Education and Information Technologies (2023) 28:14325–14346
https://doi.org/10.1007/s10639-022-11531-5

Network analysis of terms in the natural sciences insights from Wikipedia through natural language processing and network analysis

Peter Wulff[1]

Received: 21 March 2022 / Accepted: 11 December 2022 / Published online: 5 April 2023

Abstract

Scientists use specific terms to denote concepts, objects, phenomena, etc. The terms are then connected with each other in sentences that are used in science-specific language. Representing these connections through term networks can yield valuable insights into central terms and properties of the interconnections between them. Furthermore, understanding term networks can enhance assessment and diagnostics in science education. Computational means such as natural language processing and network analysis provide tools to analyze term networks in a principled way. This study utilizes natural language processing and network analysis to analyze linguistic properties of terms in the natural science disciplines (biology, chemistry, and physics). The language samples comprised German and English Wikipedia articles that are labelled according to the respective discipline. The different languages were used as contrasting cases. Natural language processing capabilities allowed us to extract term networks from the Wikipedia articles. The network analysis approach enabled us to gain insights into linguistic properties of science terms and interconnections among them. Our findings indicate that in German and English Wikipedia terms such as theory, time, energy, or system emerge as most central in physics. Moreover, the science-term networks display typical scale-free, complex systems behavior. These findings can enhance assessment of science learner's language use. The tools of natural language processing and network analysis more generally can facilitate information extraction from language corpora in the education fields.

Keywords Science education and language · Network analysis · Natural language processing · Wikipedia analyses

✉ Peter Wulff
peter.wulff@ph-heidelberg.de

[1] Physics and Physics Education, Heidelberg University of Education, Heidelberg, Germany

1 Motivation

Literacy in science has been argued to be related to language use in a fundamental sense (Norris & Phillips, 2003; von Weizsäcker, 2004). "Nothing resembling what we know as western science would be possible without text" (Norris & Phillips, 2003). Hence, studying linguistic phenomena in disciplines such as the natural sciences can point to potential affordances and challenges for becoming literate in them. In a formal sense, language is defined as a symbolic alphabet that forms words that form sentences constrained by the rules of grammar (Nowak et al., 2002). Language can be characterized by nested structures (words embedded in sentences), by hierarchical order among the elements (e.g., phrase structures) and other universal features (de Beule, 2008). Hence, language is complex by design, and potentially exhibits complex systems behavior.

From infancy, learners are confronted with linguistic stimuli in their respective communities and learn inductively to generalize the input to produce language (Wallace, 1996). Language learning was consequently characterized and shown to be a probabilistic process, where linguistic properties of the constituents of the language eventually direct the learning of it Jurafsky and Martin (2014).

Language is an important medium for representation of knowledge such as facts and relationships among concepts and terms. Science knowledge can be characterized by its hierarchical structure and interconnectedness of concepts (Nousiainen & Koponen, 2012). Science curricula in the United States (National Research Council, 2012) and Germany (KMK, 2020) stress the existence of core concepts that are central to science learning. For example, matter, system, energy, and interaction (force) are outlined in the German state physics curriculum to be pertinent to most topics in physics (KMK, 2020). It is unclear, however, to what extent these knowledge structures also manifest in science-specific language. Given the complexity and probabilistic nature of language, it would be desirable to develop principled and quantitative approaches to studying science language (Agrawal et al., 2016).

This study seeks to employ natural language processing and network analysis techniques to analyze science language in a principled (hence, reproducible) way and extract linguistic properties and interdependencies of science-related terms. To do so, we analyze widely used and well administered bodies of science language data, namely German and English Wikipedia articles that were categorized as science-related. Articles in the orders of 10 k could be collected to extract the structure and interdependencies of terms in science. Networks were then formed based on the interconnections of terms within sentences in these articles and formed the basis for our analyses.

2 Physics language

Communication and representation within the science disciplines is largely reliant on language: "Within the philosophy of science, it has typically been assumed that the fundamental representational resources are linguistic, mathematics being understood as a kind of language" (Giere, 2004). Language allows humans to "transfer

unlimited non-genetic information among individuals, and it gives rise to cultural evolution" (Nowak et al., 2002). Besides equations, graphs, and diagrams, language is one of the primary representational means to convey science ideas (Brookes & Etkina, 2009). More generally, Vygotsky's sociocultural theory of cognitive development states that learning and development is a socially-mediated process, in which language forms a primary means to convey cultural beliefs, values, and knowledge onto others (Vygotsky, 1978; 1963). Language has been labeled "the most pervasive system of semiotic resources" (Lemke, 1998). Humans use language to make sense of their science-related experiences and communicate them with others (Halliday & Matthiessen, 2007; Brookes & Etkina, 2015). Language is the means for humans to become acquainted with science contents and humans use language to make sense of their science-related experiences (Brookes & Etkina, 2015). Learning science-specific language is then essential to becoming a member of a science community (Lemke, 1998).

Science language, as language in other domains, can be characterized to be an open dynamical system: it changes over time and it is open to external influences. For example, new concepts and terms are introduced with the advent of advanced theories (Touger, 1991). Advancing theories oftentimes is accompanied by a refinement in concepts and terms used. For example, the Medieval concept of "impetus" that a once resting object carries when thrown is refined with the advent of Newtonian physics. Momentum and kinetic energy replaced this concept entirely (Halloun & Hestenes, 1985). Hence, new terms are used and old terms are abandoned. Moreover, science language is infused with everyday language. Confusion in understanding science contents is related to interferences in language that arises from using science concepts in different domains and everyday language. "Heat" is technically a process variable in science, however, it is oftentimes used as a state function in what is called a caloric metaphor in everyday language (Brookes & Etkina, 2015).

Understanding and meaning making through language is bound by the context that language appears in. The distributional hypothesis states that one understands a word by the company (of words) it keeps (Jurafsky & Martin, 2014; Harris, 1954). And within cognitive semantics, the concept of ancillary knowledge (Redish & Kuo, 2015) states that we understand the meaning of terms by a contextual web of concepts. For example, the concept "current" is understood by it's definitions as a stream of charged particles. However, the definition itself is only understood by the concepts "stream", "charged", and "particles" (Redish & Kuo, 2015). Certain concepts, then, are more central compared to other concepts and can be used as prototypes. Prototype theory posits that a bird such as a "robin" is more representative of the category bird, as, say a penguin (Rosch, 1975). Similarly, in the sciences there can be singled out core concepts that interconnect disciplines and can be hypothesized to be central in science-term networks where usage of terms is linked together. In German state curricula, conepts such as matter, system, energy, and interaction play a central role that underlie contents in physics (KMK, 2020). In the Next Generation Science Standards, some disciplinary core ideas for physics are force and motion, systems, energy, and matter (National Research Council, 2012). These concepts also function as organizing principles for curricula across many countries.

3 Modeling language

It proved intractable to specify all rules that govern language comprehension and production, and hence deterministically model language comprehension and production (Halevy et al., 2009). Information theory and complex systems theory were found to provide powerful frameworks to explain some phenomena related to language use, because of the complexity involved in any language-related phenomena. Human language is optimized to some degree to convey as much information without confusion, hence, certain words occur more frequently in order to enhance processing speed (Montemurro & Zanette, 2010). It is suspected that some form of "principle of least action" explains the complex systems behavior of language, where a vocabulary for efficient communication needs to be found such that few words are used more frequently. Most other words occur only rarely (Zanette, 2014).

