Author(s): Lahti, Lauri

Title: Educational framework based on cumulative vocabularies, conceptual networks and Wikipedia linkage

Year: 2013

Version: Post print

Please cite the original version:

Lahti, Lauri. 2013. Educational framework based on cumulative vocabularies, conceptual networks and Wikipedia linkage. London International Conference on Education 2013 (LICE 2013), London, UK, 4-6 November 2013. P. 470-478. ISBN 978-1-908320-16-2 (electronic). http://infonomics-society.org/

Lahti, L. (2013). Educational framework based on cumulative vocabularies, conceptual networks and Wikipedia linkage. Proc. London International Conference on Education 2013 (LICE 2013). 4-6 November 2013, London, UK (eds. Shoniregun, C., & Akmayeva, G.), 470–478. ISBN 978-1-908320-16-2.

Educational framework based on cumulative vocabularies, conceptual networks and Wikipedia linkage

Lauri Lahti
Department of Computer Science and Engineering
Aalto University School of Science, Finland

Abstract

We propose a new educational framework based on guided exploration in small-world networks relying on hyperlink network of the Wikipedia online encyclopedia (http://www.wikipedia.org) in which hyperlinks between articles define conceptual relationships. Educational material is presented to student with cumulative conceptual networks based on hyperlink network of the Wikipedia connecting concepts of vocabulary about current learning topic. Personalization of educational material is carried out by alternating the distribution of enabled hyperlinks connecting concepts belonging to current vocabulary according to requirements of learning objective, learning context and learner's knowledge. Besides developing a computational method to manage educational material with conceptual networks and to explore the shortest paths between concepts of vocabulary (especially highest-ranking hyperlinked concepts and strongly rising hyperlinked concepts), we have also experimentally estimated properties of conceptual networks generated based on hyperlink network of the Wikipedia between concepts retrieved from English Vocabulary Profile for cumulatively growing vocabularies corresponding to six language ability levels.

1. Introduction

Previous research has shown that small-world networks offer efficient compact link structures that seem to exist in many natural processes. Using small-world networks can help to minimize paths required to form connectivity between nodes of the network and to maintain this property also when the network grows or experiences other modifications. Small-world topology has been indentified structurally and functionally in human brain networks [1] and thus we think that representation of knowledge in form of small-world networks should be encouraged to support various knowledge management tasks and especially learning. Currently one of the biggest freely accessible knowledge resources is collaboratively built Wikipedia online encyclopedia and that has been shown to naturally represent properties of a small-world network [2]. Motivated by previous research we now propose a new framework to support learning based on knowledge structures inspired by the hyperlink network of the Wikipedia and we supply this proposal with some promising experimental results relying on our empirical analysis of properties of conceptual networks that we have generated based on the Wikipedia.

2. Previous work

It has been estimated that a human vocabulary is about 4000-5000 word families for a native 5-year-old child and grows by about 1000 word families every year to reach about vocabulary of 20000 word families for a native university graduate [3]. There are possibly well over 54000 word families in English [3] but a non-native highly educated adult can manage sufficiently already with 8000-9000 word families [4]. Understanding 95 percent of general texts has been considered sufficient for reasonable comprehension [5] corresponding to a vocabulary of 3000-5000 or just 2000-3000 word families [3].

Common European Framework of Reference for Languages (CEFRL) offers guidelines about how to measure language ability with six progressive levels that have been supplied with illustrative descriptors created and scaled with Rash modelling based on Swiss surveys in 1994-1995 covering 300 teachers and 2800 learners ([6]; [7]). These six levels of language ability in increasing order of expertise have been labeled with names A1 (Breakthrough), A2 (Waystage), B1 (Threshold), B2 (Vantage), C1 (Effective operational proficiency) and C2 (Mastery). Estimates about how many guided hours of learning are required to reach the language ability levels A2-C2 of CEFRL include 180-200 hours for A2, 350-400 hours for B1, 500-600 hours for B2, 700-800 hours for C1 and 1000-1200 hours for C2 ([8]; [9]). On the other hand, time required to adopt professional proficiency in a foreign language by a native English speaker has been estimated to range from 23-24 weeks or 575-600 class hours (language closely related to English, for example French) to 88 weeks or 2200 class hours (language which is exceptionally difficult, for example Arabic) [10].

It is possible to estimate at least coarsely the amount of information processing in learning by using measures that have been indentified for word-based information input

and output for infants, children and adults. Psychological and neurobiological experiments have given motivation to suggest that learning benefits from specific patterns of spacing of exposure and retention. According to Thalheimer [11] three or more repetitions are needed to ensure learning and spacing of exposures and spacing of retentions should be about equal and match the time required for remembering. According to Kandel [12] to activate genes establishing long-term memory stimulation of synapses can be triggered by 4-5 spaced puffs of serotonin, and according to Fields [13] at least three action potentials separated with at least 10 minutes can activate a gene for long-term memory formation in a synapse leading it to produce required proteins for about 30 minutes. By exposing marine snail to four brief trains for four days could generate memories that lasted weeks [12].

Children in ages of 2-30 months hear 12815 words per day from all adults, and there are 520 conversational turns per day for 24-month-old child in a typical family [14]. Number of daily vocalizations is for 12-month-olds 1000-1500 vocalizations, for 24-month-olds 1300-2200 vocalizations, for 36-month-olds 1600-2700 vocalizatons, and for 48-month-olds 1700-2700 vocalizations [14]. When considering 17-29-year-olds, men speak 15669 words per day and women 16215 words per day [15]. Weekly time usage pattern for children living in a family having computer with internet show that 6-8-year-olds spent about 33 h 55 min in school, 2 h 26 min for additional studying, and furthermore 1 h 28 min in reading, 12 min in being read to and 1 h 8 min in computer activities, where as 15-17-year-olds spent 30 h 21 min in school, 5 h 20 min for additional studying, and furthermore 58 min in reading, 0 min in being read to and 4 h 5 min in computer activities [16].

