically introducing the learner to the broader lexical space. In order to avoid this issue, we exclude the immediate neighbours of a word that was selected from that learning session. This supports a more distributed selection of words.

3.2 Mastery scores and updating them in the graph to capture learning

Each node in the graph is associated with a mastery score ranging from 0 to 1, with 1 indicating that the learner masters the word. We initialize the master score of each node with 0.5 and interpret this as a middle ground, where the model is uncertain about the learner's knowledge about that word.

The mastery score is updated based on the learner responses in the learning activities. To address the bottleneck that the system is tied to such a thin stream of evidence about the learner's lexical knowledge, we make use of the fact that the learner model is based on a network of semantically related word families. We use this to spread some activation from a word for which the learner has shown mastery to semantically closely related words to indicate that this word is more likely to also be known.

Let r be the learner response for a learning activity involving a word from the family F_i. Then the update to its mastery m_i is updated using $\Delta m_i = m_i * \alpha * r$. The update to the mastery score of its immediate neighbours is weighted based on the similarity score between the families $\Delta m_j = m_j * (\beta * r * w_{ij})$ where m_j is the mastery score of F_j, a neighbouring family of F_i attached with a edge weight of w_{ij}. α and β are tune-able parameters for the magnitude of an update.[2] $r \in \{-1, +1\}$ indicate the polarity of the learner's response, $+1$ for the learner responding correctly and -1 an incorrect response.

Figure 4 provides a close-up view of the graph with enlarged nodes highlighting the nodes selected for a learning activities. The figure also illustrates the color representation of the mastery level and the spreading activation to neighboring nodes. Initially, all nodes are grey, corresponding to a mastery level of 0.5. The closer the level gets to 1, the greener the node appears, and the closer to 0, the redder. A node the user has practiced with mixed success can result in a 0.5 level again, which then is shown in yellow to distinguish nodes that have already been practiced from the untouched grey ones.

[1] Currently, the word is randomly chosen from the words in the word family. One could consider selecting forms of particular relevance (e.g., irregular ones) or taking language use characteristics into account.

[2] We set both α and β to 0.3, which requires the least connected nodes to receive a minimum of two positive responses for them to count as mastered.

