INTERNATIONAL DESIGN CONFERENCE – DESIGN 2020
https://doi.org/10.1017/dsd.2020.141

EXPLORING THE APPLICABILITY OF SEMANTIC METRICS FOR THE ANALYSIS OF DESIGN PROTOCOL DATA IN COLLABORATIVE DESIGN SESSIONS

N. Becattini [1,✉], G. V. Georgiev [2], Y. Barhoush [2] and G. Cascini [1]

[1] Politecnico di Milano, Italy, [2] University of Oulu, Finland
✉ niccolo.becattini@polimi.it

Abstract

The paper presents the application of non-specialized lexical database and semantic metrics on transcripts of co-design protocols. Three different and previously analyzed design protocols of co-creative sessions in the field of packaging design, carried out with different supporting tools, are used as test-bench to highlight the potential of this approach. The results show that metrics about the Information Content and the Similarity maps with sufficient precision the differences between ICT- and non-ICT-supported sessions so that it is possible to envision future refinement of the approach.

Keywords: *text mining, big data analysis, collaborative design, research methodologies and methods, human behaviour*

1. Introduction

Analysis of verbal data produced in collaborative design activities can help to characterize and gain a deeper understanding of the occurring cognitive processes. Semantic analysis of such verbal data allows for the quantification and comparison of information concerned with the collaborative design process, and its relation to the tools and technologies used to support the design process.

Prior endeavours to semantically analyse design protocols used semantic approach of lexical chain to analyse linguistic appraisals in design, and identify discontinuities in agreement in design problem solving (Dong, 2009). Few studies semantically investigated design activities in real-world settings (Georgiev and Georgiev, 2018; Georgiev and Taura, 2014) and apply objective semantic measures to quantify the observed processes (Cash et al., 2014; Georgiev and Georgiev, 2018; Taura et al., 2012).

The paper presents the application of semantic metrics on transcripts of co-design protocols, utilizing for the calculations a non-specialized lexical database. The approach we consider in this study utilizes several semantic measures that warrant the quantification of fundamental phenomena in design, linguistics and cognitive psychology. Such measures, which showed to be successful to study idea generation and creativity in design, include polysemy, abstraction, information content (IC) and semantic similarity (Georgiev and Georgiev, 2018; Georgiev and Casakin, 2019)). Second, it employs domain-independent and systematic representation of words (i.e., WordNet database). Third, the employed measures are faster to compute compared to other semantic analysis approaches used in the context of design conversations (e.g., Dong, 2009).

1.1. Analysing verbal interactions in co-design sessions

Verbal interactions are indicative of cognitive processes involved in design activities (Hay et al., 2017). Analysis of verbal interactions occurring in co-design sessions is a way to characterize them (Georgiev and Georgiev, 2018). Identifying further relevant information about the underling cognitive processes from recorded design conversations is a challenging task because not all aspects of human creative skills are expressed, verbalized or represented at a consciously accessible level (Boden, 2004). Technologies such as NLP (Dong, 2009) and semantic-based representations (Taura et al., 2012; Georgiev and Taura, 2014) allow for quantitative understanding of cognitive processes underlying activities (e.g., Kan and Gero, 2017). Absolute measurements (e.g., multiple meanings, Georgiev and Taura, 2014) or relative measurements between different verbalizations (e.g., networks, Taura et al., 2012) are means to provide further details into the cognitive processes occurring during collaborative design activities.

1.2. Availability of data and big data

Consistently with the general trend to rely on the increasing availability of data, the literature about the analysis of design protocol is progressively getting richer of contributions and attempts to improve the understanding of design-related activities through protocol data recorded by means of a wider set of sources of acquisition (e.g. Nguyen et al., 2018). This is typically done by the introduction of additional equipment and sensors that allow researchers to also record personal data about the physiological state of the subjects involved. These allow recording a wide variety of data, such as galvanic skin resistance (e.g. Hu et al., 2015), gaze (e.g. Ruckpaul et al., 2015), heartbeat (e.g. Steinert and Jablokow, 2013) as well as brain activity (e.g. Liu et al., 2014). Nevertheless, the data acquired through more traditional approaches (e.g. audio-visual) still have margin to deep-dive the analysis. For instance, video-recorded data allows for facial expression analysis (Balters and Steinert, 2017) and gestural analysis (Becattini et al., 2017). On the other hand, audio-recorded data allows to shed light on the dynamics among co-designers (Wulvik et al., 2017), but the analysis of content still mostly relies on traditional approaches. In-depth analysis of verbal interactions in conjunction with other sources of design sessions data are promising for furthering the understanding of design processes.
The objective of this paper is to explore the applicability of semantic metrics for the analysis of design protocol data in collaborative design sessions with practicing designers. Previously the semantic analysis approach has been successfully applied analysis of conversations in design education context (Georgiev and Georgiev, 2018; Georgiev and Casakin, 2019).
The structure of the paper is as follows: next section introduces background on the semantic analysis and metrics we employ; section 3 discusses the research method detailing our data and approach; the following section focuses on analysis and discussion of the data; we conclude the paper with outlook on this approach.