A student with average score in reading test reads 12,9 minutes per day or 601000 words per year [19] whereas a student with excellent score reads 90,7 minutes per day or 4733000 words per year [19]. Reading speed for population in general is about 200 words per minute [20] and for sufficient comprehension at least about 200 words per minute is suggested [21], and with average sentence length below 20 words [22] this results in at least 10 sentences per minute. Based on previous research it has been concluded estimates that a student can yearly adopt 1000 new word families [3] or 2000-3500 new words [17] or 3000 new words [18].

Concerning amount of arriving and departing hyperlinks, in the World Wide Web, mean in-degree is 6,10 and mean out-degree is 38,11 [23], whereas in the Wikipedia mean in-degree is 20,63 and mean out-degree is 20,63, and median in-degree is 4 and median out-degree is 12 [24]. Relation between number of directed links L and articles N in the Wikipedia has been suggested to be approximately $L=N^{1,4}$ [25]. Rodero-Merino et al. [26] showed experimentally that coverage of a random walk in small-world network grows faster when the average degree of network is higher and also that an average search length grows linearly with the network size and the bigger the average degree the shortest the searches are. According to experiment of full hyperlink network of the Wikipedia on 3 March 2008 [27] on average 4,573 traversals of hyperlinks are needed to get from any article to any other article. Correspondingly in mailing experiment in USA with 296 persons the number of relationship steps connecting two persons was in range 4,6-6,1 [28], and in Facebook social network with 721 million users and 68,7 billion links between them average number of relationship steps was 4,74 [29].

3. Method

Content of the Wikipedia online encyclopedia can be quite revolutionarily edited by anyone and despite some skepticism vandalism has not prevented building relatively extensive and educationally reliable knowledge content often supplied with some convincing references and importantly all edits are saved to a log thus enabling a convenient access and possibility to revert to any previous version of an article ([30]; [31]). There is a need to develop educational resources that can be flexibly updated, shared and personalized and we think that the Wikipedia and its related sibling projects can inherently offer a promising resources for trying to find most optimal processes for building and exploring knowledge structures in learning. Our work focuses especially on the biggest language edition English Wikipedia (http://en.wikipedia.org) containing over 4,3 million articles (in September 2013).

Each encyclopedic article of the Wikipedia can be considered to represent a concept that is the title entry of the article. Text of an article is typically supplied with hyperlinks to related other articles and thus hyperlink network of the Wikipedia can be considered to form a conceptual network. Besides establishing a connectivity from start concept to end concept of hyperlink, each hyperlink is typically surrounded by a phrase in the article text of start concept that offers a brief written definition about the relationship between the connected concepts.

We think that already at the moment the Wikipedia basically contains so much useful knowledge that it could possibly cover a majority of all those situations dealing with a need of factual knowledge that a student can encounter during all his school years. However, this useful knowledge is not possibly organized and presented currently in the most optimal form to support independent cumulative adoption of knowledge that addresses student's previous knowledge and personal needs as well as to help identifying the most essential content for current learning topic and to encourage inductive and deductive reasoning with sufficiently spaced and repeated exposure and retention.

Therefore we think that there is a great potential for education in the knowledge contained already now in the Wikipedia but to enable better learning opportunities relying on the Wikipedia the research community should invest on more analysis about the properties of the Wikipedia and to develop computational methods that let to transform its knowledge to various forms of

representation to address personalized educational needs of a student.

Motivated by our earlier work [32] and previous research that has identified small-world topology structurally and functionally in human brain networks [1] as well as properties of small-world network in the Wikipedia [2] we propose a method for cumulative adoption of vocabulary supported by representations of vocabulary in knowledge structures that are based on a small-world network. We think that due to properties of small-world network emerging inherently in various instances of nature, it is possible that learning of new knowledge can get useful support if new pieces of knowledge can be added to learner's previous knowledge entities in mind in a process that can be represented by building a small-world network and through its modification and exploration.

We think that instead of just one small-world network there can be a great number of diverse parallel and partially overlapping and multidimensional small-world networks that can be used at the same time to represent knowledge both in educational material, such as texts, and in the learner's mind. We think that among students there are large individual differences in student's mental small-world networks representing his previous knowledge entity. Therefore to make new pieces of knowledge to become sufficiently fit into previous knowledge entity of student during learning process it is useful to offer personalized forms of representation of educational material.

With our method educational material is presented to student with cumulative conceptual networks based on hyperlink network of Wikipedia connecting concepts of vocabulary about current learning topic. Personalization of educational material is carried out by alternating the distribution of enabled hyperlinks connecting concepts belonging to current vocabulary according to requirements of learning objective, learning context and learner's knowledge. So far our method accepts only nouns to vocabularies since hyperlinks in the Wikipedia are typically defined to connect nouns but also other part-of-speech could be possibly exploited with a resembling approach.

Thus for life-long-learning an ultimate aim can be to reach a maximal coverage of the conceptual small-world networks representing all human knowledge and besides that even some personal contribution could be done to supplement this heritage of human knowledge through own writings and other forms of conveying new knowledge to the community. On the other hand, we think that all knowledge entities can been seen to consist of a complex collection of interconnected, overlapping and nested small-world networks so that each separate new learning topic can be considered to be learned as an own specific small-world network that becomes gradually more and more connected also to other small-world network structures held already so far in the mind of student.

When creating a hyperlink network of vocabulary based on hyperlink network of the Wikipedia we suggest extracting a relation statement for each hyperlink of Wikipedia from phrase surrounding hyperlink anchor of end concept in article text of start concept. For example for a hyperlink pointing from concept Music to concept Art one relation statement from article text of start concept Music is "Music is an art form whose medium is sound and silence." (here hyperlink anchor underlined). We suggest that during exploration in hyperlink network of vocabulary when student traverses a hyperlink between concepts learning of this relationship is supported by showing to the student a relation statement corresponding to this hyperlink. Eventually a learning session consists of a chain of traversed hyperlinks and their relation statements that can be guided to proceed in a sequential process having tailored variation and repetition computed based on theory of spaced learning, as discussed in our previous work [32].