One robust finding for language as a complex system is the power law behavior of the word occurence, called Zipf's law (Font-Clos & Corral, 2015; Alstott et al., 2014). Complex systems such as language typically comprise elementary units, called tokens, that can be grouped (by means of similarity) into larger entities, called types Font-Clos and Corral (2015). For language, tokens are the individual instances (realizations) of words and types the abstract entities, i.e., an element in the vocabulary. The frequency of word occurence can be predicted based on this power law behavior. Language dynamics also follow principles derived from evolutionary principles (Lieberman et al., 2007; Nowak et al., 2002). For example, it has been shown for English language that the half-life of an irregular verb scales with the frequency with which it is used (Lieberman et al., 2007). As such, language is characterized by regularities at small and large scale which is important in modeling language-related phenomena.

Discipline-specific language can be expected to adhere to similar regularities. As such, learners in a discipline will be confronted with central terms more often and get acquainted with words by accumulating and integrating the different meanings in different contexts (Lemke, 1998). The theory of lexical concepts and cognitive models advances a usage-based account of meaning making from language, i.e., situated meaning-construction (Evans, 2006). In this context it is suggested that words have meaning potentials that are activated as a function of the context they appear in. Learners are confronted with these different contexts to various degrees (Palmer, 1997). However, it has been argued that learners are confronted with insufficient information to explicitly learn the meaning and rules of words and concepts, which has been called the "poverty of stimulus" (Nowak et al., 2002). It is thus quite perplexing that speakers who grow up in the same speech community reliably speak the same language (Nowak et al., 2002). Language learning is in large part inductive inference (Nowak et al., 2002).

For complex systems such as language, network structures have been found to provide means to model relevant mechanisms such as information flow (Brockmann, 2021). Networks, in its simplest form, are defined by nodes (also: vertices) and edges, which connect the nodes. Networks appear in many complex systems, such as websites and social networks. The diameter of the World-Wide Web was measured by counting the average shortest distance between any two nodes (Albert et al., 1999).

The WWW appears to be only 18.59 in diameter. This means that any website can reach any other website in less than 19 steps. The social graph of Facebook was only 4.74 (Ugander et al., 2011). Moreover, real-world networks were found to follow powerlaws. Similar to Zipf's law, this means for example that many nodes in a network have few connected edges (i.e., low degree) and a non-negligible fraction of a few nodes have a large number of connecting edges (Barabási & Albert, 1999). The few well-connected nodes then dominate and mediate information flow in the networks. Behavior that follows powerlaws is also called scaling behavior, because it did not depend on the magnification with which a system is observed (West, 2017).

In discipline-based educational research network analysis has been used, among others, to analyze immersion of students in communication networks. Researchers found that position within these networks is predictive of students' performance in physics (Brunn & Brewe, 2013; Grunspan et al., 2014). Besides social networks, also the knowledge in a discipline can be represented in the form of networks (Koponen & Pehkonen, 2010). The natural sciences in particular represent disciplines where terms are logically connected and build upon each other. For example, to understand the Newtonian force concept in physics, the concepts of displacement, velocity, and accelerations have to be introduced first. Physics knowledge/curriculum structures are also hierarchical. This knowledge is stored in textbooks and curricula, and more and more in internet databases such as Wikipedia.

With regards to analyses of science language, most studies apply in-depth, qualitative research approaches such as content analysis (Brookes & Etkina, 2015; Carlsen, 2007). These approaches are based on human experts' interpretations of the language data. Even though this assures meaningful analysis of the data, it is difficult to scale this approach to larger amounts of language data in science that will become increasingly available in the future (Baig et al., 2020). Computational approaches could facilitate a more data-centered, bottom-up approach to language analysis in science education research. Natural language processing emerged as a particularly powerful tool for systematically analyzing and modelling language data. Natural language processing encompasses a wide array of tools such as part-of-speech tagging or named entity recognition (Jurafsky & Martin, 2014). All these tools can enhance computational analyses of natural language.

4 Research questions

The present study utilizes natural language processing to facilitate network analysis of science-specific language. We seek to examine linguistic properties and interdependencies of science-related terms. The goal is to identify central terms in science-specific language and examine the properties of relations among the terms through a network analysis approach. The following research questions guide this study:

RQ1: What are typical network parameter values for term networks in the natural science disciplines? In what ways are the networks for biology, chemistry, and physics similar or different?

RQ2: Which terms in biology, chemistry, and physics emerge as central based on their network properties when analyzing a large corpus of science-specific texts, respectively?

RQ3: In what ways do the contexts of the most central terms differ when used in the other considered disciplines?

5 Method

5.1 Science-related Wikipedia articles

For science, Wikipedia [1] has been proven to be a reliable source of knowledge (Giles, 2005; Agrawal et al., 2016; Ponzetto & Strube, 2007), almost as accurate as its commercial competitors (Giles, 2005). Consequently, Wikipedia has been used as a resource to automatically assist teachers in curriculum design (Agrawal et al., 2016) and to enhance natural language processing application such as coreference resolution (Ponzetto & Strube, 2007). Given the validity of science-related Wikipedia articles, we chose this as the text corpus to analyze science-related terms. We were furthermore interested in contrasting German and English science language to better understand how generalizable certain patterns are across these two western languages. Because we were interested in what science terms are central, the articles had to be cleaned to retrieve only plain text articles. Hence, mathematical formula, references, and urls were removed from the articles with the help of natural language processing tools.

Wikipedia articles are annotated in the form that categories are assigned to the articles. Only articles that were labeled as science-related (i.e., biology, chemistry, and physics) were retrieved and respectively analyzed. Table 1 displays information on the retrieved articles for physics. In this study, we will often focus physics language. In the online supplement we include the respective information for biology and chemistry.

As can be seen, German Wikipedia had overall more physics-related articles, however, physics articles in English Wikipedia were longer. Interestingly, for biology and chemistry, English Wikipedia had more artciles that were also longer. Sentence lengths in English articles were longer compared to German articles' sentences for all science disciplines. Given that German language allows for long compound nouns, we can also see that German articles had an overall greater vocabulary compared to English Wikipedia articles for all disciplines (note: for biology, the relative size adjusted by number of articles would be greater as well). The type-token-ratio for all science disciplines was greater in German compared to English. This means that German language uses more specific terms per token.

[1] The entire Wikipedia is publicly available. German Wikipedia: https://dumps.wikimedia.org/dewiki/20211001/ (access November 2021), English Wikipedia: https://dumps.wikimedia.org/enwiki/20211101/ (access December 2021).

Table 1 Characterization of the German and English Wikipedia articles for physics

Language	# Sents	Avg sent length (SD)	Type-token-ratio	Unique words	# Articles
German	119219	18.5 (8.7)	0.08	174969	9679
English	122546	25.5 (13.5)	0.03	105374	6528

5.2 Prelimiary analysis of science-related language in Wikipedia articles

To focus analyses on science-related terms, a subdataset was extracted in which only nouns were kept for the analyses. In physics ontology, typically entities, objects, and concepts are represented by nouns and processes are represented by verbs (Brookes & Etkina, 2015; Lemke, 1998). Hence, to map physics language and physics knowledge-structure, it is sensible (as a first step) to restrict analyses to nouns. Natural language processing techniques of part-of-speech tagging (as implemented in the spaCy-library for Python allows to perform this analysis in many different languages (Honnibal & Montani, 2017).