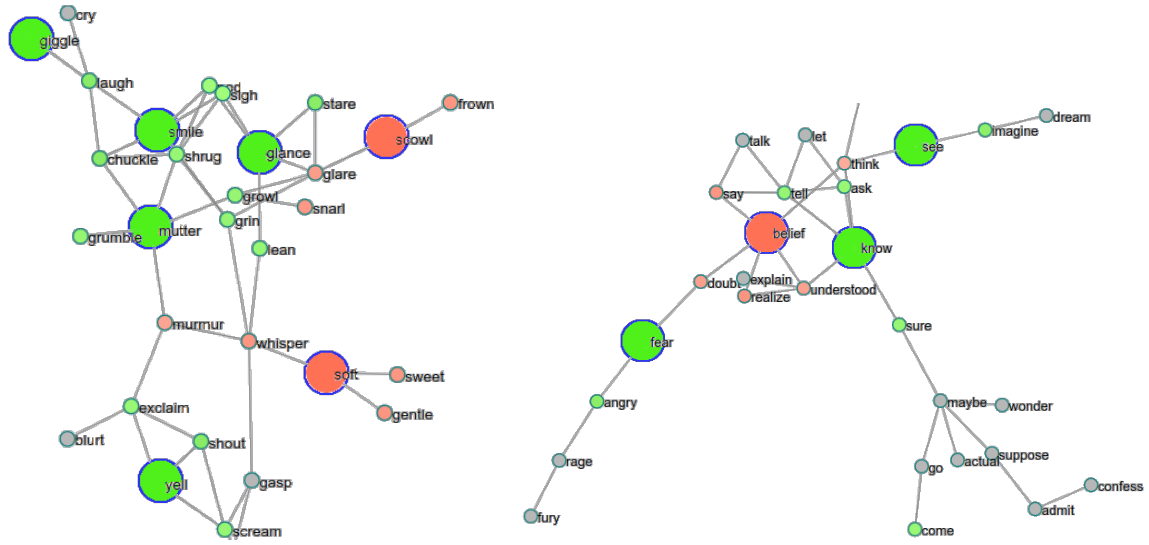

Figure 4: A close-up view of a learner model showing nodes selected for practice

While the colored visualization of the lexical space serves as an open learner model allowing learners to inspect the current state of their knowledge in relation to the lexical demands of the book, the mastery level also plays a role in selecting the next words to be practiced, with words over 0.8 no longer being selected. In the future, we plan to add a component that takes into account memory decay and the so-called spacing effect (Sense et al., 2016) to optimize when a word is selected again.

4 Putting it all together in an application

4.1 A warm start for the learner model

Given that we are targeting learners beyond the beginner stage, it is important to determine their vocabulary knowledge to avoid a cold start of the learner model. Starting from a blank slate would require many interactions with the system until the learner model reflects the learner's lexical competence – a time during which the system cannot optimally select the words to be learned next.

To avoid this cold start problem, we implemented a short web-based vocabulary Yes/No test, a format long used for vocabulary estimation (Sims, 1929; Tilley, 1936; Goulden et al., 1990). The participants see a checklist of words and select whether they know the word or not. While there is a rich literature on the test and adjustments have been proposed to counter its weaknesses as a competence diagnostic (Meara and Buxton, 1987; Beeckmans et al., 2001; Huibregtse et al., 2002; Mochida and Harrington, 2006), it readily enables learners to quickly initialize their learner model.

The words included in the test are selected from the graph using the same central node selection approach introduced in section 3.1, and the mechanism spreading activation to related nodes discussed in section 3.2 allows the system to make additional use of the information from the test.

4.2 Activity generation: practicing and testing in the target context

Content creation for vocabulary learning activities typically is a task requiring human effort so that adapting the material to individual learners' language competence and interests is beyond the reach of traditional methods. To overcome this limitation, we generate activities based on the target text for which the user wants to acquire the vocabulary. While there are multiple activity types one could consider, we implemented multi-gap activities (Zesch and Melamud, 2014), given that they make it possible for the learner to engage with a word using several sentence contexts drawn from the targeted book. Given the frequency threshold used in constructing the domain model graph, there are at least five sentences for each word in the book. We rank sentences to determine which sentences are best out of context. Fortunately, this issue has been addressed in lexicography, where authentic sentences are used in dictionaries to illustrate word usage. Kilgarriff et al. (2008) developed GDEX, a method to identify sentences that are well-suited to illustrate word meaning within a single sentence context. GDEX considers factors such as the sentence length, use of rare words and anaphora, target word occurrence in main clauses, sentence com-

pleteness, and target word collocations towards the end of sentences. Sentence length and rare word usage are the highly weighted features. We adapted GDEX for our purpose of ranking sentences for vocabulary activities and customized the rare word feature to reflect the individual learner's vocabulary knowledge as recorded in the learner model.

Learning and testing in the system are conducted in sessions. Each session consists of the top 20 central nodes from the learner model that are below the mastery score threshold. Multi-gap activities consisting of three to four sentences in which the target word chosen from the central node word family occurs are used for both learning and testing. The sentences are initially shown with the occurrences replaced by a blank. For each activity, four lexical options are provided: the target word and three distractors chosen from the book, as discussed below. Figure 5 shows an activity targeting the word family *scowl* in a learning session for the book "A Game of Thrones: A Song of Ice and Fire", after the correct word was selected by the learner.