To enable implementing educational technology for practical educational activities for students we have carried out empirical experiments to try to identify some constraints of conceptual small-world networks and to better understand behavior of their properties. Thus besides developing a computational method for exploiting conceptual small-world networks to manage and explore educational material we now also report some preliminary findings of experiments about the properties of conceptual small-world networks that we have generated based on hyperlink network of the Wikipedia connecting concepts of vocabulary about current learning topic.

4. Experiment

English Vocabulary Profile is a database aiming to represent all words and phrases learners know at each of six levels of Common European Framework of Reference for Languages (CEFRL) [33]. Table 1 shows properties of conceptual networks that we generated based on hyperlink network of the Wikipedia (as of June-July 2013) between concepts retrieved from English Vocabulary Profile (http://vocabulary.englishprofile.org/dictionary//word-list/uk/a1_c2/A) for cumulatively growing vocabularies corresponding to each of six language ability levels ranging from A1 to C2.

At the highest language ability level C2 we have the most extensive vocabulary that we call as vocabulary A1&A2&B1&B2&C1&C2 (i.e. including all six cumulative vocabularies of consecutive language ability levels A1, A2, B1, B2, C1 and C2 together) and we identified that it contains 15715 unique language items (words or phrases) that include 3710 unique nouns. Then we wanted to identify all possible hyperlinks that are connecting these 3710 unique nouns in hyperlink network of the Wikipedia and we found 25153 unique hyperlinks so that they actually connected 2880 unique nouns of these 3710 unique nouns. Therefore it seems that at language ability level C2 the Wikipedia can offer interconnected linkage for about 77,6 percent (2880/3710) of nouns belonging to current noun vocabulary. According to our

calculations each of these 2880 unique nouns of vocabulary A1&A2&B1&B2&C1&C2 has an average value of 8,7 departing unique hyperlinks and a median value of 5 departing unique hyperlinks and an average value of 8,7 arriving unique hyperlinks and a median value of 5 arriving unique hyperlinks linking it to other unique nouns belonging to same vocabulary A1&A2&B1&B2&C1&C2. In the entity of 25153 unique hyperlinks it appeared that for 4824 hyperlinks there was another hyperlink going also into opposite direction thus 2412 connections can be considered bidirectional.

Since applying the hyperlink network of the Wikipedia for educational activities relies on those nouns that actually happen to exist in hyperlinks, we wanted to estimate the properties of the conceptual networks we have generated in respect to size of noun vocabulary that is actually available for browsing in the Wikipedia along unique hyperlinks connecting unique nouns of vocabulary.

By comparing growth of values in columns of Table 1 along language ability levels from A1 to C2 we approximated that the number of unique nouns in vocabulary is about 1,3 times the number of unique nouns in unique Wikipedia hyperlinks connecting unique nouns in vocabulary, and the number of unique language items (words of phrases) in vocabulary is about 4,3 times the number of unique nouns in vocabulary, and the number of unique Wikipedia hyperlinks connecting unique nouns in vocabulary is about 8,8 times the number of unique nouns in unique Wikipedia hyperlinks connecting unique nouns in vocabulary. Based on these dependencies we extrapolated to Table 1 coarse predicted estimated values to represent four additional cases in which the number of unique nouns in vocabulary reaches such levels that have been suggested in previous research to correspond to reasonable 95 percent level comprehension (3000-5000 or just 2000-3000 word families ([3]; [5])), a non-native adult (8000-9000 word families [4]), native adult (20000 word families [3]) and general vocabulary (well over 54000 word families in English [3]).

Brezina and Gablasova [34] estimated that about 46 percent of 3000 highest-ranking words of British National Corpus are nouns which is a greater ratio than a ratio based on our just mentioned approximation that there are 23 percent (1/4,3) unique nouns in unique language items of a vocabulary. Anyway since Wikipedia hyperlinks connect now only nouns we assume that a student's explorations among 2880 unique nouns in 25153 unique hyperlinks connecting unique nouns of vocabulary A1&A2&B1&B2&C1&C2 can at least indirectly offer a conceptual exposure and coverage of 2,2-4,3 times greater amount of unique language items (i.e. containing also other part-of-speech than just nouns) meaning coverage of 6261-12522 unique language items. A student can gain this additional exposure for example by reading supplementing words in relation statements extracted from phrases surrounding hyperlink anchor in article text of start concept.

Therefore we suggest that hyperlink network of vocabulary A1&A2&B1&B2&C1&C2 containing 2880 unique nouns with 25153 unique interconnecting hyperlinks can be considered to offer sufficient knowledge structure to represent relatively reliably conceptualization of everyday human vocabulary corresponding to reasonable 95 percent level comprehension (3000-5000 or just 2000-3000 word families ([3]; [5])) that is defined based on cumulative iterative collaborative building process of Wikipedia online encyclopedia.

We carried out random path explorations in hyperlink network of 25153 unique hyperlinks connecting 2880 unique nouns of vocabulary A1&A2&B1&B2&C1&C2 so that any hyperlink can be traversed in both actual linking direction and opposite direction. A random path of 1000 steps visited only about 62 unique concepts (2,2 percent) of 2878 unique concepts that are hierarchically highest and 67 percent of visits stayed among only 10 highest unique concepts. Similarly with a random path of 10000 steps only about 80 unique concepts (2,8 percent) of 2878 unique concepts became visited and 68 percent of visits stayed among only 10 highest unique concepts.

These results seem to indicate that in hyperlink network of vocabulary exploration relying heavily on random choices of student without systematic guidance

Table 1 - Properties of conceptual networks generated based on hyperlink network of the Wikipedia between concepts for cumulative vocabularies of six language ability levels of English Vocabulary Profile ranging from A1 to C2.