To linguistically characterize the sample articles and the noun dataset in more detail and analyze potential linguistic differences between German and English Wikipedia, we examined to what extent powerlaw behavior and Zipf's law applied for the articles (Font-Clos & Corral, 2015; Clauset et al., 2009). Powerlaws are of great interest, because the distributions exhibit heavy tails, meaning that all values are expected to occur, allowing for scale-free behavior (Alstott et al., 2014; West, 2017). In the powerlaw $\nu(t) \sim t^{-\alpha}$, t refers to the word rank, $\nu(t)$ to the word count, and α to the powerlaw exponent, necessarily below zero. If the behavior of the system follows Zipf's law, the exponent should be -1 in the abovementioned representation. A slightly different, though more common, representation considers the number of words with the same number of counts ($\nu(t)$). Here, the exponent for the Zipf distribution is expected to be around -2. With maximum likelihood methods viable tools became available to analyze powerlaw behaviors. Researchers developed open source software packages for evaluating empirical data with regards to powerlaw behavior (Alstott et al., 2014). In this software package, the empirical data is mapped to a probability density distribution. A minimum value for the x-axis where the powerlaw behavior typically starts is additionally fit to consider a power law for parts of the distribution.

It is furthermore informative to contrast the powerlaw against other distributions to infer the data generating process. The data generating process for a normal distribution is adding random variables X together. Hence, many observables in the real world follow a normal distribution. For a lognormal distribution, positive random variables are multiplied together. Creating a powerlaw requires more elaborate data generating processes that are oftentimes not well understood (Alstott et al., 2014). However, mere draws from a uniform distribution of characters plus a space can reproduce some of the regularities of language related to the powerlaw behavior (Li, 1992).

Figure 1 displays the empirical and fitted distributions in a log-log-plot. Except the lower tails in the right-hand side plots, the distributions resemble linear curves as expected when powerlaw behavior is present. The exponents are close to 2 for the German articles (see Table 2). For English articles they have a greater variability, however, still close to the exponent of 2 (though not always within the error bounds, σ). Table 2 further indicates that, compared with an exponential distribution (likelihood ratio R, and significance value p), a powerlaw distribution is a statistically significantly better description of the data compared to an exponential law. This holds true for all science disciplines (see online supplement). This comparison shows that the distributions are heavy-tailed (Alstott et al., 2014). However, compared with the lognormal distribution (R_{logn}, p_{logn}), we cannot confirm a statistically significant better fit of the powerlaw distribution in all cases. For example, the negative likelihood ratio for English physics articles indicates a better fit for the lognormal distribution, which is also a heavy-tailed distribution. The equal fit of lognormal and powerlaw seems to be common in empirical data analyses (Alstott et al., 2014). Figure 1 also indicates that the heavy-tail for the noun-only dataset is smaller compared to the entire dataset.

Another important phenomenon related to language is the positive correlation between word length and rank (starting with most frequent terms as 1): longer words

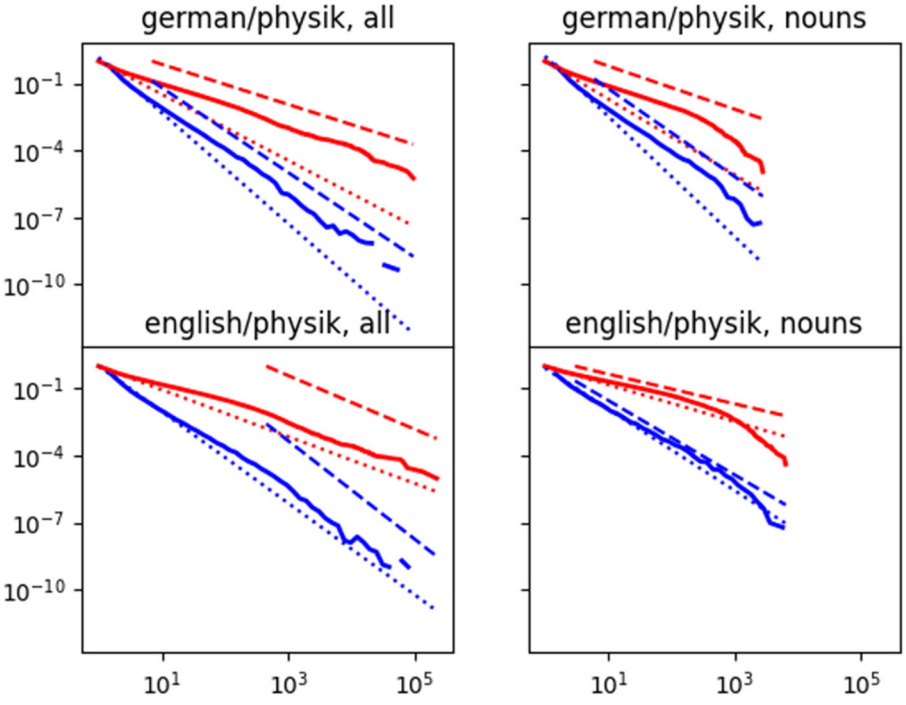

Fig. 1 Graphical representation of Zipf's law. Fitted curves are dashed. Upper left: all articles of German Wikipedia; upper right: all articles of English Wikipedia; lower left: only nouns in German Wikipedia; lower right: only nouns in English Wikipedia. Blue lines: probability density function; red lines: complementary cumulative probability density function

Table 2 Parameters for power law fits and comparison with other distributions

Language	Dataset	$x_{\min}$	α	σ	D	R	p	R_{logn}	p_{logn}
German	All	7	1.90	0.01	0.022	27359	<.001	0	0.67
German	Nouns	6	1.97	0.01	0.025	6909	<.001	0	0.6
English	All	453	2.20	0.04	0.023	701	<.001	0	0.15
English	Nouns	3	1.67	0.01	0.039	11088	<.001	−10	0.0028

tend to be less frequent (Zanette, 2014). We examined this relationship in the present datasets. For German, Pearson correlations were .23 and .23 for all words and only nouns, respectively. For English, correlations were .21 and .18 for all words and only nouns, respectively. Corellations for other disciplines were: .25,.26,.20,.19 (biology) and .25,.25,.23,.19 (chemistry). Hence, we can confirm that science-related language also exhibits this general relationship in languages.

6 Parameters of science-term networks

In RQ1 typical parameters of the different science-term networks for the disciplines will be displayed and compared. First, the different frequencies of nouns and the overall vocabulary size, i.e., number of nodes in the network, will be displayed. Alongside the number of nodes we will also display the number of edges between the nodes. A common property of networks is the density (Grunspan et al., 2014). The density is calculated as the proportion of realized edges to the number of possible edges. Another interesting property of the networks is the average shortest paths. We indicated that for social network graphs this distance is typically rather low, which refers to the property that with only few steps from any node (here: person) any other node can be reached. Finally, the transitivity will be calculated. Transitivity is a measure of cohesion (Grunspan et al., 2014). It measures the number of realized triads in relation to the number of possible triads in the network.

Another important phenomenon for real-world networks is the scale-free behavior of node degrees (Brockmann, 2021). As with the linguistic properties analysis above (Zipf's law), the frequency of node degrees can be similarly described with a powerlaw. We will use a similar analysis as outlined above. The frequency of nodes for a certain degree will be plotted as histrograms and probability density distribtuions with their respective fitted curves. The analysis of the scaling parameter α and the comparison with exponential and lognormal distribution will reveal properties of the data distribution.

