NETWORK BASED ANALYSIS OF POLICY DOCUMENTS ON SCIENTIFIC LITERACY

ABSTRACT

We're still working on this article.

INTRODUCTION

Scientific literacy is a complex concept. It has many different interpretations as used in policy documents, research articles, and as understood by teachers (Roberts, 2007). As part of the project Mind the Gap (Bruun, Mind The Gap, 2009), we have developed at method for finding structure in documents on scientific literacy. This article focuses on the network and information theoretical part of the method. Questions relating to how the method is incorporated in discussions about scientific literacy are addressed elsewhere (Bruun, Evans, & Dolin, 2009).

A person conveying information in writing has different ways of emphasizing the importance of different parts of the information. If a particular concept is important, the person may choose to use a word or a set of words describing it many times. Another way of emphasizing is to link the concept to many other concepts. In any of these two cases the concept will stand out in the text. A reader going through the document will encounter this concept many times or in a lot of different contexts. This is the basic assumption of our method.

Lately, authors (Masucci & Rodgers, 2006)have worked with language as a form of complex network. In these networks, labeled linguistic networks, the words are nodes, and a link from node *k* to node *l* exists, if *k* precedes *l*. See figure 1.

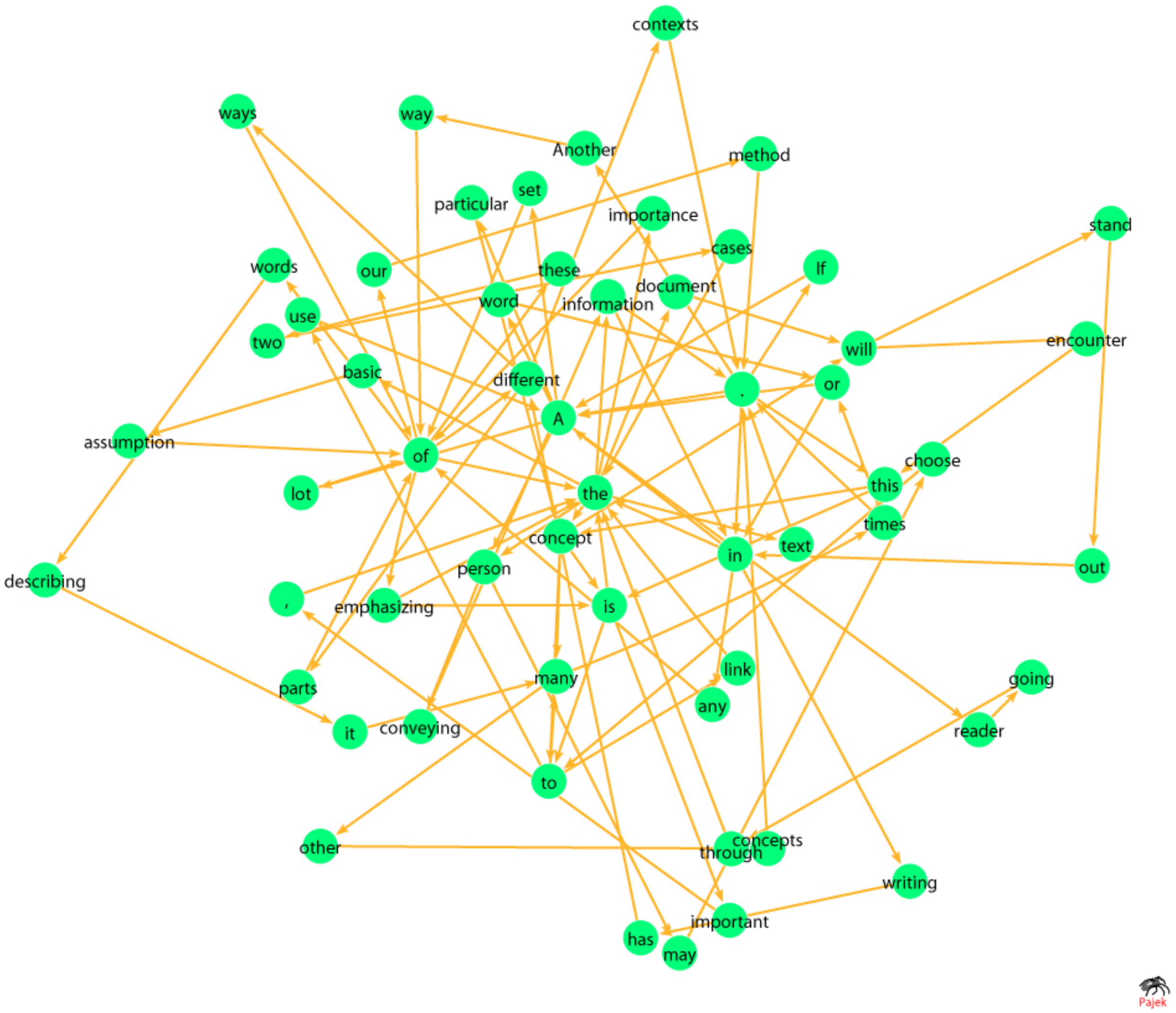


FIGURE 1: THE SECOND PARAGRAPH OF THE ARTICLE AS A LINGUISTIC NETOWRK. IF YOU CAN FOLLOW THE ARROWS, YOU CAN ACTUALLY READ THE SECOND PARAGRAPH USING THIS GRAPH.