Vocabulary of language ability level reached so far (predicted = only extrapolated estimates)*	*Unique language items (words or phrases) in vocabulary*	*Unique nouns in vocabulary*	*Unique Wikipedia hyperlinks connecting unique nouns in vocabulary*	*Unique nouns in unique Wikipedia hyperlinks connecting unique nouns in vocabulary*
A1	785	305	1007	248
A1&A2	2382	880	3868	707
A1&A2&B1	5327	1761	9566	1376
A1&A2&B1&B2	9502	2707	17448	2123
A1&A2&B1&B2&C1	11908	3 198	21410	2472
A1&A2&B1&B2&C1&C2	15715	3 710	25153	2880
3000-5000 unique nouns (reasonable 95 percent level comprehension), predicted*	12900-21500 *	3000-5000 *	20308-33846 *	2308-3846 *
8000-9000 unique nouns (non-native adult), predicted*	34400-38700 *	8000-9000 *	54154-60923 *	6154-6923 *
20000 unique nouns (native adult), predicted*	86000 *	20000 *	135385 *	15385 *
54000 unique nouns (general vocabulary), predicted*	232200 *	54000 *	365538 *	41538 *

can lead to very limited pedagogic gain due to visiting only very limited subsection of all unique concepts and their unique connecting hyperlinks. Thus we suggest that pedagogically rewarding exploration in hyperlink network of vocabulary should actively exploit traversing the shortest paths connecting pairs of unique concepts of vocabulary. We think that in adoption of new knowledge the learner benefits from an opportunity to see intuitively the shortest connectivity between pieces of knowledge thus helping contextually to filter out less relevant things that might disturb concentration by excessive cognitive load, and using the shortest paths enables also highlighting clustering structure of conceptual relationships to the student and generating a systematic efficient process to traverse in hyperlink network of vocabulary with an extensive diverse coverage.

We suggest that to support adoption of vocabulary a student's guided exploration in hyperlink network of vocabulary could proceed pedagogically rewardingly if exploration of the shortest paths gradually moves to cover new concepts related to concepts that have been adopted already earlier. On coarser level of granularity this gradual moving can be implemented by moving from vocabulary A1 to A1&A2 and then from vocabulary A1&A2 to A1&A2&B1 and so on. On finer level of granularity the guided exploration should gradually introduce new concepts belonging to current vocabulary and its most related subset of concepts concerning current learning topic while still also helping to refresh previously adopted concepts, with sequential tailored variation and repetition computed based on theory of spaced learning (see details in [32]).

We also suggest that these new concepts should particularly include highest-ranking concepts of the topics that are intended to be learned so that exploration in hyperlink network of vocabulary could be performed especially by traversing the shortest paths between the highest-ranking concepts of previously adopted concepts and highest-ranking concepts of new concepts. In addition we suggest that, when available, parallel alternative shortest paths should be traversed between pairs of concepts to learn better the diversity of conceptual relations. With these suggestion we expect to establish efficient connectivity covering old and new concepts relying on dominant concept clusters of hyperlink network shown to student and that could then be also easier to conceptualize by the student.

We suggest that according to the needs of the learner new cumulative sets of vocabularies along gradually increasing adoption of new knowledge can be gained by generating high-frequency word lists from suitable text samples concerning intended learning topic or for example retrieving a desired set of words from resources such as British National Corpus [35].

We analyzed a sample of 102 Wikipedia articles selected to match 102 highest-ranking terms in texts generated by students. These 102 articles had together hyperlinks to 20512 end concepts of which 14907 were unique and an article had on average 201 (median value 152) departing hyperlinks. When analyzing all 422 unique hyperlinks existing between these 102 Wikipedia articles (as of 3 March 2008) we found out that each start concept of a hyperlink had on average 4,1 (median value 3,5) different end concepts. Furthermore among all hyperlinks between these 102 Wikipedia articles (as of 3 March 2008) we indentified that there were on average 1,5 (median value 1) parallel hyperlinks (i.e. a certain end concept having more than one hyperlink anchors in article text of start concept) from each start concept to its end concept. For example, an article having two departing unique hyperlinks will on average have one of these two unique hyperlinks duplicated (1,5*2=3). In addition in all 422 unique hyperlinks existing between the set of 102 Wikipedia articles in the article text of start concept the end concept was mentioned on average 7,4 (median value 3) different times. On the other hand we indentified that in the article text of each 102 articles on average 21,3 (median 20) different concepts corresponding to other 101 article titles were mentioned (i.e. resembling end concept).

Thus based on this sample of 102 articles it seems that when considering a noun vocabulary interconnected by Wikipedia hyperlinks, on average Wikipedia article has 1,5 hyperlink anchors for each hyperlink and the end concept of each hyperlink occurs 7,4 times in article text of start concept. Furthermore while about 4 percent (4,1/101) of concepts belonging to vocabulary can be actually reached via hyperlink from Wikipedia article it appears that about 21 percent (21,3/101) of concepts belonging to vocabulary are anyway mentioned in article text of an average Wikipedia article, meaning that number of potential relationships becomes multiplied with about 5.

These results suggest that besides actually existing unique hyperlinks between concepts of a vocabulary and possible exploitation of parallel hyperlinks there exists a passive potential to extend current linking by establishing additional supportive cross-linking between all occurrences of concepts of vocabulary in all Wikipedia article texts of concepts of vocabulary. These findings suggest concerning vocabulary A1&A2&B1&B2&C1&C2 that hyperlink network which we so far managed to get to contain 2880 unique nouns with 25153 unique interconnecting hyperlinks can be extended progressively to contain much more hyperlinks, and using multiplication factors (1,5; 7,4 and 5) motivated above leads to an estimated range of 37730-186132 hyperlinks. By generating these supplementing hyperlinks we expect to increase diversity of linkage thus offering extended variation in exposure and coverage of a student's exploration in hyperlink network to adopt conceptual relationships and knowledge in general.

