Development of an Annotation Schema for the Identification of Semantic Uncertainty in DIN Standards

Jörn Stegmeier[1(✉)], Jakob Hartig[2], Michaela Leštáková[2], Kevin Logan[2], Sabine Bartsch[1], Andrea Rapp[1], and Peter F. Pelz[2]

[1] Institute of Linguistics and Literary Studies, Technische Universität Darmstadt, Dolivostraße 15, 64293 Darmstadt, Germany
joern.stegmeier@tu-darmstadt.de

[2] Chair of Fluid Systems, Technische Universität Darmstadt, Otto-Berndt-Straße 2, 64287 Darmstadt, Germany

Abstract. This paper presents the results of a pilot study carried out in cooperation between Linguistics and Mechanical Engineering, funded by the collaborative research centre (CRC) 805 "Beherrschung von Unsicherheit in lasttragenden Systemen des Maschinenbaus". Our goal is to help improve norm compliant product development and engineering design by focusing on ambiguous language use in norm texts (= "semantic uncertainty"). Depending on the country and product under development, industry standards may be legally binding. Thus, standards play a vital role in reducing uncertainty for manufacturers and engineers by providing requirements for product development and engineering design. However, uncertainty is introduced by the standards themselves in various forms, the most notable of which are the use of underspecified concepts, modal verbs like *should*, and references to texts which contain semantically uncertain parts. If conformity to standards is to be ensured, the person using the standards must interpret them and document the interpretation. In order to support users in these tasks, we

1. developed an annotation schema which allows the identification and classification of semantically uncertain segments of standards,
2. used the schema to create a taxonomy of semantic uncertainty in standards,
3. developed a proof-of-concept information system.

The results of this project can be used as a starting point for automated annotation. The information system alerts users to semantically uncertain segments of standards, provides background information, and allows them to document their decisions how to handle the semantically uncertain parts.

Keywords: Information system · Taxonomy · Semantic uncertainty

P. F. Pelz and P. Groche (Eds.): ICUME 2021, LNME, pp. 23–34, 2021.
https://doi.org/10.1007/978-3-030-77256-7_3

1 Introduction

Standards and Their Role in Product Development. Technical standards helped with rationalisation and quality management of the production of goods in the 20th century by organising and standardising the shape, size and design of products and processes in a meaningful way [25]. Today a plethora of international, national and regional organisations develop and publish technical standards to unify rules for the exchange of information, ensuring compatibility and reducing the variety of products, services, interfaces and terms [22]. Technical standards therefore play a role in many processes in the manufacturing industry as well as in product development processes.

The application of standards is voluntary, but can be mandatory by law or contract [22]. In all cases non-compliance with standards, at least in the European Union, is associated with high risks for manufacturers since in the case of product liability the burden of proof is on the manufacturer. When compliant with norms, the burden of proof is reversed [30]. To ensure compliance, standards have to be written clearly and concisely [5]. This is in stark contrast to the findings in [9]. Among users of technical standards there is a considerable lack of knowledge of how technical standards must be interpreted.

We attribute this difference to the need of technical standards to be applicable for a wide range of contexts, situations and new technical developments.

Uncertainty in Standards. While the main purpose of standards is to unambiguously regulate products and product development, they can not be entirely strict. One the one hand, there are aspects which defy complete strictness, such as design or different solutions to a problem which yield the same result. On the other hand, standards need to allow for innovation, which is only possible with a certain degree of flexibility and thus rules out complete strictness. However, standard compliance is only achievable if any and all uncertain parts are resolved and the solution is not only documented but also communicated to all persons involved.

Uncertainty in technical standards is foremost a lack of information and, hence, a lack of knowledge which makes resolving it primarily a matter of researching and understanding further information. Resolving uncertain parts adds to the to-do list and should be addressed in an early stage of the project to ensure compliance. Identifying and classifying uncertain parts in standards should be regarded as a form of division of labor. It is less time consuming to have a dedicated team analyze and annotate all standards relevant for a project than having each engineer go through them on their own.