7 Identification of central terms through network analysis

The more frequent terms in a language are crucial for processes of language acquisition, language perception, and production. An intuitive way to analyze frequent terms

in physics-specific languages would be to count occurences of types in the Wikipedia articles. However, most frequent words in English such as "the" would be uninformative. Hence, only nouns were considered. The nouns were split by their linguistic function as being a subject or object, given that this adds information to what extent a term is used as agent (subject) in a sentence. The Python library spaCy was used to generate these datasets.

The structure of sentences in English and German is organized in phrases, where the subject (in a noun phrase) determines the agent in a sentence which is related to objects via verbs. In our analysis of central terms, we therefore extracted every subject for the respective Wikipedia articles and linked them to their nouns in a sentence. To create a network representation, every subject and object was stored as a node in the network. Extracting subjects and objects was again performed with the spaCy library in Python. Each link between a subject and object was stored as an edge. A network representation can then be generated through complex modeling. For example, the spring layout utilizes the concept from physics where each edge is a spring and an equilibrium distribution is to be found through optimization techniques.

Retrieving central nodes, i.e., central terms, can be done in multiple ways. A simple approach is counting the incoming and outgoing edges. However, PageRank algorithm has been found to be more performant to detect important nodes. Based on the observation that a node with fewer links from otherwise more important other nodes should be ranked higher than a node with many links from irrelevant nodes (Page et al., 1999). Hence, we will use the Python library networkx's implementation of PageRank to identify central terms in science language (RQ2).

In RQ3 we will use the most central terms in the physics articles and analyze how they are used in the different disciplines. We compared the use of physics terms in the disciplines biology, chemistry, and politics as contrasting cases. Given that English and German analyses yield similar results with regard to central terms, we will focus the English articles in this analysis. For the new disciplines, we also retrieved all articles from Wikipedia and the respective noun dataset. We then analyzed the links that each term from the physics terms had with other terms in the respective discipline. This analysis will eventually yield differences in contexts in which terms are used in the disciplines.

8 Findings

8.1 Parameters of the science-term networks

In RQ1 we calculated important parameters of the respective science-term networks (see Table 3). As a baseline comparison, we depict the number of nodes in the networks. It can be seen that the German networks have more nodes compared to the English networks. However, the English networks have more edges (i.e., links between the nodes) compared to the German networks. Hence, the density for the English networks is more than twice the density for the German physics network. In all networks, the average shortest path length is little above 3 for English and around 4 for German language. This means that any term can on average be reached from any

Table 3 Summary of network parameters for the different groups

Group	Number nodes	Number edges	Density	Avg. shortest path	Transitivity
English, Biology	31593	405123	0.0008	3.21	0.12
English, Chemistry	21942	226068	0.0009	3.27	0.12
English, Physics	24610	368815	0.0012	3.15	0.16
German, Biologie	53868	205572	0.0001	3.98	0.04
German, Chemie	53226	169642	0.0001	4.09	0.03
German, Physik	90868	426893	0.0001	3.83	0.05

other term with only few steps. The transitivity (cohesion) in the English networks is greater compared to the German networks. This is likely related to the greater density for the English networks that indicate that everything is closer tied together. Overall, the within-group differences in a language seem to be smaller compared to the within-group differences in a domain.

The node degrees of these real-world term networks for science discriplines show scale-free behavior, i.e., they follow a powerlaw distribution (see Table 4). The scaling parameter α is around 2 for all networks. The R and p values indicate that the data is better approximated by a powerlaw distribution as compared to an exponential distribution. Hence, the upper tail is populated with nodes that have many connections (i.e., high degree). The comparison with the lognormal distribution (R_{logn}, p_{logn}) is less conclusive: sometimes the lognormal is a better fit, and sometimes the powerlaw. However, both distributions (powerlaw and lognormal) have a heavy tail.

The powerlaw behavior can be observed in the distributions as well (see Fig. 2). The histograms indicate that many nodes have a low degree. Following powerlaw scaling, few nodes have large degrees. This indicates that few terms are central in the networks and function as hubs that connect the different regions of the networks.

8.2 Central terms in networks

In RQ2 we evaluate what science terms emerge as central from analyzing filtered German and English Wikipedia. The resulting network for German physics-related

Table 4 Powerlaw parameters for the science-term networks

Group	$x_{\min}$	α	σ	D	R	p	R_{logn}	p_{logn}
English/Biology	5	1.70	0.01	0.036	7705	<.001	−58	6.7e−13
English/Chemistry	5	1.78	0.01	0.038	5431	<.001	−38	9.6e−09
English/Physics	5	2.05	0.01	0.019	17291	<.001	−1	0.24
German/Biologie	7	2.10	0.01	0.025	4601	<.001	0	0.62
German/Chemie	6	1.75	0.01	0.040	7660	<.001	−71	8.8e−15
German/Physik	8	2.17	0.01	0.025	3082	<.001	0	0.74

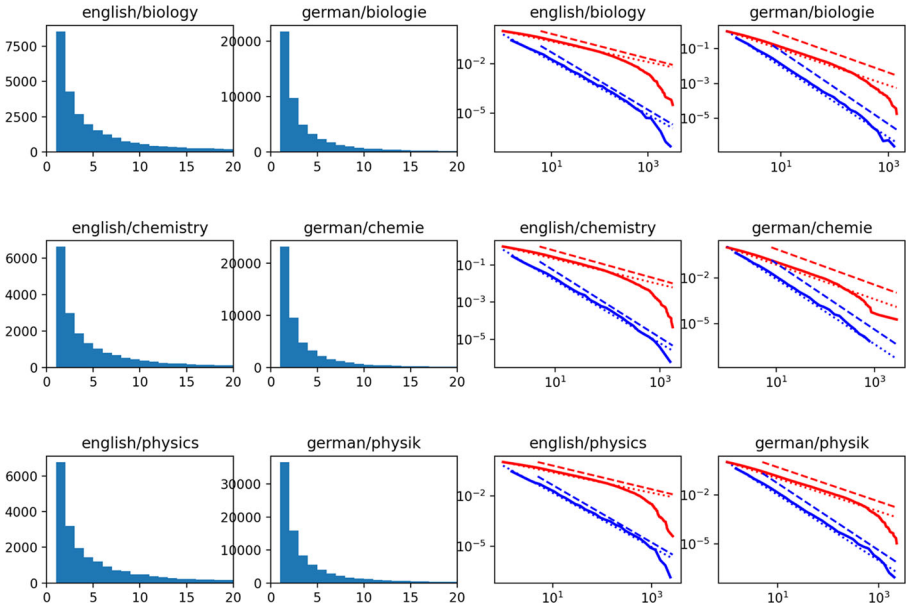

Fig. 2 Powerlaw behavior of the science-term networks. Histograms on the left represent the frequency over number of degrees for the nodes in the respective networks. Probability density distributions on the right represent the probability density (red: cumulated, blue: probability; solid: empirical, dotted: $x_{\min}$ fixed to 1, dashed: $x_{\min}$ free to vary)

Wikipedia articles can be seen in Fig. 3.[2] Only the most connected nodes are represented to make the network readable. The highlighted edges represent the strongest links between any two terms. The 20 most central terms are: Energie, Teilchen, Zeit, Physik, Theorie, Begriff, Arbeiten, Form, Eigenschaften, Teil, Elektronen, Arbeit, Entwicklung, Masse, System, Bereich, Körper, Materie, Größe, Beispiel[3].