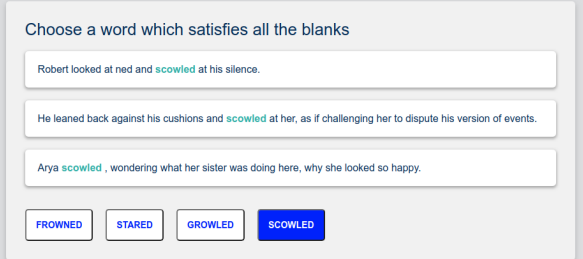

Figure 5: An example activity

In the learning mode, the learners are provided with learning aid such as dictionary lookup, translations, and word usage examples from within and outside the targeted book. The mastery scores in the learner model are not updated during training mode. In the testing mode, no such support is provided and the score for the target family and its neighbors is updated based on the user responses.

Distractor generation is a critical part of multi-gap learning activities. We are interested in distractors that require some cognitive effort to discriminate, actively engaging the learner with the choices in the different sentence contexts. Thus in the step of choosing distractors, morpho-syntactically appropriate forms are used to focus the choice on the meaning rather than grammatical surface cues. To identify challenging distractors, we select the appropriate forms from neighboring graph nodes. We empirically established that edge weights between 0.5 to 0.8 seems to be a suitably challenging distractor. This avoids synonyms that are too closely related to be distinguishable from the target, but also semantically unrelated words that are too easy to rule out. The edge weight for nodes that are not immediate neighbors is computed as the product of the edge weights connecting the nodes. Often the best distractors turn out to be two hops away. We are considering combining such a distractor generation based on the domain model with other strategies discussed by Zesch and Melamud (2014).

5 Evaluation

To empirically evaluate the approach, we first want to establish that it works as described. There are a number of components and parameters involved in generating the graph, selecting the nodes to be practiced, and updating the learner model. So we ran experiments with simulated users who performed the activities at different levels of accuracy. In addition to a a cold-start setting, where the system initially knows nothing about the learner, we also performed warm-start simulations for users at different proficiency levels. As a second step, we envisage conducting studies with language learners in authentic learning contexts. Testing educational tools in real-life contexts is crucial for establishing that an approach is effective in the complex authentic education contexts with the rich set of cognitive, motivational, and social variables at stake there. While in Meurers et al. (2019) we illustrate the feasibility of conducting such randomized controlled field studies, this clearly is an endeavor of its own, beyond the scope of this paper.

In the first set of experiments, we cold start the system with the learner model set to the default .5 chance level for every word family, and we simulated learners with performance levels of 60%, 70%, 80%, 90%, and 100%. As baseline approach for comparison, we include a traditional flashcard setup tackling words in a linear fashion, one by one, where each word is independent of the other words. Throughout, we assume that mastery is achieved when a word has reached .8 or more.

Table 2 shows the number of learning sessions, each consisting of 20 words, that the user would need to complete to fully master all learning targets in the given books in our and the baseline setup. We see that under the 100% accuracy condition, where the learner successfully completes each activity they work on, the baseline approach requires

	Learning targets	Setup	# of sessions given accuracy rate of 100%	90%	80%	70%	60%
a)	1.7k	our	85	100	120	195	1280
		baseline	170	235	340	650	3150
b)	1.2k	our	65	75	90	140	735
		baseline	120	165	245	475	2490
c)	3.7k	our	165	190	240	395	2850
		baseline	370	505	745	1390	7230

Table 2: Number of interactions required to master the vocabulary for simulated learners at given accuracy rate

exactly twice the number of learning targets when compared to our graph based approach spreading the activation to semantically related words and taking semantically transparent word families into account. The difference becomes even more pronounced when the accuracy for completing the activities is set to more realistic levels between 60 and 90%. Note the steep increase in the number of sessions needed by learners performing exercises with only 60% accuracy. This showcases that the ability to interpret lexical material in context, based on an understanding of the domain of the book from which the exercises are drawn, is important for determining which book one can successfully prepare for. Overall, while the simulation experiments clearly are based on a very simple model of learning, the observations reported should carry over to more sophisticated learning models in which initial learning gains are higher than later ones and also modeling forgetting of what has been learned.