We carried out experiments to identify how the shortest paths in hyperlink network of vocabulary evolve when observed vocabulary is cumulatively expanded thus introducing new interconnecting hyperlinks and intermediary concepts that enable emergence of gradually shorter paths between pairs of concepts of vocabulary as well as increase in the number parallel alternative paths. We experimented with vocabularies ranging from

vocabulary A1 with 1007 unique interconnecting hyperlinks to vocabulary A1&A2&B1&B2&C1&C2 with 25153 unique interconnecting hyperlinks and the results seemed to support suggested pedagogic gains of using the shortest paths to guide educational exploration for adoption of new knowledge.

For example we analyzed how the available shortest paths evolve between start concept "question" and end concept "school" when expanding observed vocabulary cumulatively from A1 to A1&A2&B1&B2&C1&C2. With vocabulary A1 the shortest paths require traversing eight consecutive hyperlinks and there is only one path of this length: question -> problem -> business -> restaurant -> food -> supermarket -> book -> homework -> school. With vocabulary A1&A2 the length of the shortest path has decreased to three hyperlinks and again there is only one path of this length: question -> quiz -> game -> school. With vocabularies bigger than A1&A2 the length of the shortest path does not anymore decrease from three hyperlinks but new alternative parallel paths emerge thus introducing diversity to express the characteristics of relationship of concepts (please note that those shortest paths found with smaller vocabularies remain available also with bigger vocabularies). With vocabulary A1&A2&B1 two new alternative parallel paths emerge including question -> grammar -> education -> school and question -> information -> education -> school, and with vocabulary A1&A2&B1&B2 five new paths include question -> philosophy -> psychology -> school, question -> philosophy -> government -> school, question -> theory -> education -> school, question -> theory -> psychology -> school and question -> concept -> psychology -> school. With vocabulary A1&A2&B1&B2&C1 one new alternative parallel path emerges including question -> proposition -> psychology -> school but vocabulary A1&A2&B1&B2&C1&C2 does not introduce any more new paths (i.e. vocabulary A1&A2&B1&B2&C1&C2 offers nine parallel paths) which can possibly even indicate that already with this size of vocabulary some kind of saturation has been reached in formation of somewhat optimal connectivity between these two concepts of human knowledge in respect to shortness of paths and diversity of parallel paths.

5. Discussion and future work

We have now explained our experiments creating estimates about the sizes of hyperlink networks that can match with language ability levels from A1 to C2 of English Vocabulary Profile, and also estimates about the sizes of hyperlink networks that can match with sizes of vocabularies covering language usage needs for reasonable 95 percent level comprehension, non-native adults, native adults and general vocabulary. We have also estimated how already hyperlink network of vocabulary A1&A2&B1&B2&C1&C2 containing 2880 unique nouns with 25153 unique interconnecting hyperlinks can be extended to offer much more hyperlinks based on article texts defining unused potential relationships and possible exploitation of parallel hyperlinks. We have also experimentally identified very limited coverage gained with random paths in hyperlink network of vocabulary and thus we have suggested using the shortest paths to guide educational exploration for adoption of new knowledge, and with cumulatively growing vocabularies the length of shortest paths can usefully decrease and alternative parallel paths offering diversity can be gained. We do not know any previous research proposing same kind of approach and results that we have presented here and we hope that our suggestions can open promising new perspectives to learning. Based on our experiments we next explain some further suggestions for educational use of hyperlink network of vocabulary and we hope these ideas can offer inspiration for future work in both on research agenda and in real-life application to support personalized learning.

It is pedagogically useful that when observing the shortest paths to two opposite directions between a pair of concepts there often emerges two different routings offering new perspectives. For example with vocabulary A1&A2&B1&B2&C1&C2 from concept "love" to concept "memory" the shortest paths have two hyperlinks and there is only one path of this length: love -> psychology -> memory, and from concept "memory" to concept "love" the shortest paths have three hyperlinks and there are three alternative parallel paths of this length including memory -> psychology -> emotion -> love, memory -> psychology -> motivation -> love and memory -> learning -> emotion -> love. Besides identifying the shortest paths in both directions between a pair of concepts we suggest that additional pedagogic potential of diversity and possibly even shorter paths become available when identifying the shortest paths in hyperlink network of vocabulary also so that any hyperlink can be traversed in both actual linking direction and opposite direction. When enabling these bidirectional hyperlink traversals in hyperlink network of vocabulary A1&A2&B1&B2&C1&C2 between concepts "love" and "memory" the shortest paths have length of two hyperlinks and there are three alternative paths of this length: love -> psychology -> memory, love -> loneliness -> memory and love -> mind -> memory.

We think that pedagogically useful exploration in hyperlink network of vocabulary could benefit from exploring especially those shortest paths that exist between the highest-ranking hyperlinked concepts and strongly rising hyperlinked concepts of vocabulary, and some of them are shown in Table 2 for cumulative vocabularies of six language ability levels of English Vocabulary Profile ranging from A1 to C2.

Column High of Table 2 lists some of the highest-ranking hyperlinked concepts (occurrences indicated in parenthesis), i.e. those concepts that have the highest number of departing unique hyperlinks (in case of highest-ranking as being a start concept) or arriving unique hyperlinks (in case of highest-ranking as being end concept). Column Rising of Table 2 lists some of strongly rising hyperlinked concepts, i.e. concepts that seem to

strongly rise in ranking position from previous smaller vocabulary to current bigger vocabulary in respect to number of departing or arriving unique hyperlinks (for example which of the concepts belonging to vocabulary A1 seem to get among the biggest increase in ranking position when observing these same concepts again in vocabulary A1&A2). We created shown list of rising concepts (change in ranking position indicated in parenthesis, suffix -s indicating shared ranking position) by browsing highest-ranking concepts in decreasing order and selected such concepts which increased their ranking position by at least value 0,01 when for all vocabularies the ranges of ranking values had been first transformed to equal range of closed interval from zero to one (i.e. [0,1]).