In these networks the degree of a node is proportional to the number of times the word is used in the text. In such a network, words like *the*, *a*, *and* become very important hubs, since they are used frequently. However, it has been shown, that these words are randomly distributed in *fictional literature*, meaning that in a given word sample the probability distribution of these words are Gaussian. For this reason, it is reasonable to shortcut such words. They serve to maintain a semantic structure in the network. We call these words structural words.

Besides the structural words, a text will have an effective vocabulary of words which are used to convey information to the reader of the text. Knowing only the common structural words will not provide the reader with information about the content of the document, while the structure of the text vocabulary will. Compressing the text to a network in which every word in the vocabulary is used only once, provides a map where the structure is captured by the links between words and their link strengths.

Documents explaining scientific literacy tend to have sentences with many messages. In this method we manually parse the messages in to one message sentences, which are then mapped to a network using the thoughts explained above.

METHOD

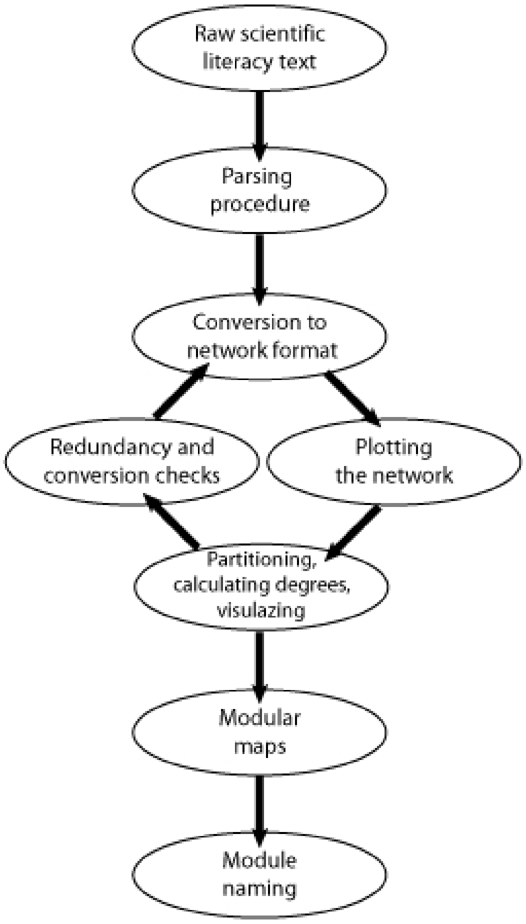

Raw scientific literacy text
Parsing procedure
Conversion to network format
Redundancy and conversion checks
Plotting the network
Partitioning, calculating degrees, visulazing
Modular maps
Module naming

Figure 2: Method overview

Figure 2 shows the overview of our method. A raw text on scientific literacy (SL text) is parsed to a state, in which all sentences contain one statement. The parsed SL text is then converted into a network where non structural words are nodes and links between nodes exists if they are adjacent to each other or if a set of structural words separate them. Now, an iterative process begins, where we plot the network, label the words according to function, calculate degrees of words, visualize, and check for redundant words and conversion errors. In the final stages we use an information theoretical computer algorithm simulating the information flow in the network (Rosvall & Bergstrom, 2008), to group the words in modules, and finally we name the modules according to the most prominent nodes in them. The whole process should be considered iterative, since it is possible to compare both types of networks (modular and linguistic) to the original text.

We illustrate the method by an example. We want to make a map of the PISA 2006 definition of Scientific Literacy. According to this, scientific literacy is:

An individual's scientific knowledge and use of that knowledge to identify questions, to acquire new knowledge, to explain scientific phenomena, and to draw evidence-based conclusions about science-related issues, understanding of the characteristic features of science as a form of human knowledge and enquiry, awareness of how science and technology shape our material, intellectual, and cultural environments, and willingness to engage in science-related issues, and with the ideas of science, as a reflective citizen.

OECD (2006). ***Assessing Scientific, Reading and Mathematical Literacy – A Framework for PISA 2006,*** **OECD, Paris.**

PARSING

In this definition many words are connected to other words in list form. We split sentences with this property up, leaving us with many sentences containing only one statement. We do this by using our parsing rules iteratively[1]. We wish to incorporate the individual, so we introduce the

[1] See appendix A for the rules.

scientifically literate person and make necessary changes. Finally commas, and ands used for listing adds a statement to the sentence. Sentences including *commas* and *ands* used for listing are therefore split up. The final result is:

- *A scientifically literate person should have scientific knowledge.*
- *A scientifically literate person uses scientific knowledge to identify questions about science related issues.*
- *A scientifically literate person uses scientific knowledge to acquire new knowledge about science related issue.*
- *A scientifically literate person uses scientific knowledge to explain scientific phenomena about science related issues.*
- *A scientifically literate person uses scientific knowledge to draw evidence based conclusions about science related issues.*
- *A scientifically literate person understands the characteristic features of science as a form of human knowledge*
- *A scientifically literate person understands the characteristic features of science as a form of enquiry.*
- *A scientifically literate person is aware of how science shapes our material environments.*
- *A scientifically literate person is aware of how science shapes our intellectual environments.*
- *A scientifically literate person is aware of how science shapes our cultural environments.*
- *A scientifically literate person is aware of how technology shapes our material environments.*
- *A scientifically literate person is aware of how technology shapes our intellectual environments.*
- *A scientifically literate person is aware of how technology shapes our cultural environments.*
- *A scientifically literate person is willing to engage in science-related issues, as a reflective citizen.*
- *A scientifically literate person is willing to engage with the ideas of science, as a reflective citizen.*

We note here that this particular way of parsing the sentences is biased. It builds on our interpretation of the text, and the text is ambiguous. This could be analyzed by testing different interpretations of the text and noting differences. In this way the method would provide a visualization of different interpretations.