In the second set of experiments, we assume an accuracy of 90% and instead consider the effect of proficiency differences as indicated by the learner's CEFR level. Instead of simulating the web-based book-specific vocabulary test we implemented as discussed in section 4.1, we base are simulation experiments on Meara and Milton's estimation of the knowledge of the most frequent 5000 lemmatized English words for learners at different CEFR levels as reported in Milton and Alexiou (2009). Simplifying their estimates for the number of known words distributed over the frequency bands to the upper bound given for the number of words learned in the first four proficiency levels (A1: 1500, A2: 2500, B1: 3250, B2: 3750), we started the simulation by setting the mastery score of those nodes to 0.75 for which the head word of the family occurs frequently enough (as determined by reference to SUBTLEX-US; Brysbaert and New, 2009) to be included in the set for the given proficiency level. The learner model thus encodes that the learner is likely to know the word, but the positive bias alone is not sufficient to cross the 0.8 level indicating mastery so that additional evidence is required to mark them as known. Table 3 sums up the results of the second set of experiments.

	Learning targets	Setup	# of sessions at given proficiency A1	A2	B1	B2
a)	1.7k	our	80	75	69	68
		baseline	175	147	139	123
b)	1.2k	our	54	47	43	42
		baseline	92	77	72	70
c)	3.7k	our	169	157	149	142
		baseline	421	380	376	357

Table 3: Number of interactions required to master the vocabulary for simulated learners with 90% accuracy when starting out at the specified proficiency level

For example, an A2 learner only needs 47 sessions to master the learning targets for book b) assuming an accuracy of 90% in completing the exercises, whereas in the cold start condition we saw in Table 2, one would need 75 sessions. The number of sessions estimated for this warm-start condition seems realistic for using the approach in practice, especially considering that one naturally does not need to learn all of the words to be able to read a book (Schmitt et al., 2011).

6 Related work

While the experimental evaluation provides some insights into the practical viability of the approach, given the conceptual nature of our proposal, we here also contextualize and compare the approach with related work to discuss where it conceptually advances the state of the art. Our approach can be characterized by the following aspects: First, the user can select what they want to learn the vocabulary for; they pick the text of the book they want to be able to read, i.e., the functional task goal. Second, the system automatically creates a domain model graph representing the lexical semantic space to be learned. Third, a learner model is created as an overlay of the domain model graph and records the mastery of the concepts by the learner, with updates to the learner model spreading activation through the graph to indirectly activate related concepts as a way to avoid explicit interaction for every word. Fourth, it determines in which order the words can be learned in such a way that the lexical space is efficiently explored, prioritizing the words that are central nodes. Fifth, the system compactly represents word families to allow the

visualization and open learner model to be concise and usable with minimal number of interactions. Sixth, the system supports learning of the words using multi-gap activities using sentences drawn from the actual book to be read.

Putting this approach into context of the related work on vocabulary learning, there is a large number of applications designed to support vocabulary learning – though, as we will see, the above characteristics clearly seem to set our approach apart from what is offered in this domain.

Foreign language textbooks systematically provide a list of vocabulary items per chapter and there are many specialized or general file card applications for memorizing these sentences including *Phase-6.de*, *Quizlet.com*, or *Ankiweb.net*. Other tools offer more language-related functionality.

Lextutor (https://lextutor.ca) is a website offering a collection of tools to learn vocabulary using lexical resources such as frequency-based vocabulary lists and corpus data. *List learn* supports learners in choosing words from frequency-based word lists and working with corpus concordances. *Grouplex* lets the learner select from a 2k crowd-sourced word list and practice them in fill-in-the-blank activities, with hints based on dictionary definition and POS tags. *Flash* employs cards showing words on one side and lexical support on the other. Apart from word meaning and usage, *MorphoLex* supports learning regular inflectional and derivational affixes based on the word family levels of Bauer and Nation (1993). Other lextutor tools target reading texts with support from concordances and dictionaries. *Resource assisted reading* lets the user choose a pre-processed book, but *Hyper text* allows the learner to upload their text. While Lextutor offers a variety of tools and corpus resources, none of them offer personalized learning, performance tracking, or structured vocabulary spaces.