The terms refer to sociology/philosophy of science (Theorie [theory], Begriff [term], Untersuchungen [investigation], Wissenschaftler [scientist][4]) and to discipline-specific concepts (Energie [energy], Teilchen [particle], Zeit [time], Elektronen [electron], Kraft [force]). We will focus on the discipline-specific terms. Energy emerged as the most central term. It is linked to many other terms. For example, energy is linked to Teilchen (particle), Form (form), System (system), Masse (mass). This is well in line with our expectations. Forms (of energy) are a common approach to teaching energy. Furthermore, discussing particles (Teilchen) oftentimes involves energy. In elementary particle physics, energy is an important concept

[2]The network graphs for biology and chemistry can be found in the onlines supplement. Please find English translations of the German term networks in parentheses within the network figures. Translations are automated through the `googletrans`-Python library.

[3]English: energy, particle, time, physics, theory, concept, works, form, attributes, part, electrons, work, development, mass, system, domain, body, matter, size, example.

[4]Note that there appears a gender bias in German language, because only the male form of scientist is represented.

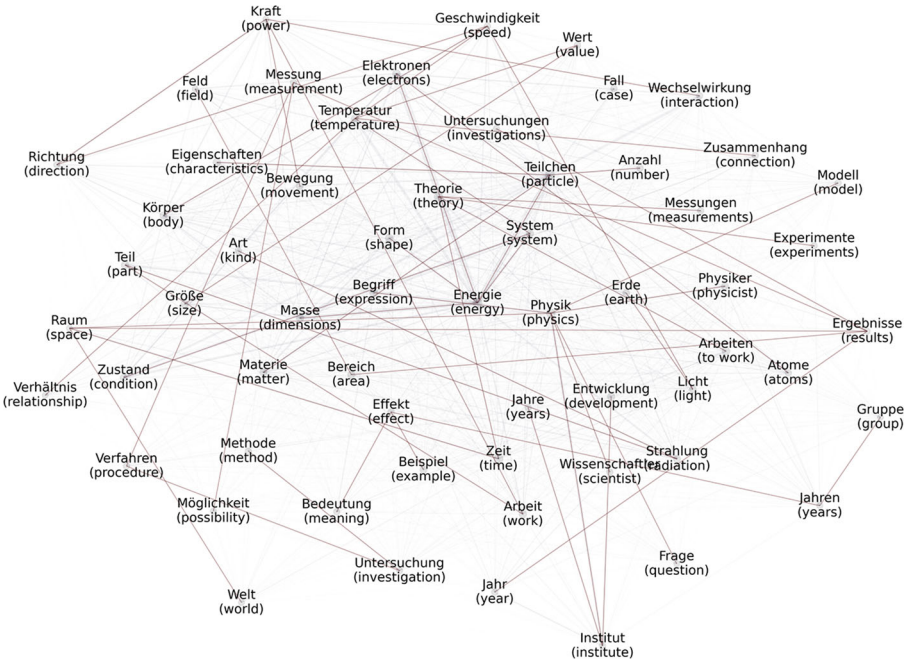

Fig. 3 Network representation of central physics terms in German Wikipedia

(alongside with momentum) to analyze experiments and detect potential new particles. Energy is also linked to system and system is then linked to state. These links express the importance of system identification when dealing with energy. Furthermore, linking it to state suggests that energy is a state function that is independent of trajectories (Brookes & Etkina, 2015).

Particle then is linked to matter. This link expresses the fact that matter is made up of particles. As an example of a particle, electron is linked to particle. It is noteworthy that electron emerged as a specific particle. The electron is a well-studied object in physics. For once, eletromagnetism in large parts is concerned with the behavior of electrons as a negatively-charged elementary particle (fermion). Furthermore, the study of the electron in the early days of quantum mechanics (e.g., the Dirac equation) and the observation that electrons as particles exhibit wave-like properties were crucial to advance physics. Electron is then linked to atoms, because atoms comprise electrons. Understanding the behavior of electrons in atoms enables the prediction of properties of molecule formation and chemical reactions.

Temperature is linked to velocity. This attributes to the fact that temperature is defined by the average velocity of microscopic particles in a system. Space and time are also linked. This might be attributed to the strong conncetion of these two concepts in the realm of relativity theory.

For English Wikipedia, the 20 most central physics terms were: theory, time, energy, system, example, number, field, particles, effect, work, model, process, physics, result, state, equation, method, experiment, form, part. As is evident from

this list, many terms in English Wikipedia can be mapped to German Wikipedia. Also in English Wikipedia energy emerges as the most central physics concept (see Fig. 4). Energy is linked to system. Space and time are also linked in English Wikipedia as in the German Wikipedia.

However, many other connections also emerge in the English Wikipedia. For example, equation is highly interlinked with other terms. Laws are commonly linked to equations as for example Ohm's law, or Newton's second law that can be encapsulated in an equation. Terms is also linked to equations, which refers to the mathematical constituents of an equation. Furthermore, field and density are linked to equation. Field equations are common in gravitational theory, electromagnetism, and quantum field theory, among others. Density could be linked to equation via the continuity equation.

Finally, time is well interlinked in the English Wikipedia network. Besides space, time is also connected to process. This attributes to the fact that processes are inherently time-bound. Time is also linked to equation. This might be attributed to the fact that dynamical equations in physics model processes in time, i.e., time is a parameter in these equations. The link between time and concepts could indicate that time itself is oftentimes introduced as a concept. In general relativity theory, time is exposed as fundamentally bound to space. Recent quantum theories such as relational quantum mechanics and predecessors treat time as a fundamental ingredient in physics theories.

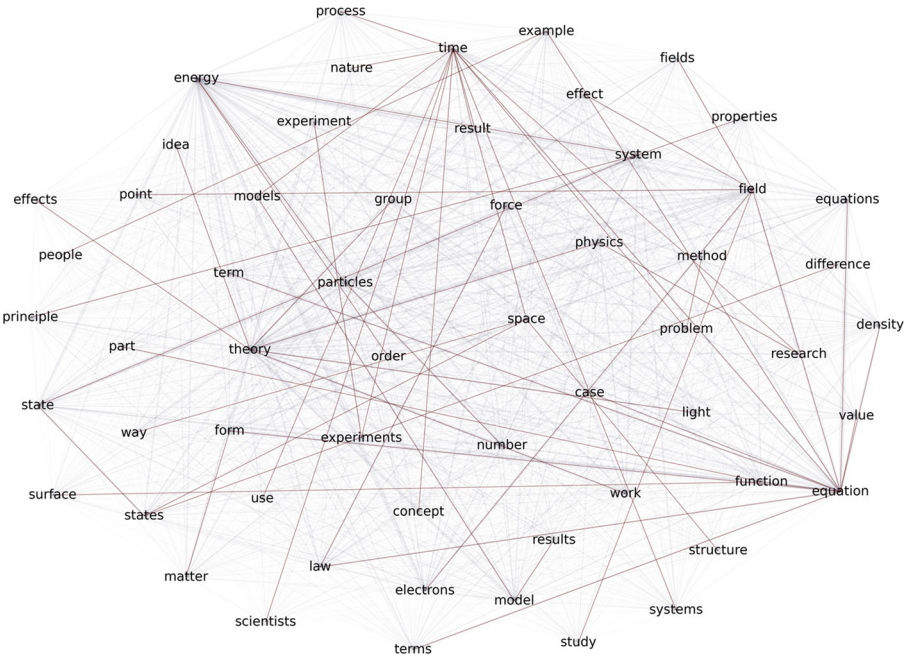

Fig. 4 Network representation of central physics concepts in English Wikipedia

8.3 Contrasting term use by discipline

In RQ3 we now use the 20 most central terms (extracted by PageRank) in physics English Wikipedia and determine in what contexts they appear in other disciplines than physics. The disciplines biology, chemistry, and politics were considered, because biology, and chemistry are closely related to physics and politics is rather different in terms of concepts and used language. Table 5 shows the counts of each physics term in the other domains. It is noteworthy that all physics terms could be found in all other disciplines. The counts varied. For example, "number" was the most frequently encountered physics term in politics articles, whereas in biology and in chemistry "process" was the most frequent physics-derived term. The less used terms were particles, physics, and physics for politics, biology, and chemistry respectively.