CONVERSION TO NETWORK FORMAT

To illustrate how we build the network we have plotted the first sentence and the first and second sentences. First the sentences are put in to a temporary format in a spread sheet.

Table 1: The words in the sentences above are put in a spread sheet counting the number of times one word follows another. This number is listed in the third column. The fourth column is for structural words, which go between the two words of the first rows. This table shows the spread sheet for the first two sentences above.

Scientifically literate person	has	1	
has	scientific	1	

scientific	knowledge	2	
Scientifically literate person	use	1	
use	scientific	1	
knowledge	identify	1	to
identify	question	1	
question	science-related	1	about
science-related	issue	1	

Now, we use a computer program[2] to convert the spread sheet format in to a format readable for the Pajek software. Finally, we put in the structural words as labels (de Nooy, Mrvar, & Batagelj, 2005). The resulting file formats can be seen in the textboxes below.

```
*Vertices 9

1 "Scientifically literate person"
2 "has"
3 "scientific"
4 "knowledge"
5 "use"
6 "identify"
7 "question"
8 "science-related"
9 "issue"
*Arcs
1 2 1
2 3 1
3 4 2
1 5 1
5 3 1
4 6 1
6 7 1
7 8 1
8 9 1
```

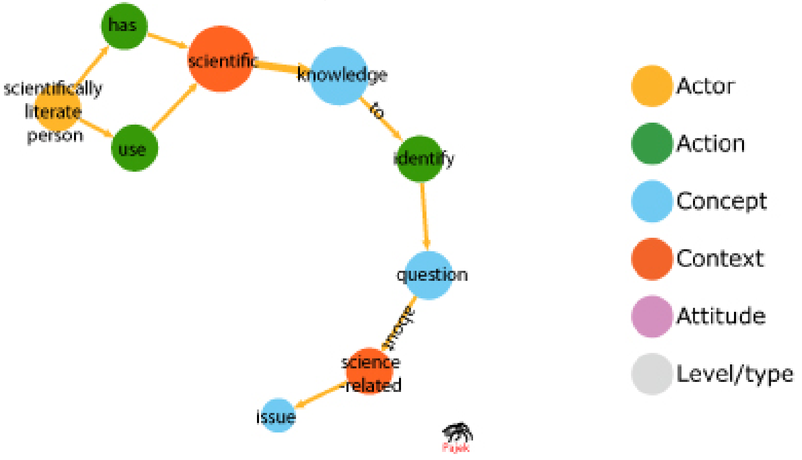
```
*Vertices 9

1 "Scientifically literate person"
2 "has"
3 "scientific"
4 "knowledge"
5 "use"
6 "identify"
7 "question"
8 "science-related"
9 "issue"
*Arcs
1 2 1
2 3 1
3 4 2
1 5 1
5 3 1
4 6 1 l "to"
6 7 1
7 8 1 l "about"
8 9 1
```

Finally, we plot the network using Pajek. The size of a node is proportional to the number of connections entering and leaving the node. Thus, *issue*, is the smallest, since it only has one connection going in, while *scientific* is the largest, since it has 2 going in and 2 leaving. The two connections leaving have been collected in one arrow signifying the strength between scientific and

[2] Txt2Pajek can be downloaded from: ….

knowledge. Finally, the colors of nodes show our labeling. The details of this visualization are discussed in the next section.



LABELING, VISUALIZING AND CHECKING

We label the different words according to their function in the sentence. We operate with seven different categories:

1. Actor. The student, citizen, scientifically literate person.
2. Action. Verbs describing physical or cognitive processes of the actor.
3. Concept. Nouns describing what the actor acts on.
4. Context. Primarily adjectives but also nouns which put the actor-action-concept relation in to some context.
5. Attitude. Verbs and adjectives which describe attitudes.
6. Level/type. Words which define a type or level of concept, context, action, or attitude.
7. Structural words. Discussed above.

The categories above have been chosen on the basis of previous studies in scientific literacy (Roberts, 2007) to allow for a visual grasping of the network. As such they are only meaningful categories in the context of scientific literacy, the exception being structural words.

After assigning values to the words on the basis of the sum of connections going in and out from them (thus implying their rank (REF ZIPF)), we have a graph of the network. The actual realization of the graph on *Figure 3* is due a specific energy algorithm, where the links between nodes are treated as springs and nodes as particles. The algorithm then searches for an equilibrium state of the network. Other plotting algorithms are possible (de Nooy, Mrvar, & Batagelj, 2005).

The in-degree (out-degree) of a node in this network is the total strength of in-coming (out-going) links. Most links will have their out-degree equal to their in-degree. This does not hold true for *beginning* nodes nor for *end* nodes. Beginning nodes start sentences, while end nodes finish them. One way of checking whether the conversion from the parsed state to the network state is consistent is by calculating the difference between in- and out-degree for all nodes. Only end and beginning nodes should have non-zero values here.