We think that a person's ability to adopt new knowledge based on the shortest paths between concepts is affected for example by the length of the shortest paths, the number of alternative parallel shortest paths and the number of different concepts belonging to intermediary concepts along paths. We think that among parallel paths those shortest paths that have highest number of shared intermediary concepts and especially such intermediary concepts that occur most often among paths are important paths to define meaning of relationship between a pair of concepts. On the other hand to express diversity of meanings those shortest paths are important which have most distinctive routing among parallel paths (i.e. minimizing sharing). Also longer paths than the shortest paths can complement meanings of conceptual relationships.

We think that to adopt new knowledge a successful pedagogical exploration in hyperlink network of vocabulary could possibly benefit from such mental processes of student that have resemblance to traversing average search paths in network. Thus we suggest that conceptualization in the student's mind could benefit from having such guided exploration in conceptual networks that enables many explorations that do not explore directly only the shortest paths between concepts but instead extend to cover also some sidetracks and even dead-ends.

Motivated by previous research showing that in a small-world network of 10000 nodes has an average search path of 950 steps for average degree of 10 and average search path of 200 steps for average degree of 30 [26] and that Wikipedia has mean out-degree 20,63 (median value 12) [24], we thus coarsely estimate that in hyperlink network of vocabulary A1&A2&B1&B2&C1&C2 having average out-degree 8,7 (median value 5) and containing 2880 unique nouns to enable the student to at least weakly conceptualize a single relationship between a pair of concepts could possibly require exploring about 300 steps in the hyperlink network of vocabulary. Since previous research showed that in the Wikipedia on average 4,573 hyperlink steps are between a pair of concepts [27], and similarly in Facebook social network the average number of relationship steps between two users is 4,74 [29], our coarse estimate of exploring 300 steps is about 66 times the average length of the shortest path between a pair of concepts in hyperlink network of vocabulary.

It thus seems that the student's conceptualization of conceptual relationships can require many times more exploration steps in the hyperlink network than belong to exploring just the shortest paths. On the other hand, it is possible that when traversing one exploration path several concepts that become encountered along the path can be cumulatively conceptualized in parallel, and it is also possible that the number of steps needed in later explorations can decrease as some kind of memories about previous explorations help to guide later explorations.

Since earlier research estimates that children are daily exposed to hear about 12815 words [14] and produce 1000-2700 vocalizations [14], and adults speak daily about 15669-16215 words [15], it seems to us that human learning ability apparently can easily manage knowledge adoption at least through listening at a daily rate of about 12815-16215 words. Based on earlier research it seems that knowledge adoption through reading can have somewhat lower levels than listening but still managing daily rate of about 1647-12967 words [19] corresponding with an average length of 20 words in sentence [22] to reading 80-648 sentences which can take with a suggested reading speed 200 words per minute [20] about 8-65 minutes. Motivated by these estimates we concluded based on earlier research, we suggest that adoption of vocabulary by exploration in hyperlink network of vocabulary can be usefully carried out in a daily process that resembles reading 80-648 sentences.

Since each hyperlink in the Wikipedia typically has its own phrase (in article text surrounding hyperlink anchor) defining the relationship between start concept and end concept, and since the shortest path between a pair of concepts has on average 4,573 hyperlink steps in the Wikipedia [27], knowledge adoption of 80-648 sentences per day can be considered to correspond to traversing shortest paths of about 17-142 average pairs of concept in hyperlink network of vocabulary. Based on previous recommendations of about 3-4 spaced exposures to enable fertile learning ([11]; [13]; [12]), it seems that traversing 17-142 shortest paths can be considered to correspond (i.e. when dividing the number of shortest paths by 3 or 4 to enable 3-4 repetitions) an aim to learn connectivity relying on the shortest paths for about 4-47 pairs of concepts with every daily session of exploring hyperlink network of vocabulary. This result can be contrasted with and seems to resemble earlier estimates that a student can adopt daily about 4-9 new words ([17]; [18]; [3]).

We hope that the proposed framework can open new possibilities for developing innovative methods of computer-assisted learning relying on knowledge structures managed with small-world networks. We suggest that personalization of learning activities can benefit from exploring collaboratively built and gradually updated free knowledge resources of the Wikipedia online encyclopedia that inherently offers diverse collection of hyperlinks defining conceptual relationships usable for varied pedagogic purposes. We think that the principle of cumulatively expanding hyperlink networks covering more and more linkage between concepts of gradually growing

Table 2. - Some of the highest-ranking hyperlinked concepts and strongly rising hyperlinked concepts for cumulative vocabularies of six language ability levels of English Vocabulary Profile ranging from A1 to C2.