Example. The phrase ‘allgemein anerkannte Regeln der Technik’ [*generally acknowledged rules of technology*] is a good example for uncertainty that arises through ambiguous language use. It hinges on various assumptions:

1. There are rules of technology,
2. there is a kind of review process for these rules the result of which has merit for everybody,

3. there is a possibility to know which rules of technology are considered to be generally acknowledged.

The phrase leaves the reader in a state of uncertainty, since it does not provide enough information to know which specific way of behaviour is part of the generally acknowledged rules and which is not. Only if there were a closed list of accepted rules of technology would this phrase not be uncertain. Since such a list would stand in the way of innovation, it cannot be provided even if it could be compiled. From this perspective, this phrase is also a good example for the need of uncertainty in technical standards. The authors of technical standards are completely aware of this phrase's ambiguity as is evident from DIN 45020 [8] where 'acknowledged rule of technology' is defined as 'technical provision acknowledged by a majority of representative experts as reflecting the state of the art' [8, entry 1.5] and 'state of the art' is defined as 'developed stage of technical capability at a given time as regards products, processes and services, based on the relevant consolidated findings of science, technology and experience' [8, entry 1.4]. Both definitions do not provide specific enough information to decide without further steps how to handle a given task.

Scope and Aims. The project was designed as a pilot study which means that proof-of-concept took precedence over depth. The project's main aim was to develop an annotation schema for uncertainty in the language of DIN standards, a taxonomy of uncertainty based upon it, and an information system which provides access to the categorized instances of uncertainty. Annotating has a long-standing tradition in the humanities and can be regarded both as a part of knowledge acquisition and as a scholarly primitive [17,29]. Basically any form of data enrichment, from writing notes in the margin of a manuscript to computationally classifying sentences or words, can be regarded as annotation. Developing an annotation schema is an iterative process in which classes and subclasses are created based upon concrete instances in the documents (see Sect. 3 for some details on the process). It makes sense to use the same environment for both annotating and the development of the annotation schema. We used the application Inception for both tasks [14]. The backend for the information system is a MySQL database where we stored information about the documents as well as the annotated instances of ambiguous language use. We chose the series DIN 1988, consisting of the parts DIN 1988-100, DIN 1988-200, DIN 1988-300, DIN 1988-500, DIN 1988-600 since these standards play a role in the work of the CRC 805, see e.g. [16].

2 Meaning, Knowledge, and Uncertainty

Words and Meaning. There are numerous theories and approaches concerning meaning in language which are subsumed (for an overview, see [2,3,21,23]). One of the most seminal models of the relationship between words and meaning is the 'semiotic triangle' [21, p. 11] (see Fig. 1).

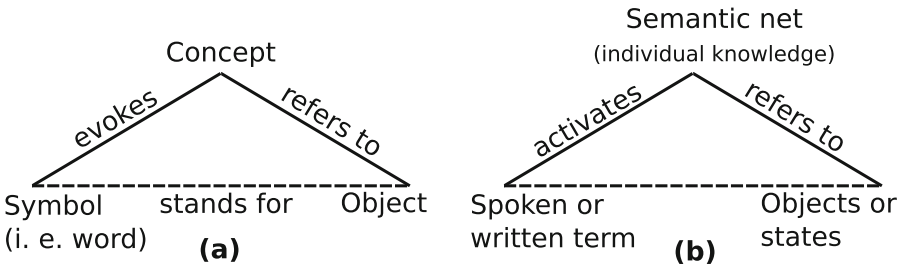

Fig. 1. Relationship between words and meaning. The *semiotic triangle* in (a) refers to language as a whole while the adaptation in (b) aims at an individual language user.