To analyze differences in contexts, the three most connected nouns for each physics term were determined (see politics/biology/chemistry 1/2/3 in Table 5). The term "system" in politics is linked to representation, government, and method. This relates to a political system and its function of representation. In biology, "system" is connected to cells, species, and systems. The cell is described as a structurally separable, autonomous, and self-sustaining system. Hence its close relation to system. Finally, in chemistry "system" is connected to equilibrium, state, and example. Equilibrium chemistry is concerned with systems where involved chemical entities do not change with time. This is typically an important assumption in order to model processes and phenomena mathematically. Even though the underlying concepts of system in all inspected disciplines share commonalities (e.g., a whole comprised of parts), the connected words are entirely different. Similarly, the term theory in politics refers to government, in biology to evolution, and in chemistry to orbitals (atoms). This indicates that language learners in different disciplines get acquainted with different contexts for the same term.

9 Discussion

This study sought to analyze science-specific language in a principled way with the help of natural language processing and network analysis methods. To retrieve a representative body of science-related language, Wikipedia articles were analyzed that were categorized with the labels biology, chemistry, and physics. The respective German and English versions of Wikipedia were analyzed as contrasting cases. German Wikipedia had more articles overall in physics, however, in English Wikipedia the articles were longer. In terms of linguistic properties, both languages followed Zipf-law behavior. This means that few terms appear very often. This is a well documented phenomenon for languages (Moreno-Sánchez et al., 2016).

In RQ1 we analyzed the general properties of the different science-term networks. English science-term networks appear to have fewer nodes compared to the German networks, however English networks have more connections between the nodes, hence, they are denser compared to the German networks. This raises the cohesion of the English science-term networks. The average shortest path between nodes in English and German science-term networks is approximately 3 and 4, respectively.

Table 5 Word contexts of different disciplines with regards to the most central terms in physics

Concept	# Politics	Politics 1	Politics 2	Politics 3	# Biology	Biology 1	Biology 2	Biology 3	# Chemistry	Chemistry 1	Chemistry 2	Chemistry 3
Particles	2	Part (1)	Model (1)	Inheritance (1)	90	Molecules (7)	Size (6)	Air (6)	246	Time (19)	Size (17)	Surface (13)
Physics	3	Industry (2)	System (1)	Means (1)	3	Thermo-dynamics (1)	Surface (1)	Success (1)	16	Chemistry (8)	Energy (3)	Approach (2)
Equation	6	Antisemitism (2)	Capital (2)	Time (1)	89	Fraction (14)	Life (14)	Planets (12)	289	Reaction (38)	Solution (22)	Energy (20)
Experiment	25	People (3)	Absence (3)	Democracy (3)	181	Moths (8)	Acids (8)	Protein (8)	102	Molecules (7)	Compounds (7)	Acids (7)
Energy	27	Fuels (3)	Use (3)	Coal (2)	68	Bonds (5)	Membrane (4)	Water (4)	304	Molecule (28)	State (28)	System (25)
Field	63	Candidates (6)	Study (5)	Number (4)	155	Biology (22)	File (10)	Development (10)	139	Chemistry (19)	Direction (7)	Research (7)
Method	147	Seats (10)	Form (9)	Representation (8)	541	Cells (38)	Sequences (25)	Methods (23)	562	Use (25)	Process (21)	Molecules (20)
Form	209	Government (33)	Democracy (13)	People (10)	196	Species (12)	Cells (8)	Traits (7)	147	Acid (12)	Molecule (9)	Conditions (8)
Model	235	Capitalism (11)	Individuals (11)	Form (10)	394	Species (23)	Evolution (11)	Time (11)	242	Theory (15)	Molecules (14)	Electrons (13)
Effect	264	Electorates (15)	People (11)	Election (11)	301	Species (18)	Cells (18)	Population (13)	315	Reaction (13)	Molecules (13)	Increase (11)
Example	331	Description (18)	Government (11)	State (11)	515	Species (47)	Evolution (22)	Cells (21)	479	Reaction (44)	Water (34)	Acid (27)
Time	341	Election (21)	Years (13)	War (9)	100	Size (7)	Days (5)	Field (5)	84	Years (7)	Number (5)	Pressure (4)

Table 5 (continued)

Concept	# Politics	Politics 1	Politics 2	Politics 3	# Biology	Biology 1	Biology 2	Biology 3	# Chemistry	Chemistry 1	Chemistry 2	Chemistry 3
Part	439	County (32)	Area (24)	Line (18)	141	Structure (11)	Body (11)	Sequence (10)	65	Coefficient (5)	Particle (4)	Term (4)
Theory	496	Government (18)	Society (18)	Power (16)	588	Evolution (99)	Selection (44)	Organisms (33)	260	Orbitals (21)	Atoms (18)	Chemistry (15)
Process	688	Place (27)	Government (23)	Law (23)	682	Cells (61)	Cell (33)	Time (31)	597	Reaction (38)	Energy (32)	Place (20)
Work	756	History (18)	Time (18)	Women (17)	502	Evolution (23)	Development (22)	Theory (17)	364	Chemistry (38)	Development (27)	Compounds (22
Result	815	Majority (128)	Seats (113)	Election (108)	189	Discussion (15)	Number (11)	Molecule (9)	149	Equation (10)	Solution (8)	Number (7)
System	1263	Represen-tation (79)	Government (63)	Method (58)	614	Cells (41)	Species (24)	Systems (21)	399	Equilibrium (59)	State (34)	Example (19)
State	1835	States (96)	Law (78)	Government (68)	67	States (7)	Nature (4)	Compounds (4)	231	Energy (29)	Reaction (19)	Electrons (17)
Number	2489	Members (275)	Votes (184)	People (170)	583	Species (75)	Cells (31)	Individuals (23)	426	Atoms (42)	Electrons (35)	Components (34)

This number is well in line with other networks such as social networks, e.g., Facebook has an average shortest path of around 5 (Brockmann, 2021). The distribution of node degrees was shown to follow a powerlaw distribution. This indicates that the science-term networks follow scale-free behavior and have heavy tails (Barabási & Albert, 1999). This means that few terms form central hubs in these networks and dominate the information flow. Arguably, these terms have to be specifically accounted for in educational efforts and curricula.

In the context of RQ2 we sought to identify these central terms. To do so, the subjects in each sentence were linked to their respective objects, similar to network analyses where people are linked to other people or websites are linked in search engines. The PageRank algorithm was used to extract the most important terms in German and English Wikipedia. The most central terms in German and English are almost identical for physics, and also for biology and chemistry, respectively. The terms were also linked to other terms in similar ways in both languages. For physics in particular, the analyzed terms referred to philosophy of science/sociology and physics-specific concepts. The physics-specific terms particulary match expectations as expressed in physics state curricula. It is interesting to note that no domains in physics appear as central terms, e.g., thermodynamics, mechanics, optics, etc. We hypothesize that probably the core concepts (interaction, system, energy, force) are more important also across domains.