English Vocabulary Profile: A1				*English Vocabulary Profile: A1&A2*				*English Vocabulary Profile: A1&A2&B1*			
As start concept		As end concept		As start concept		As end concept		As start concept		As end concept	
High	Rising	High	Rising	High	Rising	High	Rising	High	Rising	High	Rising
food (22)	N/A	animal (21)	N/A	food (38)	water (9,5s->2)	water (48)	sun (11,5s ->5,5s)	human (61)	time (12->7)	human (68)	science (20->6)
month (19)	N/A	water (20)	N/A	water (33)	shoe (13,5s->5)	animal (41)	fruit (9,5s->5,5s)	food (60)	book (15,5s->14)	animal (67)	music (32->18,5s)
supermarket; party (18)	N/A	food (17)	N/A	toy (30)	game (18,5s->7,5s)	food (34)	television (19,5s->7,5s)	entertainment (57)	painting; kitchen (15,5s->15,5s)	water (66)	plant (32->22,5s)
plant; bread; meal (16)	N/A	fish; rice (16)	N/A	supermarket; nature; shoe (27)	soup (8->7,5s)	wood (32)	sugar (11,5s->7,5s)	water (56)	fruit (20,5s->17,5s)	earth (62)	art (22->22,5s)
soup (15)	N/A	day; year; milk (15)	N/A	soup; game (25)	bread (6->9,5s)	fruit; sun (29)	bird (19,5s->9)	transport (50)	artist (46,5s->21,5s)	turkey (53)	business (36,5s->28,5s)
water; house (14)	N/A	month; fruit (14)	N/A	month; bread (24)	time (13,5s->12)	sugar; television (26)	meat (13->11,5s)	nature (46)	sausage (30->21,5s)	science (49)	computer (41->31,5s)
lunch (12)	N/A	sugar; sun (13)	N/A	time; plant; party (23)	kitchen (40->15,5s)	bird (25)	milk (7->11,5s)	time (41)	sky; wind (20,5s->21,5s)	food (45)	time (26->31,5s)
fruit; time; garden; shoe (11)	N/A	meat (12)	N/A	book; red; kitchen; painting (22)	book (18,5s->15,5s)	meat; milk; leather; fish (24)	paper (31->20)	shoe (40)	pizza (26->26,5s)	wood (44)	history (78,5s->34,5s)
day; year; book; drink; november; game (10)	N/A	cheese; tea; sheep (11)	N/A	meal (21)	wind (54,5s->20,5s)	light; plastic; temperature; rice (23)	radio (49,5s->26)	soup; toy (37)	sea (63,5s->30)	sun (43)	physics (32->34,5s)
milk; grass; tomato (9)	N/A	plant; time; bird; horse; television; computer (10)	N/A	fruit; wind; salad; sky (20)	fruit (13,5->20,5s)	insect (22)	wine (31->26)	technology; mind (35)	meat (36,5s->30)	religion (42)	language (66->40)

English Vocabulary Profile: A1&A2&B1&B2				*English Vocabulary Profile: A1&A2&B1&B2&C1*				*Eng.Voc.Prof.: A1&A2&B1&B2&C1&C2*			
As start concept		As end concept		As start concept		As end concept		As start concept		As end concept	
High	Rising	High	Rising	High	Rising	High	Rising	High	Rising	High	Rising
human (98)	competition; science (48,5s->18,5s)	water (93)	law (26,5s->10,5s)	human (112)	abuse (50->21)	water (102)	genetics (65->50)	human (121)	philosophy (74->44)	animal (108)	war (65,5s->41)
food (78)	meat (30->18,5s)	animal (91)	government (50,5s->25,5s)	food (86)	evolution (69->38)	animal (100)	medicine (70,5s->53)	food (93)	cancer (80->58)	human; water (106)	logic (116->82,5s)
water (72)	reason (56,5s->24)	human (90)	language (40->34,5s)	water (82)	oxygen (56->39)	human (97)	crime (84->59,5s)	water (85)	ship (94->70)	earth (101)	death (127->88)
entertainment (68)	crime (93->30,5s)	earth (84)	biology (44,5s->40)	nature (75)	life (124->43)	earth (94)	statistics (90->65,5s)	nature (79)	death (92->74)	mammal (98)	police (127->95)
transport (66)	future (48,5s->30,5s)	carbon dioxide (70)	chemistry (50,5s->47)	entertainment (71)	title (77,5->64)	mammal (91)	aluminium (95,5s->84)	entertainment (74)	evaluation (100->81)	psychology (92)	profession (108->95)
nature (65)	music (37,5s->34)	philosophy; turkey (69)	disease (65,5s->51,5s)	transport (70)	rainforest (91,5s->67)	philosophy (83)	experiment (117->93,5s)	transport (72)	invasion (117->82)	philosophy (90)	contract (168->112)
mind (52)	culture (65->44)	religion (68)	technology (50,5s->51,5s)	infrastructure (64)	bird (124->79)	protein (81)	blood (108->93,5s)	nutrition (68)	creativity (109->87)	law (86)	climate change (148->112)
technology (48)	skin (80->50)	psychology (65)	society (56->55,5s)	nutrition (63)	cancer (143,5s->80)	religion (78)	child; heart (108->101)	mind (66)	civilization (155->100)	religion; protein (85)	system (138->112)
time; shoe (47)	insect (65->50)	culture; law (64)	knowledge; war (109,5s->59,5s)	globalization; mind (58)	milk (91,5s->81)	psychology; carbon dioxide (76)	trade (158->108)	infrastructure (65)	mask (133->112)	science; carbon dioxide (80)	fear (168->121,5s)
plant (45)	garden; writer (56,5s->61,5s)	science (63)	money (99,5s->65)	shoe (54)	death (143,5s->92)	turkey (75)	risk; tool (139,5s->116)	globalization (63)	cloud (127->119)	turkey (79)	mind (157,5s->121,5s)

vocabulary can enable an efficient and intuitive way to explore and adopt new knowledge meaningfully as well as to develop new kind of educational games that can be extended to manage diverse content besides text like images, videos, and tasks with augmented reality and tracking kinetic activities.

References

[1] Wang, J., Zuo, X., & He, Y. (2010). Graph-based network analysis of resting-state functional MRI. Frontiers in Systems Neuroscience, 4:16.

[2] Ingawale, M., Dutta, A., Roy, R., & Seetharaman, P. (2009). The small worlds of Wikipedia: implications for growth, quality and sustainability of collaborative knowledge networks. Proc. Americas Conference on Information Systems (AMCIS 2009).

[3] Nation, P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In Schmitt, N., & McCarthy, M. (eds.), Vocabulary: Description, acquisition, pedagogy. Cambridge University Press, New York, USA, 6-19.

[4] Nation, I. (2006) How large a vocabulary is needed for reading and listening? Canadian Modern Language Review 63, 1, 59-82.