Listing nodes alphabetically offers a way to determine redundant nodes. Some redundancies stem from difference in spelling, while others originate in differences in past and present tenses, or between singular and plural forms. We dispose of redundant nodes by reducing all spellings and forms to the same form as the word would have in a dictionary.

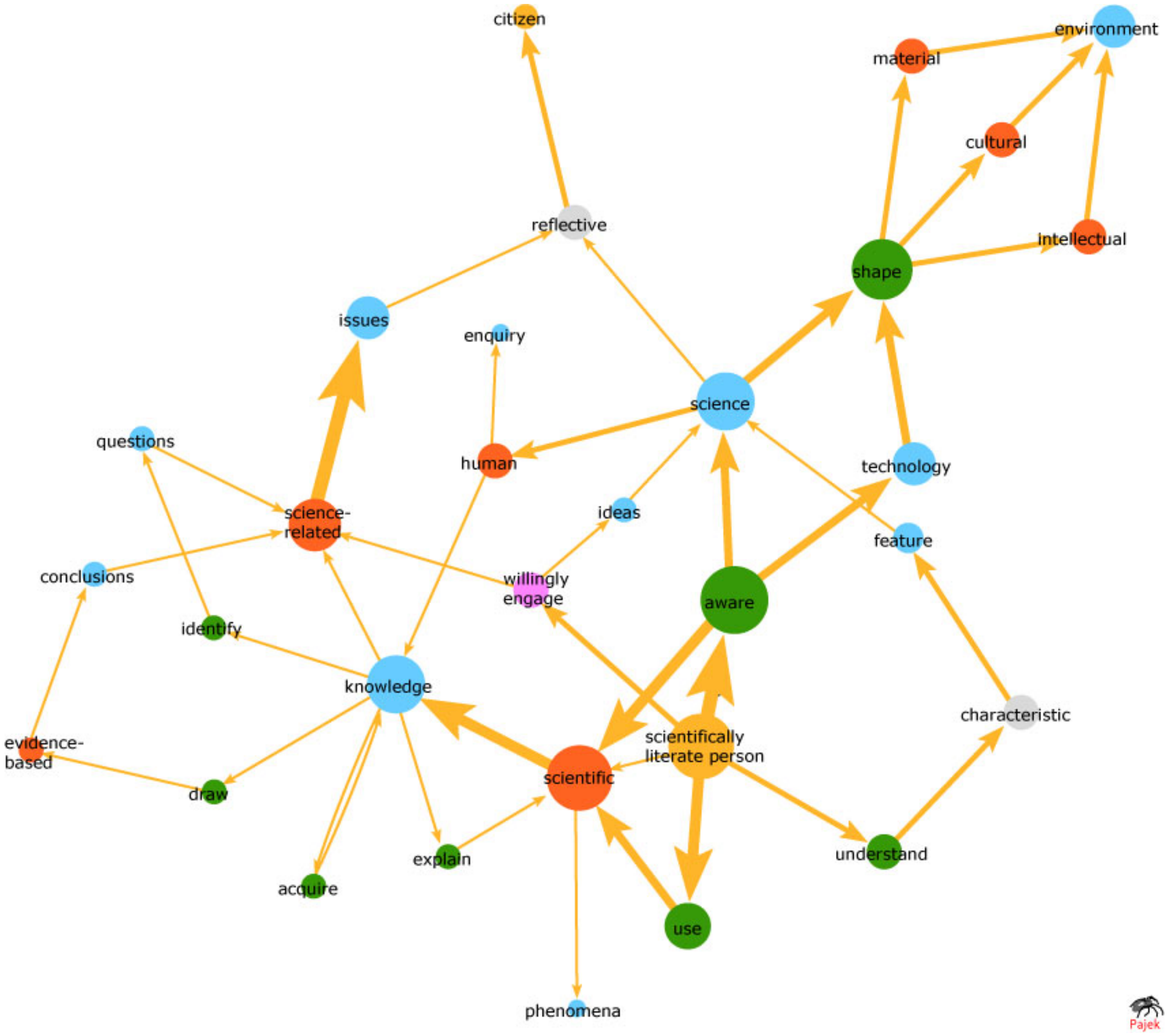


Figure 3: A possible map of the PISA definition of scientific literacy. The sizes represent the number of times a word is mentioned. Structural words have been left out. The size of the arrows represents how many times one word is mentioned after another. The colors represent a labeling scheme.

Figure 2 shows one possible map of the PISA definition of scientific literacy following this approach. Initially, the nodes had the same color and size. Analyzing a number of texts, we built up and constantly updated a vocabulary of words used in texts on scientific literacy.

Each node has a number of connections going in and out. For each node we noted both in connections, out connections, and the sum of in and out connections. In the analysis, we plotted nodes with their sizes proportional to the sum of in and out connections. This allowed us to compare the same node (word) in different maps (contexts). Also, we used the differences in *in* and *out* connections to find end nodes (words at the end of sentences) and to check for errors.

Comparing the size of a word, groups of words, or links between words in different maps, allowed for a structural analysis of the texts, part of which came from the MDL data.