In RQ3 we applied the most common terms which were found in physics to other domains in order to examine differences in contexts with reference to the disciplines. We found that all terms were also found in the other domains (politics, biology, chemistry). The contexts of the terms (i.e., the words with which they appear in a sentence), however, varied considerably across the disciplines and matched expected terms in the respective disciplines. For example, the term "theory" in biology was related to evolution, whereas in chemistry it was related to orbitals and atoms. This finding hints to the challenges of meaning-making in different disciplines: The same terms are used in different contexts with similar, yet non-identical meaning. This creates the cognitive challenge for learners to always consider the context when encountering a specific term. In fact, effects of framing and context for reasoning have been investigated in science education research (Palmer, 1997).

10 Limitations

Our study has several limitations that limit generalizability of our findings. We submit that oftentimes concepts are captured in a term, however, there are important concepts that are represented as more than one word. For example, in German, "Freier Fall" (free fall) refers to free fall of a moving body. Free fall is a specific physics concept, where assumptions such as no friction are made. These cases are not necessarily captured in our approach. Identification of bigrams or including the entire noun chunk of the subject can enhance this analysis. We also could not verify if there are representational biases in the Wikipedia articles. The marked differences between English and German Wikipedia articles might be attributable to the fact that German

language more uniquely captures concepts in compound nouns which account for greater number of nodes and, potentially, less dense networks.

Our analyses focused on linguistic properties of terms in science language. However, cognition, expertise, and learning in science is in large part visual. Allegedly, Kekulé discovered the structure of benzene by visual means. Forming a visual representation of a science problem is an indicator of science expertise (Singh, 2008). Hence, the integration of visual and language-related modes of representation would be crucial to fully understand development of expertise in science. This interconnection of visual experience and language development in forming physics intuitions and conceptions is an important future direction for science education research.

Finally, we can only hypothesize that this scale-free behavior and the central terms will emerge similarly in non-western languages. We restricted our analyses to two western languages with well-developed and curated Wikipedia databases for the natural sciences. However, these languages cover only a fraction of the existing languages and speakers. It would be certainly worthwhile to replicate these analyses in other languages. All described libraries and computational tools will allow for these kinds of analyses in other non-western languages.

11 Conclusions

Applications of network analyses have been found useful in many domains such as education where large unstructured datasets could be systematically analyzed to identify active learning and knowledge acquisition (Grunspan et al., 2014; Brunn & Brewe, 2013). With the help of natural language processing, in particular, these network analyses can be extended to systematically analyze large corpora of language data. Our approach in particular showed how term-based networks can be extracted from large text corpora. Thus, we provide a template for knowledge-oriented networks that are expected to enhance social network analysis (Brass, 2022). Moreover, social network analysis in conjunction with language networks capture essential aspects of Vygotsky's sociocultural theory of cognitive development insofar as they enable to account for the complex interdependencies of socially-mediated knowledge and language acquisition. The language networks particularly capture aspects of the interrelated knowledge in science. This can enhance curriculum design (Agrawal et al., 2016) and diagnostics of beliefs, values, and knowledge (Wulff et al., 2022).

Our analyses also indicate that complex and dynamic systems analyses can play an important role in the methods portfolio of education researchers (Brass, 2022; Hilpert & Marchand, 2018). Educational research related to learning and language can focus on micro, meso, or macro processes, e.g., individual learning, group learning in classes, or learning as a cultural phenomenon on the societal level. In any layer, learning appears to be complex and even interrelated to the other layers. E.g., learning on the individual level is impacted by societal discourses, but also individual cognitions such as beliefs and values. It is intricately difficult to extract relationships, laws, and even theories under these circumstances (Halevy et al., 2009). However, complex and dynamic systems analyses provide a means to extract underlying relationships and laws. This has been documented extensively for complex systems in the natural world. Crickets, birds, cells, or even human crowds were shown to behave

like complex systems (Strogatz, 2003). Even though their behavior appears to be chaotic, complexity science revealed that simple laws govern these natural systems that give rise to complex behaviors (West, 2017; Strogatz, 2003; Brockmann, 2021; Wolfram, 2002). Our analyses showed that simple laws such as the powerlaw underlie the science-term networks extracted from Wikipedia. These laws allow educational researchers to delineate normative distributions and patterns of language use. In some way or another, experts' language can be hypothesized to approximate these normative language distributions of terms, given the exposure of experts to science-related language over extended periods of time (Ericsson, 1998). Complex systems analyses can facilitate detection of outliers, i.e., novices who potentially lack important terms in the distributions. Such linguistic modules could be implemented in intelligent tutoring systems (Graesser et al., 2004), that are language-bound in large part and need any form of evidence to diagnose competences of the tutees.

Abbreviations STEM, science, technology, engineering, and mathematics.

Acknowledgements We acknowledge language editing by [blinded].

Author Contributions Not applicable

Funding Open Access funding enabled and organized by Projekt DEAL. The work is funded by the open access fonds of Heidelberg University of Education.

Availability of data and materials Data is freely available. See references in text. Querys regarding code should be directed to the corresponding author.

Declarations

Ethics approval and consent to participate Not applicable

Consent for Publication Not applicable

Competing interests The authors declare that they have no competing interests.

References

Agrawal, R., Papalexakis, E., & Golshan, B (2016). Toward data-driven design of educational courses: a feasibility study. *Journal of Education Data Mining*, *8*(1), 1–21.

Albert, R., Jeong, H., & Barabási, A. -L. (1999). Diameter of the world-wide web. *Nature*, *401*(6749), 130–131.

Alstott, J., Bullmore, E., & Plenz, D. (2014). Powerlaw: a python package for analysis of heavy-tailed distributions. *PloS one*, *9*(1), 1–11. ISSN 1932-6203.

Baig, M. I., Shuib, L., & Yadegaridehkordi, E. (2020). Big data in education: a state of the art, limitations, and future research directions. *International Journal of Educational Technology in Higher Education 17*,(1).

Barabási, A.-L., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512.

Brass, D. J. (2022). New developments in social network analysis. *Annual Review of Organizational Psychology and Organizational Behavior*, *9*(1), 225–246. ISSN 2327-0608.

Brockmann, D. (2021). Im Wald vor lauter Bäumen:Unsere komplexe Welt besser verstehen. dtv, München. ISBN 9783423282994. https://www.perlentaucher.de/buch/dirk-brockmann/im-wald-vor-lauter-baeumen.html..

Brookes, D. T., & Etkina, E. (2009). "Force," ontology, and language. *Physical Review Special Topics - Physics Education Research*, *5*(1), 643. ISSN 1554-9178.

Brookes, D. T., & Etkina, E. (2015). The importance of language in students' reasoning about heat in thermodynamic processes. *International Journal of Science Education*, *37*(5-6), 759–779.

Brunn, J., & Brewe, E. (2013). Talking and learning physics: predicting future grades from network measures and force concept inventory pretest scores. *Physical Review Special Topics Physics Education Research, 9*(020109).

Carlsen, W. S. (2007). Language and science learning. In Sandra K. Abell, & N. Lederman (Eds.) *Handbook of research on science education, Lawrence Erlbaum Associates Publishers, Mawhah, New Jersey*.

Clauset, A., Shalizi, C. R., & Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, *51*(4), 661–703. ISSN 0036-1445.

de Beule, J. (2008). Compositionality hierarchy and recursion in language: a case study in fluid construction grammar: dissertation.