Memrise.com is a commercial flashcard-based vocabulary learning application focused on beginners, with learning units grouped by theme with little freedom for the learner to choose contents of interest. *Duolingo.com* is a strictly guided application supporting the users to learn a foreign language using various learning activities offering some gamification elements but no personalized vocabulary learning for texts or domains of personal interest. *Vocabulary.com* is a gamified free vocabulary list learning application that lets learners choose from collections and the literature to practice the words in multiple-choice questions activities to choose the correct meaning phrase for the given word usage. The literature only is a source of vocabulary though, it is not used as testing context or learning goal, and the vocabulary domain is not semantically structured or to construct a structured learner model. *Cabuu.app* supports learning of vocabulary lists scanned from books by associating each item with gestures.

Overall, while there is a rich landscape of applications supporting vocabulary learning, the six characteristics of the method presented in this paper set our approach apart – especially the use of distributional semantic methods to create a graph representation for any book or text the user wants to read, to efficiently organize and individually support and track the learning in this lexical space.

7 Conclusion

In this paper, we discussed the methodological basis and realization of a tool allowing the learner to systematically learn the lexical material needed to be able to read a book they are interested in. Automatically structuring the lexical space and sequencing the learning is achieved through distributional semantic methods, the automatic identification of word families, and concepts from network analysis. The graph-based domain model that is automatically derived from the given book serves as the foundation of a learner model supporting the selection of an efficient learning path through the lexical space to be acquired. Multi-gap activities are automatically generated from the targeted book and used for practice and testing activities.

In addition to self-guided learning for people interested in reading specific books, which may be particularly useful in the context of so-called intensive reading programs, the approach is particularly well-suited for the English for Specific Purposes context, where both the language and the particular content domain are of direct importance. Given this kind of integration of language and content learning, a similar affinity exists to so-called Content and Language Integrated Learning (Coyle et al., 2010).

Acknowledgements

We would like to thank Himanshu Bansal for his contribution to the initial stages of this project, and we are grateful to the reviewers for the helpful suggestions and pointers they provided.

References

Laurie Bauer and Paul Nation. 1993. Word families. *International Journal of Lexicography*, 6(4):253–279.

Renaud Beeckmans, June Eyckmans, Vera Janssens, Michel Dufranne, and Hans Van de Velde. 2001. Examining the yes/no vocabulary test: Some methodological issues in theory and practice. *Language Testing*, 18(3):235–274.

Phillip Bonacich. 1987. Power and centrality: A family of measures. *American Journal of Sociology*, 92(5):1170–1182.

Stephen P. Borgatti. 2005. Centrality and network flow. *Social Networks*, 27(1):55–71.

Stephen P. Borgatti and Martin G. Everett. 2006. A graph-theoretic perspective on centrality. *Social Networks*, 28(4):466–484.

Marc Brysbaert and Boris New. 2009. 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. *Behavior Research Methods*, 41(4):977–990.

Susan Bull and Judy Kay. 2010. Open learner models. In R. Nkambou, J. Bourdeau, and R. Mizoguchi, editors, *Advances in intelligent tutoring systems*, pages 301–322. Springer.

Maria Chinkina and Detmar Meurers. 2016. Linguistically-aware information retrieval: Providing input enrichment for second language learners. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 188–198, San Diego, CA. ACL.

Do Coyle, Philip Hood, and David Marsh. 2010. *Content and language integrated learning*. Ernst Klett Sprachen.

John P. D'Angelo and Douglas B. West. 1997. Mathematical thinking. *Problem Solving and Proofs*.