There is no direct connection between words and objects in the world. Words do not mean anything by themselves, rather, they trigger or activate parts of the knowledge store in our mind. The word *tree* does not contain a tree, it evokes the concept of a tree in the mind of the language user which is an abstraction of and a reference to the trees or a specific tree in the world. The semiotic triangle, which is also the basis for the general principles regarding concepts and terms in DIN 2330 [7], aims to illustrate the relationships between words and meaning in language in general, i.e. language as a system. However, language and language use (communication) are interdependent [2, p. 360]. On an individual level, words and their meaning are handled by the 'mental lexicon', which 'can be regarded as an individual network containing different kinds of personalized information on known words' [28, p. 6]. This also means that 'a word does not simplistically relate to a concept [...], but to a network of interrelated and overlapping distinct "senses", related to background world-knowledge' [19, p. 12] or, in other words, a semantic net.

For the purposes of this project, we understand uncertainty as a condition a) in which it is impossible to comply with the standards and b) which necessitates further steps of knowledge acquisition (see Fig. 2 below). We further consider this kind of uncertainty to be a result of ambiguous language use in technical standards.

Uncertainty enters language in various forms, the most notable of which are polysemy and underspecification. Polysemy occurs when a term activates multiple nodes of the network in the mental lexicon at once, for example the term 'mouse'. For a modern user of English, there are at least two concepts or senses activated upon hearing or reading this term. 1. *rodent.* 2. *peripheral computer device.* Usually, polysemy is resolved by taking into account the neighbouring terms (co-text) or the communicative setting (context) [13, cf. p. 7 f.].

Language, Knowledge, and Knowledge Acquisition. Even though language as whole can be regarded as a system shared and shaped by its users, the realms where individual language users are active are subsystems of language as a whole. These subsystems are formed and determined by (combinations of) socio-demographic factors like age, region, education, and, most notably for our purposes, occupation, specialization, and experience (these phenomena are studied

in detail in sociolinguistics [18], and LSP, languages for special purposes, [15]). Hence, the knowledge and 'senses' available in an individual's mental lexicon are in part determined by the same factors. Specific fields of knowledge like linguistics or engineering create and constantly reshape their own specialized subsystem of language as a whole in order to accurately denote objects and how they relate to each other (mathematics and formal logic can be regarded as a part of these specialized subsystems or as subsystems in their own right). The constant reshaping brings about a shift in meaning for some words and phrases since the concepts they refer to undergo change. For a member of a specific field to keep track of theses shifts in meaning, constant knowledge acquisition is in order.

For our purposes, we draw on [1,24] and regard knowledge acquisition to be a cognitive process which involves the following steps: Sources need to be found and (after evaluation) used to gather data presumed to be pertinent to the project in question. The data needs to be pre-processed (both computationally and cognitively) to transform it into information which in turn can be cognitively understood, which results in knowledge. The newly acquired knowledge needs to be applied, which entrenches it into the mind and adds to the explicit and implicit knowledge. All of these steps draw on previous knowledge which is why we regard knowledge acquisition to be an ongoing iterative process (see Fig. 2).

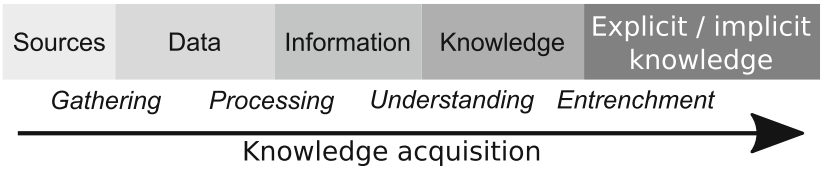

Fig. 2. Knowledge acquisition.