We use the minimum description length (MDL) approach (Rosvall & Bergstrom, 2008) to find sentence bits which are closely connected. In the MDL approach, a random walker traverses the network. When going from one node to an adjacent node, the walker is allowed to follow the arrows alone. If more than one arrow leaves a node the walker chooses one of the arrows with a probability proportional to the thickness (strength) of the arrows. A random walker may easily get stuck, so the MDL algorithm gives the walker a probability for jumping to any node in the network at each step of the walk. This jumping random walker is called a random surfer (Rosvall & Bergstrom, 2008). After a sufficient amount of surfing each node has been visited with a certain frequency. It is possible to describe the path of a surfer with a code, and by grouping nodes into modules in which the surfer spends a lot of time, the MDL algorithm minimizes the description length. In the context of reading a text, the random surfer can be interpreted as a person skimming through the text. We have used the tool on texts from three different countries to allow for comparison.

The MDL data is modules of nodes. Each module thus becomes a node in a new network. Each module has a size proportional to the time the surfer spends visiting nodes in the module and two modules are connected when the surfer goes from one module to another. In our analysis of the MDL data, we have named modules according to the nodes in them. Figure 4 shows two modular versions of the network in figure 3.

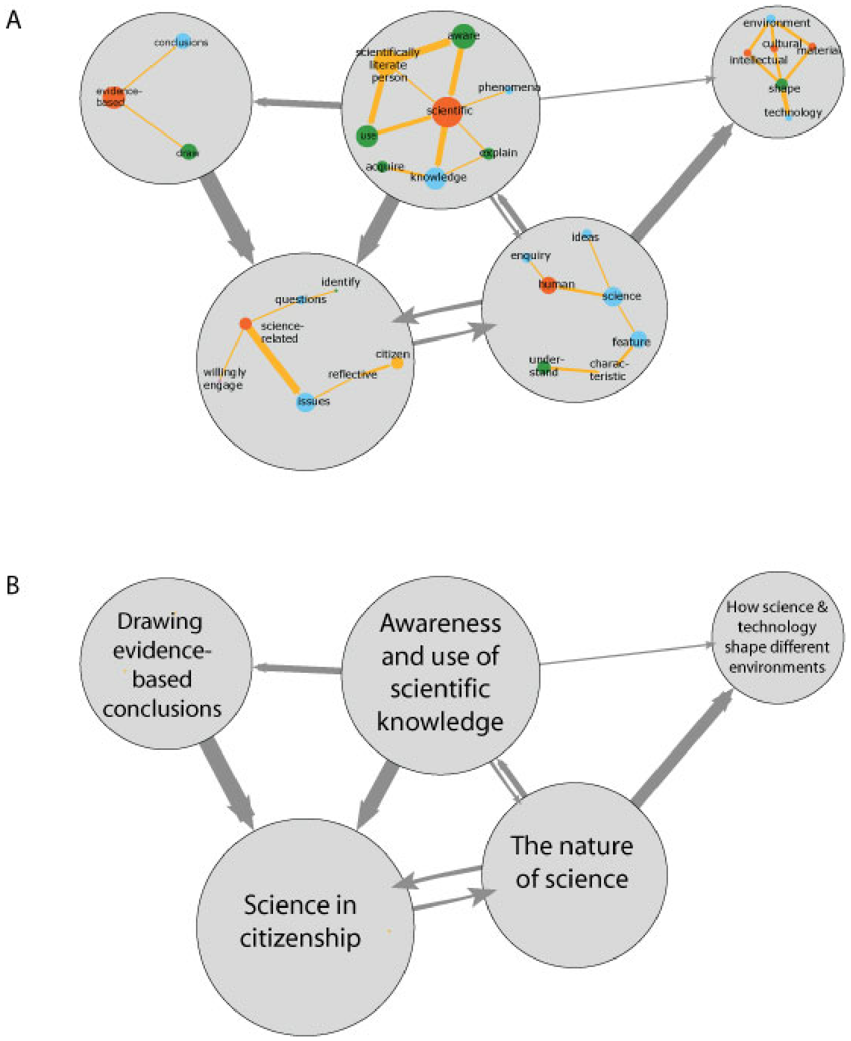


Figure 4: The PISA definition after the MDL algorithm has been applied. Graph A shows the nodes in the different modules. Graph B shows the modules after our naming. The sizes of the modules represent the amount of time a random surfer spends in each module. The arrows between modules represent the probability for travelling between the modules.

ANALYSIS AND RESULTS

It is possible to analyze the network before and after the creation of modular maps. Policy documents from Hungary, England, and Denmark have been mapped so far - as shown in figures 5-6.