[5] Laufer, B. 1989. What percentage of text-lexis is essential for comprehension? In C. Lauren and M. Nordman (eds.), Special Language: From Humans Thinking to Thinking Machines. Clevedon: Multilingual Matters.

[6] Council of Europe (2001). Common European Framework of Reference for Languages: Learning, Teaching, Assessment. Council of Europe. Cambridge University Press, ISBN 0521803136. http://www.coe.int/t/dg4/linguistic/Source/CECR_EN.pdf

[7] North, B. 1996/2000: The development of a common framework scale of language proficiency. PhD thesis, Thames Valley University. Reprinted 2000, New York, Peter Lang.

[8] European Commission (2012). Language competences for employability, mobility and growth. Commission staff working document. http://ec.europa.eu/education/news/rethinking/sw372_en.pdf

[9] Cambridge English for Speakers of Other Languages (ESOL) / Cambridge English Language Assessment (2013). http://www.cambridgeesol.org/about/standards/can-do.html, redirected to http://www.cambridgeenglish.org/about-us/what-we-do/international-language-standards/ (as of 18 August 2013).

[10] Sanatullova-Allison, E. (2009). Less commonly taught languages: often overlooked but equally important. Language Association Journal, 60(2).

[11] Thalheimer, W. (2006). Spacing learning events over time: what the research says. Publication of Work-Learning Research Inc., March 2006, http://willthalheimer.typepad.com/files/spacing_learning_over_time_2006.pdf

[12] Kandel, E. (2001). The molecular biology of memory storage: a dialog between genes and synapses. Nobel Lecture, 8 December 2000. Bioscience Reports, 21(5).

[13] Fields, R. (2005). Making memories stick. Scientific American, 292 (February 2005), 74-81.

[14] Gilkerson, J., & Richards, J. (2009). The power of talk, 2nd Edition. Impact of adult talk, conversational turns, and TV during the critical 0-4 Years of child development. LENA Technical Report LTR-01-2. LENA Research Foundation. http://www.lenababy.com/pdf/The_Power_of_Talk.pdf

[15] Mehl, M., Vazire, S., Ramirez-Esparza, N., Slatcher, R., & Pennebaker, J. (2007). Are women really more talkative than men? Science, 317, 82.

[16] Juster, F., Ono, H., & Stafford, F. (2004). Changing times of American youth: 1981-2003. Institute for Social Research, University of Michigan, Ann Arbor, Michigan, USA (November 2004), http://www.ns.umich.edu/Releases/2004/Nov04/teen_time_report.pdf

[17] Lehr, F., Osborn, J. & Hiebert, E. (2004). Research-based practices in early reading series: a focus on vocabulary. Regional Educational Laboratory, Pacific Resources for Education and Learning. http://vineproject.ucsc.edu/resources/A%20Focus%20on%20Vocabulary%20PREL.pdf

[18] Kuhn M., & Stahl S. (1998). Teaching children to learn word meanings from context: a synthesis and some questions. Journal of Literacy Research, 30(1), 119-138.

[19] Anderson, R., Wilson, P., & Fielding, L. (1988). Growth in reading and how children spend their time outside of school. Reading Research Quarterly, 23, 285–303.

[20] Lewandowski, L., Codding, R., Kleinmann, A., & Tucker, K. (2003). Assessment of reading rate in postsecondary students. Journal of Psychoeducational Assessment, 21, 134-144.

[21] Anderson, N. (1999). Improving reading speed. English Teaching Forum, 37(2).

[22] DuBay, W. (2004). The principles of readability. Impact Information, Costa Mesa, California, USA. http://www.impact-information.com/impactinfo/readability02.pdf

[23] Najork, M., Zaragoza, H., & Taylor, M. (2007). HITS on the Web: how does it compare? Proc. 30th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 471-478.

[24] Kamps, J., & Koolen, M. (2009). Is Wikipedia link structure different? Proc. Second ACM International Conference on Web Search and Data Mining (WSDM 2009), 232-241.

[25] Zlatic, V., Bozicevic, M., Stefancic, H., & Domazet, M. (2006). Wikipedias as complex networks. Physical Review E 74, 016115 (2006).

[26] Rodero-Merino, L., Fernández Anta, A., López, L., & Cholvi, V. (2010). Performance of random walks in one-hop replication networks. Computer Networks 54 (2010) 781–796.

[27] Dolan, S. (2011). Six degrees of the Wikipedia, Stephen Dolan, Trinity College, Dublin, Irland. http://mu.netsoc.ie/wiki/

[28] Travers, J., & Milgram, S. (1969). An experimental study of the small world problem. Sociometry 32: 425-443.

[29] Backstrom, L., Boldi, P., Rosa, M., Ugander, J., & Vigna, S. (2011). Four degrees of separation. Proc. 4th ACM International Conference on Web Science (WebSci).

[30] Chesney, T. (2006). An empirical examination of Wikipedia's credibility. First Monday, 11(11).

[31] Konieczny, P. (2012). Wikis and Wikipedia as a teaching tool: five years later. First Monday, 17(9).

[32] Lahti, L. (2012). Educational framework for adoption of vocabulary based on Wikipedia linkage and spaced learning. Proc. Global Learn 2012 (pp. 8-13), online conference on 6 November 2012 organized by AACE. http://www.editlib.org/p/42033/. spt vltvvt wjsbkluly ulzwwl dlwlsbt atpw lcötyvz ölvvlcopabl

[33] Capel, A. (2013). Completing the English Vocabulary Profile: C1 and C2 vocabulary. English Profile Journal, 3, e1.

[34] Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the New General Service List. Applied Linguistics, fist published online August 26, 2013. doi: 10.1093/applin/amt018.

[35] Leech, G., Rayson, P., & Wilson, A. (2001). Word frequencies in written and spoken English: based on the British National Corpus. Longman, London, United Kingdom. ISBN 0582-32007-0. A companion web site: http://ucrel.lancs.ac.uk/bncfreq/flists.html