Rod Ellis. 2009. Task-based language teaching: Sorting out the misunderstandings. *International Journal of Applied Linguistics*, 19(3):221–246.

Linton C. Freeman. 1978. Centrality in social networks conceptual clarification. *Social Networks*, 1(3):215–239.

Javad Gholami and Sima Khezrlou. 2014. Semantic and thematic list learning of second language vocabulary. *CATESOL Journal*, 25(1):151–162.

Robert Godwin-Jones. 2019. Riding the digital wilds: Learner autonomy and informal language learning. *Language Learning & Technology*, 23(1):8–25.

Yoav Goldberg and Omer Levy. 2014. word2vec explained: Deriving Mikolov et al.'s negative-sampling word-embedding method. *arXiv preprint arXiv:1402.3722*.

Robin Goulden, Paul Nation, and John Read. 1990. How large can a receptive vocabulary be? *Applied Linguistics*, 11(4):341–363.

Ineke Huibregtse, Wilfried Admiraal, and Paul Meara. 2002. Scores on a yes-no vocabulary test: Correction for guessing and response style. *Language Testing*, 19(3):227–245.

Adam Kilgarriff, Milos Husák, Katy McAdam, Michael Rundell, and Pavel Rychlỳ. 2008. Gdex: Automatically finding good dictionary examples in a corpus. In *Proceedings of the XIII EURALEX international congress*, pages 425–432. Universitat Pompeu Fabra Barcelona, Spain.

Paul Meara and Barbara Buxton. 1987. An alternative to multiple choice vocabulary tests. *Language Testing*, 4(2):142–154.

Detmar Meurers, Kordula De Kuthy, Florian Nuxoll, Björn Rudzewitz, and Ramon Ziai. 2019. Scaling up intervention studies to investigate real-life foreign language learning in school. *Annual Review of Applied Linguistics*, 39:161–188.

James Milton and Thomaï Alexiou. 2009. Vocabulary size and the Common European Framework of Reference for languages. In *Vocabulary studies in first and second language acquisition*, pages 194–211. Springer.

Kira Mochida and Michael Harrington. 2006. The yes/no test as a measure of receptive vocabulary knowledge. *Language Testing*, 23(1):73–98.

Malti Patel, John A Bullinaria, and Joseph P Levy. 1998. Extracting semantic representations from large text corpora. In *4th Neural Computation and Psychology Workshop, London, 9–11 April 1997*, pages 199–212. Springer.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543.

Norbert Schmitt, Xiangying Jiang, and William Grabe. 2011. The percentage of words known in a text and reading comprehension. *The Modern Language Journal*, 95(1):26–43.

Florian Sense, Friederike Behrens, Rob R. Meijer, and Hedderik van Rijn. 2016. An individual's rate of forgetting is stable over time but differs across materials. *Topics in Cognitive Science*, 8(1):305–321.

Verner Martin Sims. 1929. The reliability and validity of four types of vocabulary tests. *The Journal of Educational Research*, 20(2):91–96.

Harvey C. Tilley. 1936. A technique for determining the relative difficulty of word meanings among elementary school children. *The Journal of Experimental Education*, 5(1):61–64.

Stanley Wasserman, Katherine Faust, et al. 1994. *Social network analysis: Methods and applications*. Cambridge University Press.

Julie Weeds, David Weir, and Diana McCarthy. 2004. Characterising measures of lexical distributional similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING)*, pages 1015–1021.

Yorick Wilks and Mark Stevenson. 1998. The grammar of sense: Using part-of-speech tags as a first step in semantic disambiguation. *Natural Language Engineering*, 4(2):135–143.

Jane Willis and David Willis. 2013. *Doing task-based teaching*. Oxford University Press.

Torsten Zesch and Oren Melamud. 2014. Automatic generation of challenging distractors using context-sensitive inference rules. In *Proceedings of the Ninth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pages 143–148, Baltimore, Maryland. Association for Computational Linguistics.