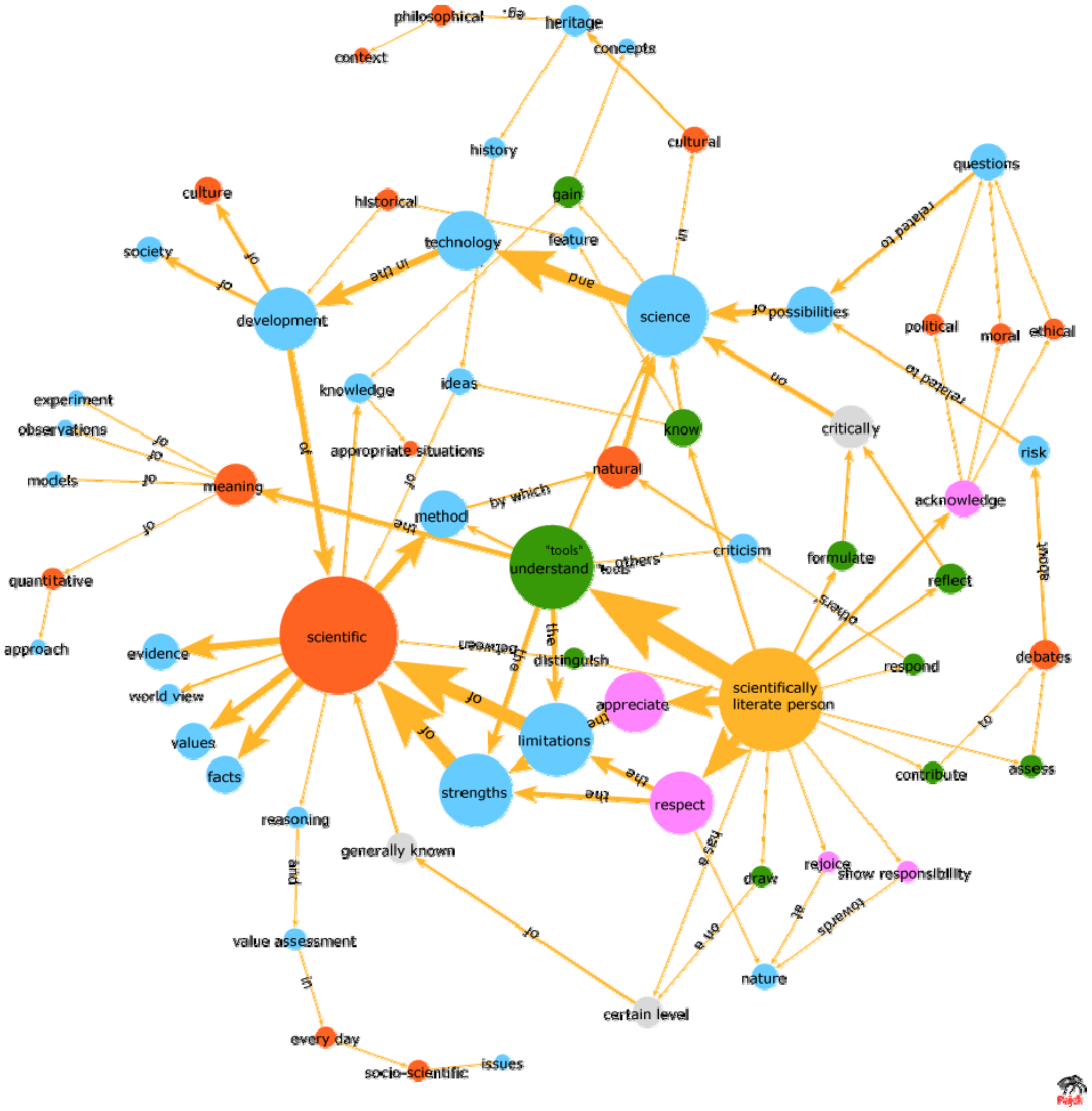


FIGURE 7: MAP OF SCIENTIFIC LITERACY IN THE DANISH POLICY DOCUMENT: FREMTIDENS NATURFALIGE UDDANNELSER (UVM, 2006).