3 Taxonomy of Uncertainty

The taxonomy is the result of iteratively identifying and annotating (= assigning a class of uncertainty) instances of ambiguous language use in the technical standards. Identifying uncertain parts hinged upon the definition of uncertainty given above in Sect. 1, namely the answer to the question whether there was information missing in a sentence or the co-text of the sentence. Within each iteration, we inspected the emerging classes of uncertainty to ensure that they accurately reflected all instances of ambiguous language use and that they were sufficiently distinct from each other to avoid overlap. Both, the final annotations schema and the final annotations were validated by one last round of annotating, carried out by three engineers. Even though we focused on uncertainty arising from language use, we knew from previous experience with technical standards that there is at least one class of uncertainty which arises from conflicting knowledge rather than from lack of information conveyed by the text of a technical

standard. Consider the following example: An engineer who is familiar with a specific technical standard operates on the knowledge already present in his mind but is not aware that there is a newer version of the technical standard available in which something has changed. Let's assume that the changes themselves are unambiguous but in conflict with the previous version of the standard. This constellation leads to uncertainty which is independent from language use. Therefore we distinguish evident uncertainty from hidden uncertainty as first sub-classes of uncertainty and regard evident uncertainty to be any form of uncertainty that arises from language use.

Our analysis of the standards yielded the following classes of uncertainty (Fig. 3):

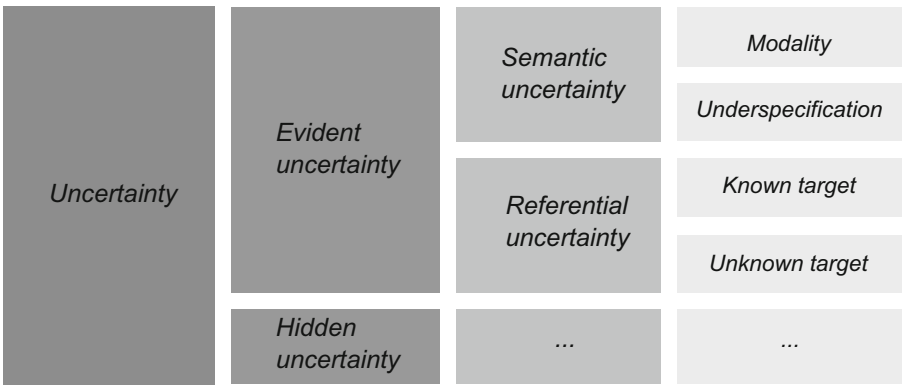

Fig. 3. Taxonomy of uncertainty.

Uncertainty that is grounded in terms and phrases is either modal or underspecified in nature. Modal uncertainty arises (intentionally) from any use of 'should' or 'can' leaving the decision which steps to take up to the standard user. Underspecification comprises any other case of ambiguous language use, ranging from phrases like 'the generally acknowledged rules of technology' to single words like 'bedürfen' in the following example: 'Dies gilt insbesondere für Apparate, die einer regelmäßigen Inspektion und Wartung bedürfen.' ['In particular, this applies to devices that are in need of scheduled inspection and maintenance.'] [6, p. 38]. To resolve the uncertainty, the maintenance needs for each device have to be checked. The instances of ambiguous language found in the technical standards comprise a *vocabulary of uncertainty* which will be the basis for the enhancements described below in Sect. 5. For a more detailed account of the taxonomy see [27].

4 Information System

Based on the taxonomy of uncertainty, we developed a proof-of-concept information system, which is targeted at engineers who work in a project where technical standards play a crucial role and annotating the documents is part of the project work. It is designed to provide the following features:

- a description of the taxonomy used to categorize the uncertain parts
- an overview over all standards that are relevant for the project
- a list of all uncertain parts of the annotated standards with the possibility to take notes
- inbuilt additional information on specific underspecified concepts
- possibility to add project specific information like for example instances of hidden uncertainty

Description of the Taxonomy of Uncertainty. The information system provides a detailed description of the taxonomy which offers the possibility to add project specific information. This is especially targeted at users who would like to redefine (parts of) the taxonomy or use project specific examples for the description to improve the project's internal communication and understanding.

Overview Over Standards Used. The overview is rendered as a network graph generated by the relationships between technical standards and a) their references to other technical standards which are listed as 'normative references' in each document, and b) other documents pertinent to the uncertain parts of the technical standards in question.