Ericsson, A. K. (1998). The scientific study of expert levels of performance:general implications for optimal learning and creativity 1. *High Ability Studies*, *9*(1), 75–100.

Evans, V. (2006). Lexical concepts, cognitive models and meaning-construction. *Cognitive Linguistics 17*(4). ISSN 0936-5907.

Font-Clos, F., & Corral, Á. (2015). Log-log convexity of type-token growth in zipf's systems. *Physical Review Letters*, *114*(23), 238701.

Giere, R. N. (2004). How models are used to represent reality. *Philosophy of Science*, *71*(5), 742–752.

Giles, J. (2005). Internet encyclopaedias go head to head. *Nature*, *438*(7070), 900–901.

Graesser, A. C., McNamara, D., Louwerse, M. M., & Cai, Z. (2004). Coh-metrix: analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers*, *36*(2), 193–202.

Grunspan, D. Z., Wiggins, B. L., & Goodreau, S. M. (2014). Understanding classrooms through social network analysis: a primer for social network analysis in education research. *CBE Life Sciences Education*, *13*(2), 167–179. ISSN 1931-7913.

Halevy, A., Norvig, P., & Pereira, F. (2009). The unreasonable effectiveness of data. IEEE Intelligent Systems, 8–12.

Halliday, M. A. K., & Matthiessen, C. M. I. M. (2007). An introduction to functional grammar. Hodder Education. London, 3. ed. [nachdr.] edition.

Halloun, I. A., & Hestenes, D. (1985). The initial knowledge state of college physics students. *American Journal of Physics*, *53*(11), 1043–1055.

Harris, Z. S. (1954). Distributional structure. *WORD*, *10*(2-3), 146–162.

Hilpert, J. C., & Marchand, G. C. (2018). Complex systems research in educational psychology: aligning theory and method. *Educational Psychologist*, *53*(3), 185–202. ISSN0046-1520.

Honnibal, M., & Montani, I. (2017). spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Jurafsky, D., & Martin, J. H. (2014). Speech and language processing. Always learning. Pearson Education, Harlow, 2. ed., pearson new internat. ed. edition, ISBN 9781292025438.

KMK (2020). Bildungsstandards im fach physik für die allgemeine hochschulreife: Beschluss der kultusministerkonferenz vom 18.06.2020. https://www.kmk.org/fileadmin/Dateien/veroeffentlichungen_beschluesse/2020/2020_06_18-BildungsstandardsAHR_Physik.pdf.

Koponen, I. T., & Pehkonen, M. (2010). Coherent knowledge structures of physics represented as concept networks in teacher education. *Science & Education*, *19*(3), 259–282. ISSN 0926-7220.

Lemke, J. L. (1998). Teaching all the languages of science: words, symbols, images, and actions: La caixa conference on science education. https://www.researchgate.net/publication/270904608_Teaching_All_the_Languages_of_Science_Words_Symbols_Images_and_Actions.

Li, W. (1992). Random texts exhibit zipf's-law-like word frequency distribution. *IEEE Transactions on Information Theory*, *38*(6), 1842–1845. ISSN 00189448.
Lieberman, E., Michel, J.-B., Jackson, J., Tang, T., & Nowak, M. A. (2007). Quantifying the evolutionary dynamics of language. *Nature*, *449*, 713–716.
Montemurro, M. A., & Zanette, D. (2010). Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, *13*(02), 135–153. ISSN 0219-5259.
Moreno-Sánchez, I., Font-Clos, F., & Corral, Á. (2016). Large-scale analysis of zipf's law in english texts. *PloS one*, *11*(1), e0147073. ISSN 1932-6203.
National Research Council. (2012). *A framework for K-12 science education: practices, crosscutting concepts, and core ideas*. Washington: The National Academies Press. https://doi.org/10.17226/13165.
Norris, S. P., & Phillips, L. M. (2003). How literacy in its fundamental sense is central to scientific literacy. *Science Education*, *87*(2), 224–240. ISSN 00368326.
Nousiainen, M., & Koponen, I. T. (2012). Concept maps representing knowledge of physics: connecting structure and content in the context of electricity and magnetism. *Nordic Studies in Science Education*, *6*(2), 155–172. ISSN 1504-4556.
Nowak, M. A., Komarova, N. L., & Niyogi, P. (2002). Computational and evolutionary aspects of language. *Nature*, *417*(6889), 611–617.
Page, L., Brin, S., Motwani, R., & Winograd, T. (1999). The pagerank citation ranking: Bringing order to the web. http://dbpubs.stanford.edu:8090/pub/showDoc.Fulltext?lang=en&doc=1999-66&format=pdf.
Palmer, D. (1997). The effect of context on students' reasoning about forces. *International Journal of Science Education*, *19*(6), 681–696.
Ponzetto, S. P., & Strube, M. (2007). Knowledge derived from wikipedia for computing semantic relatedness. *Journal of Artificial Intelligence Research*, *30*, 181–212.
Redish, E. F., & Kuo, E. (2015). Language of physics, language of math: disciplinary culture and dynamic epistemology. *Science & Education*, *24*(5-6), 561–590. ISSN 0926-7220.
Rosch, E. (1975). Cognitive reference points. *Cognitive Psychology*, *7*, 532–547.
Singh, C. (2008). Assessing student expertise in introductory physics with isomorphic problems. i. performance on nonintuitive problem pair from introductory physics. *Physical Review Special Topics - Physics Education Research*, *4*(1), 191. ISSN 1554-9178.
Strogatz, S. H. (2003). *SYNC: the emerging science of spontaneous order: how order emerges from chaos in the universe, nature, and daily life*. New York: Thesia.
Touger, J. S. (1991). When words fail us. *The Physics Teacher, 29*(90).
Ugander, J., Karrer, B., Backstrom, L., & Marlow, C. (2011). The anatomy of the facebook social graph. *arXiv*.
von Weizsäcker, C. F. (2004). Der begriffliche Aufbau der theoretischen Physik: Vorlesung gehalten in Göttingen im Sommer 1948. Hirzel, Stuttgart and Leipzig, 1. aufl. edition, ISBN 9783777612560.
Vygotsky, L. (1963). *Thought and language*. Cambridge, MA: Press, MIT.
Vygotsky, L. S. (1978). *Mind in society: the development of higher psychological processes*. Cambridge, MA: Harvard University Press.
Wallace, C. S. (1996). Mml inference of predictive trees, graphs, and nets. In *Alexander Gammerman, editor, computational learning and probabilistic reasoning,* pp. 43–66*. John Wiley & Sons, Chichester and New York*.
West, G. (2017). Scale: The universal laws of growth, innovation, sustainability, and the pace of life in organisms, cities, economies, and companies. Penguin Press, New York, ISBN 1594205582. http://www.penguin.com/book/scale-by-geoffrey-west/9780143110903.
Wolfram, S. (2002). A new kind of science. Wolfram Media, Champaign, Ill., 1. edition, ISBN 1579550088.
Wulff, P., Buschhüter, D., Westphal, A., Mientus, L., Nowak, A., & Borowski, A. (2022). Bridging the gap between qualitative and quantitative assessment in science education research with machine learning — a case for pretrained language models-based clustering. *Journal of Science Education and Technology*. https://doi.org/10.1007/s10956-022-09969-w.
Zanette, D. (2014). Statistical pattern in written language. arXiv:1412.3336.