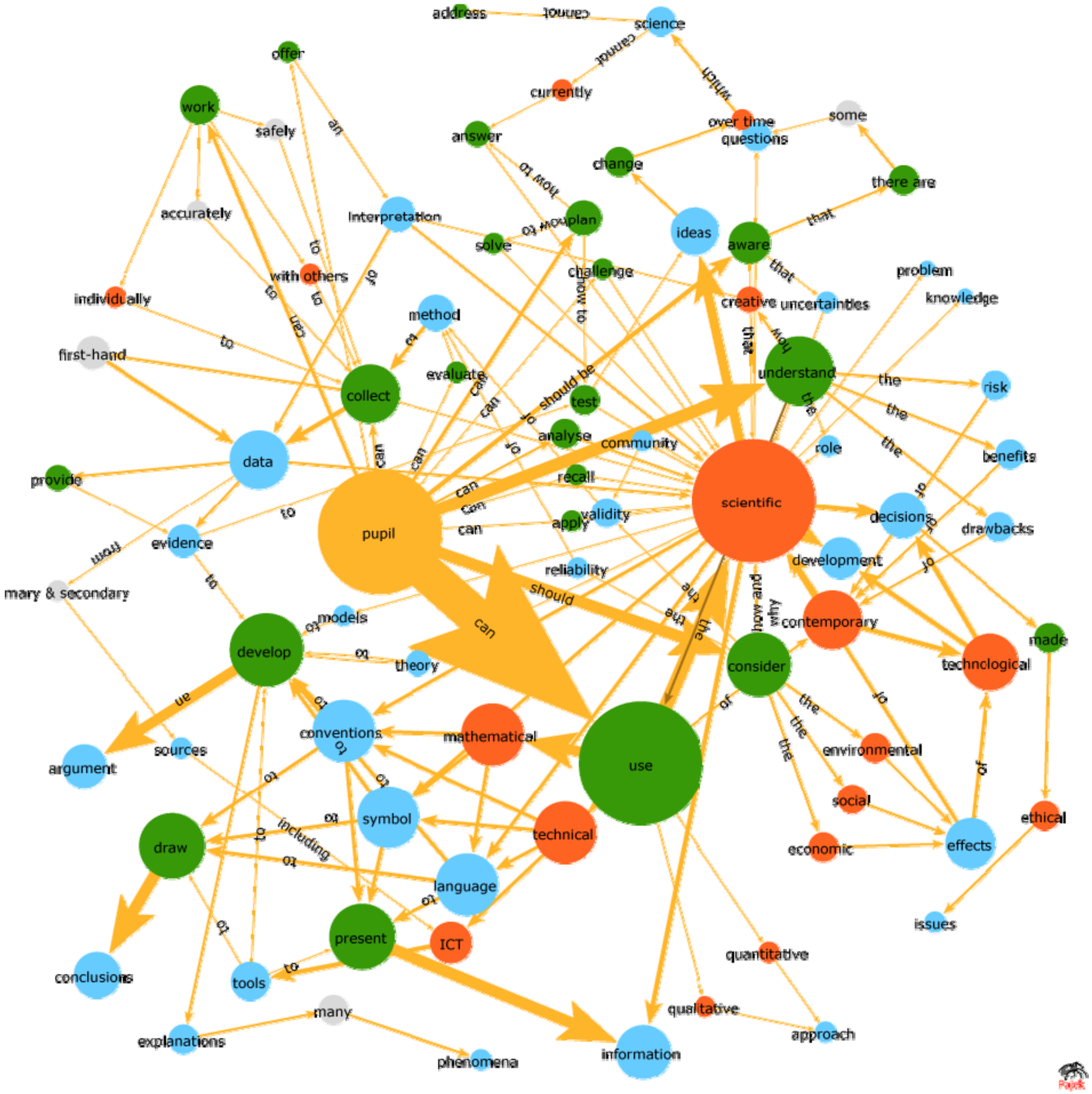


FIGURE 8: MAP OF SCIENTIFIC LITERACY IN THE ENGLISH NATIONAL CORE CURRICULUM KEY STAGE 4.