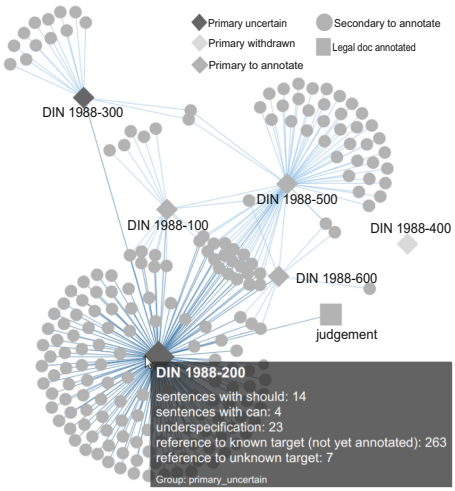

Fig. 4. Standards referenced by primary standards (edited screenshot of information system).

It not only shows which documents are linked to each other but also gives information about the group a document belongs to and about the annotation results (see Fig. 4). The groups are freely configurable to match the needs of a specific project. For our study we chose the following categories:

- *primary to annotate*: a technical standard directly pertinent to a given project
- *primary uncertain*: a technical standard directly pertinent to a given project which has been already annotated and contains uncertain parts
- *primary withdrawn*: a technical standard that is no longer valid but part of the series directly pertinent to a given project
- *secondary to annotate*: a not yet annotated technical standard which is linked to a primary document
- *legal doc annotated*: legal documents that contain information which helps to resolve some of the uncertain parts in the technical standards (here: a judgment)

As is evident from the categories, the information system is not only targeted at managing technical standards (= sources of uncertainty) but also any other documents which contain useful information. As an example for this, we chose a judgment which deals with a case where a newly installed drinking water system needed to be cleaned repeatedly and with enormous effort because the thread cutting agent used for cutting the pipes did not adhere to regulations [26]. We included this judgment for its descriptions of the steps taken to clean the pipes because they can be understood as an instance of following the 'generally acknowledged rules of technology'.

List of Classified Instances of Ambiguous Language Use. The core functionality of the information system is to display all uncertain parts in a structured way and provide a possibility to take notes on how to deal with specific instances of uncertainty in the technical standards in question. The default view shows all instances of all classes of uncertainty for all annotated technical standards. The tables on the top of the page provide links to more specific queries. Currently, these can be used to display

1. all instances of all classes of uncertainty found in a specific technical standard (first column of left table in Fig. 5)
2. all instances of a specific class of uncertainty found in a specific technical standard (second column of left table in Fig. 5),
3. and all instances of a specific class of uncertainty (first column of right table in Fig. 5).

The screenshot in Fig. 5 shows an excerpt of all uncertain items annotated as 'underspecified'. To limit this to underspecified items found in DIN 1988-200 the user just needs to click on *underspecification*.

Any specifications can be accessed via the link provided by the information system. The specifications provide a summary as well as an excerpt of the original document, and a link to the original document.

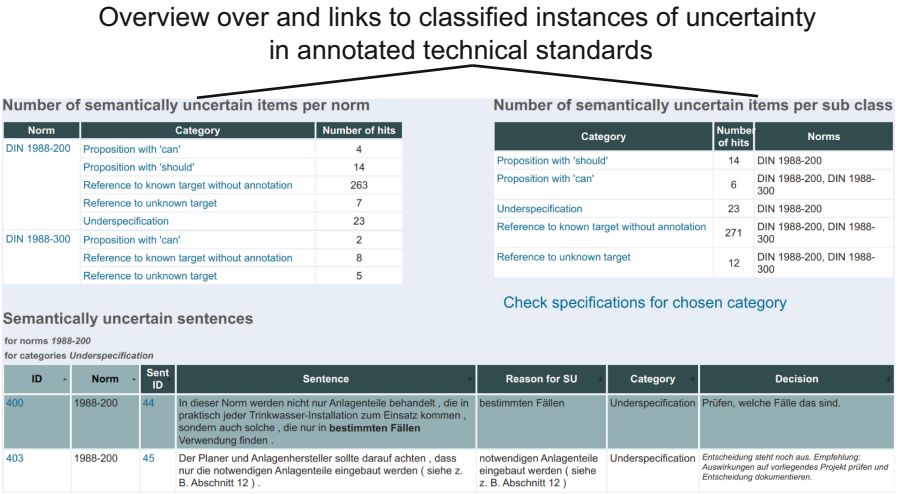

Fig. 5. Display of uncertain items in the information system.

5 Conclusion and Outlook

In the future, we will enhance the project in two ways. On the one hand, we will further develop the taxonomy of uncertainty and on the other hand, we will focus on automation, especially on automated annotation. To develop the taxonomy in a suitable manner, we will create a gold standard of correctly annotated instances of uncertainty, which means that we will annotate a larger number of carefully chosen technical standards. Both, determining the number of annotated instances and determining which technical standards to annotate requires time and consideration. The number of annotated instances needs to be high enough to yield significant results for rule-based automated annotation. The technical standards to annotate need to be representative for a given field of mechanical engineering and balanced with regard to aspects like document type, for example national vs. international codes. This brief outline of how we will proceed follows the best practices for corpus linguistic projects (for a more detailed account, cf. the section on methodological considerations in [4]). The gold standard of annotations will in turn allow us to make use of recent developments in computational linguistics with regard to automated classification and annotation, especially trainable classification systems like the ones provided by Inception [14]. Additionally, resources made available by lexicographical projects will be used to automatically retrieve synonyms for the instances of uncertainty (possible resources include for example [10–12, 20]. After evaluation with regard to their context dependent meanings,

these synonyms will be used to extend the vocabulary of uncertainty and, hence, the lexical material available for automated annotation.

References

1. Ackoff, R.L.: From data to wisdom. J. Appl. Syst. Anal. **16**, 3–9 (1989)
2. Acquaviva, P., Lenci, A., Paradis, C., Raffaelli, I.: Models of lexical meaning. In: Pirrelli, V., Plag, I., Dressler, W.U. (eds.) Word Knowledge and Word Usage: A Cross-Disciplinary Guide to the Mental Lexicon. De Gruyter Mouton (2020)
3. Allan, K., Jaszczolt, K.: The Cambridge Handbook of Pragmatics. Cambridge University Press, Cambridge (2012)
4. Biber, D., Reppen, R.: The Cambridge Handbook of English Corpus Linguistics. Cambridge University Press, Cambridge (2015)
5. DIN 820-2 Standardization - Part 2: Presentation of documents. Beuth Verlag GmbH (2018)
6. DIN 1988-200: DIN 1988-200:2012-05, Technische Regeln für Trinkwasser-Installationen – Teil 200: Installation Typ A (geschlossenes System) – Planung, Bauteile, Apparate, Werkstoffe; Technische Regel des DVGW. Beuth Verlag GmbH (2012)
7. DIN 2330: DIN 2330:2013-07. Begriffe und Benennungen: Allgemeine Grundsätze. Concepts and Terms: General principles. Concepts et termes: Principes généraux. Beuth Verlag GmbH (2013)
8. DIN 45020: DIN 45020:2007-03. Normung und damit zusammenhängende Tätigkeiten – Allgemeine Begriffe (ISO/IEC Guide 2:2004); Dreisprachige Fassung EN 45020:2006 Standardization and related activities – General vocabulary (ISO/IEC Guide 2:2004); Trilingual version EN 45020:2006 Normalisation et activités connexes – Vocabulaire général (ISO/IEC Guide 2:2004); Version trilingue EN 45020:2006. Beuth Verlag GmbH (2007)
9. Drechsler, S.: Kompetenzbedarfe von Maschinenbauingenieuren in Bezug auf Richtlinien, Normen und Standards zur Ausübung ihrer beruflichen Tätigkeit. Ph.D. thesis, Karlsruher Institut für Technologie (2016)
10. Fankhauser, P., Kupietz, M.: DeReKoVecs. IDS, Institut für deutsche Sprache Mannheim (2017)
11. Hamp, B., Feldweg, H.: GermaNet - a Lexical-Semantic Net for German. In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications, pp. 9–15 (1997)
12. Heid, U., Schierholz, S., Schweickard, W., Wiegand, H.E., Gouws, R.H., Wolski, W.: Das Digitale Wörterbuch der Deutschen Sprache (DWDS). DE GRUYTER, Berlin (2010)
13. Henrich, V.: Word Sense Disambiguation with GermaNet. Dissertation, Universität Tübingen (2015)
14. Klie, J.-C., Bugert, M., Boullosa, B., Castilho, R.E., de Gurevych, I.: The inception platform: machine-assisted and knowledge-oriented interactive annotation. In: Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations. Association for Computational Linguistics, pp. 5–9 (2018)
15. Humbley, J., Budin, G., Laurén, C.: Languages for Special Purposes: An International Handbook. De Gruyter Mouton, Berlin (2018)