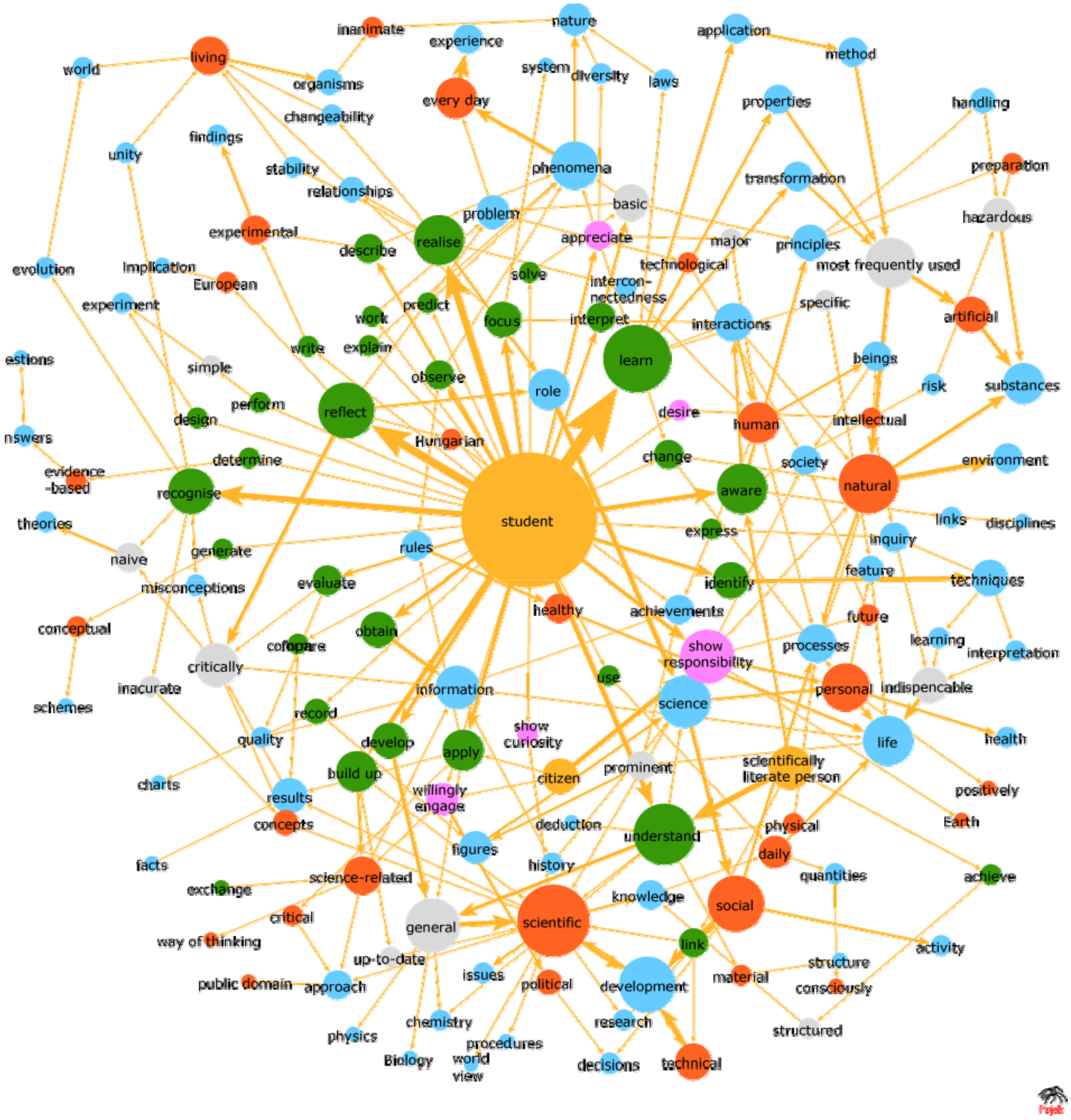


FIGURE 9: MAP OF SCIENTIFIC LITERACY IN A HUNGARIAN POLICY DOCUMENT.

In all three maps, the word *scientific* is the most prominent node apart from the chosen primary actor (student, pupil, scientifically literate person). Comparing the Danish and English maps, we see a huge emphasis on use in the English map, while the Danish map emphasizes understanding. Also, the English map has no attitudinal nodes, while the Danish map emphasizes *respect* and *appreciate*. The Hungarian map consists of many nodes which are only used once or twice, while the Danish and English maps consist of few nodes which are used many times.

Table SUMMARY summarizes the properties of the three networks

TABLE SUMMARY

DISCUSSION

CONCLUSION

REFERENCES

Bruun, J. (2009, January 16). *Mind The Gap.* Retrieved January 16, 2009, from Institut for Naturfagenes Didaktik: http://www.ind.ku.dk/Forskning/forskningsprojekter/MTG/parsing.doc

Bruun, J., Evans, R., & Dolin, J. (2009). Investigating Scientific Literacy with linguistic network analysis (submitted). *ESERA 2009 book.* Istanbul: ESERA 2009 Conference.

Ferrer, R., Solé, R. V., & Köhler, R. (2004). Patterns in syntactic dependency networks. *Physical Review E* .

Masucci, A. P., & Rodgers, G. J. (2006). Network properties of written human language. *Physical Review E* , (026102-1)-(026102-8).

Roberts, D. A. (2007). Scientific Literacy/Science Literacy. In S. K. Abell, N. G. Lederman, & (eds.), *Handbook of research on science education* (pp. 729-780). Mahwah, New Jersey: Lawrence Erlbaum Associates, Inc.

Rosvall, M., & Bergstrom, C. T. (2008). Maps of random walks on complex networks reveal community structure. *PNAS* , 1118-1123.

APPENDIX A: PARSING RULES

DEFINITIONS

SENTENCE

Normal definition from grammatical theory

PARSED SENTENCE

A sentence which has been extracted from the text using the rules and procedures described below.

LIST COMMA AND LIST *AND*

A list comma is used to list a number of different features or properties which all belong to a set of actors, contexts, and/or concepts. They can be identified by the following rule: If you replace a list comma with an "*and"* the meaning is unaltered. The list commas and "*ands*" can separate *list items*. A list can be placed in the beginning of, in the middle of or at the end of a sentence.

LIST ITEMS

Items of a list: single words, groups of words or even sentence structures.

SENTENCE SPLITTING

A sentence containing a class of n_i list commas and list "*ands*" can be split in to $n_i + 1$ sentences, each of which contains one of the list items. The parts of the sentence which do not belong to the list are reused in each of the split sentences.

ACTIVE AND PASSIVE FORMS

Active form: The student uses a mathematical model.

Passive form: A mathematical model is used by the student.

NOUNS DERIVED FROM VERBS

Awareness may be viewed as a derivative of *being aware*. So "The student has awareness of" can be changed to "The student is aware of". *Knowledge* could in the same manner be viewed as a derivate of to *know*. However, "*having knowledge of*" seems to be qualitatively different from "*to know*"

SENTENCE SPLITTING PROCEDURE

1. Identify sentence
2. Find all commas in the sentence
3. Identify all list commas and list "*ands*"
4. Classify list commas and list "*ands*"
5. Split a sentence with a class of n_i commas in to $n_i + 1$ new sentences
6. Repeat steps 1-5 until all list commas and list "*ands*" have been eliminated.

SENTENCE ADJUSTMENT PROCEDURE

1. Identify subject of sentence
2. Change passive forms to active forms
3. Change sentence subject to relevant actor
4. Change nouns derived from verbs to active verb form

CHANGING GOALS TO DESCRIPTIONS

Some SL-texts express goals for the student. In our parsing procedure we change all verbs describing goals to verbs describing the scientifically literate actor. Thus for example, "*should have*" is changed to "*has*".

SENTENCE ELIMINATION (OPTIONAL)

In some texts statements describing the scientifically literate person (the SL-person) may be weaved in between statements describing other things – for example teaching methods. If only statements describing the SL-person are necessary, other sentences should be eliminated.

FINAL CHECK

1. A parsed sentence may have precisely one statement.
2. A parsed sentence must start with an actor.
3. The statement describes features about the actor.
4. A parsed sentence has no passive forms of verbs.
5. A parsed sentence has no nouns derived from verbs.
6. The parsed sentence must be grammatically meaningful.
7. (optional) All sentences describe the actor.