16. Philipp, L., Lena, C.A., Pelz, P.F.: Energy-efficient design of a water supply system for skyscrapers by mixed-integer nonlinear programming. In: Operations Research Proceedings 2017: Selected Papers of the Annual International Conference of the German Operations Research Society (GOR), Freie Universiät Berlin, Germany, 6–8 Sept 2017, S. 475–481, Springer (2017). ISBN 978-3-319-89920-6
17. Meister, J.C.: From TACT to CATMA or a mindful approach to text annotation and analysis. In: Rockwell, G., Sinclair, S. (eds.) Festschrift for John Bradley (Forthcoming)
18. Mesthrie, R.: The Cambridge Handbook of Sociolinguistics. Cambridge University Press, Cambridge (2011)
19. Murphy, G.L.: The Big Book of Concepts, 1. MIT Press paperback ed. MIT Press, Cambridge [u.a.] (2004)
20. Naber, D.: OpenThesaurus: ein offenes deutsches Wortnetz. In: Gesellschaft für Linguistische Datenverarbeitung (Germany), Fisseni, B. (ed.) Sprachtechnologie, mobile Kommunikation und linguistische Ressourcen: Beiträge zur GLDV-Tagung 2005 in Bonn. Lang, New York, Frankfurt am Main (2005)
21. Ogden, C.K., Richards, I.A.: The Meaning of Meaning. Harcourt, Brace & World, New York (1923)
22. Feldhusen, J., Grote, K.H. (eds.): Pahl/Beitz Konstruktionslehre: Methoden und Anwendungerfolgreicher Produktentwicklung. Springer Vieweg, Berlin (2013)
23. Riemer, N.: The Routledge Handbook of Semantics. Routledge, London (2016)
24. Rowley, J.: The wisdom hierarchy: representations of the DIKW hierarchy. J. Inf. Sci. **2**, 163–180 (2007). https://doi.org/10.1177/0165551506070706
25. Spur, G., Krause, F.: Das virtuelle Produkt: Management der CAD-Technik. Hanser, München (1997)
26. Kullmann, S., Dressler, L.M.: VI ZR 229/93 (1994)
27. Stegmeier, J. Hartig, J., Bartsch, S., Leštáková, M., Logan, K., Rapp, A., Pelz, P.F.: Linguistic analysis of technical standards to identify uncertain language use. In: Groche, P., Pelz, P.F., et al. (eds.) Mastering Uncertainty in Mechanical Engineering, Springer (Forthcoming)
28. Trautwein, J.: The Mental Lexicon in Acquisition: Assessment, Size and Structure. Universität Potsdam, Potsdam (2019)
29. Unsworth, J.: Scholarly Primitives: what methods do humanities researchers have in common, and how might our tools reflect this? (2000)
30. Wendt, J., Oberländer, M.: Product Compliance: Neue Anforderungen an sichere Produkte. In: Zeitschrift für Energie- und Technikrecht (ZTR), vol. 2016, no. 02, S. 62–70. Pedell (2017). ISSN: 2224-6819