2. Background on semantic analysis approach and semantic metrics

Semantic networks model human memory as an associative system wherein each concept (represented by a node) can lead (by a link) to many other concepts (Boden, 2004). In the field of artificial intelligence, the semantic networks are utilized as computational structures that represent meaning in a simplified way within a conceptual space. Semantic networks were used to computationally model conceptual associations and structures in design (Taura et al., 2012).
Existing approaches to the analysis of verbal interactions are general purpose and typically not focused on fundamental phenomena in design. In this study, we consider an alternative semantic analysis approach to the existing ones (Dong, 2009). The metrics (Georgiev and Georgiev, 2018; Georgiev and Casakin, 2019), are easy to compute, focus on fundamental phenomena in design (e.g., design creativity, Georgiev and Casakin, 2019), and offer a systematic representation.
These metrics focus on nouns, as noun-noun combinations and noun-noun relations play an essential role in designing (Dong, 2009), furthermore, similarity or dissimilarity of noun-noun combinations is found to be related to creativity through yielding emergent properties of generated ideas (Wilkenfeld

https://doi.org/10.1017/dsd.2020.141 Published online by Cambridge University Press

and Ward, 2001). Four semantic metrics are used to quantify fundamental process with regard to the design process, which includes polysemy, abstraction, information content, and semantic similarity.
Polysemy is defined as the quality of a word having multiple meanings. It has been identified as an essential manifestation of the flexibility, adaptability, and meaning potential, in relation to creativity (Fauconnier and Turner, 2003). Abstraction is defined as a generalisation from specific instances that carries a lower level of detail in information, and where common features are identified or abstracted, process which can lead to novel ideas (Ward et al., 2004). Information content (IC) is defined as the amount of information transmitted by a particular unit of language in a specific context (Georgiev and Georgiev, 2018). IC measures the degree of informativeness of a unit, therefore, units with higher IC have a lower probability of occurrence. Semantic similarity is used to quantify the strength of semantic relationships between units or instances of language (e.g., Resnik, 1995). It allows quantifying how alike are two words. It has been useful in the identification and representation of fundamental processes in design thinking (Georgiev and Georgiev, 2018).
These metrics have been already proved as effective in performing protocol analysis in context of design education (Georgiev and Georgiev, 2018; Georgiev and Casakin, 2019). Our original contribution here is applying these metrics to design sessions using different technological support (comparing the impact of different design representations on the verbal interactions between co-design team members). This is it relevant because these are real design sessions outside design education relying on technology. On a longer perspective, this will help analysing the efficacy of a design-supporting tool in comparison of process without tool.
There are other potentially relevant applications of the semantic in different contexts. For example, the most typical measures used in natural language processing are those related to semantic similarity (e.g., Resnik, 1995).

3. Research method

In order to explore the potential of semantic analysis and grab relevant information out of them by means of related metrics, the authors decided to rely on three existing design protocols that were previously transcribed, segmented and coded by 6 persons that received a tailored training for this purpose. This will be essential to run a preliminary check of the potential of applying semantic metrics to verbal/textual content, as it is necessary to refer the results of the semantic analysis to known outcomes. The adoption of already existing, coded and analysed protocols, in fact, allows the ex-post interpretation of the results of the semantic analysis under the light of previous findings emerged from the same protocols.

3.1. Sessions to study

Three different and previously analysed design protocols of co-creative sessions in the field of packaging design carried out with different supporting tools, are used as test-bench to highlight the potential of this approach (SPARK Project consortium, 2018). Table 1 shows the main differences and similarities between sessions.

Table 1. Summary of the already analysed protocols used for the semantic analysis

Session Name	Design domain	Design Task	Technology support	Participants	Markers in the sessions
ARTEFICE AR	Packaging design	Pumpkin soup packaging	ICT: Augmented Reality	2 Designers (A,B) + 2 clients (a,b)	Screenshots for good ideas
ARTEFICE SAR	Packaging design	Pumpkin soup packaging	ICT: (Spatial Augmented Reality)	2 Designers (C,D) + 2 clients (c,d)	Screenshots for good ideas
ARTEFICE NO ICT	Packaging design	Pumpkin soup packaging	None	2 Designers (E,F) + 2 clients (e,f)	No Markers

More precisely the three sessions were carried out for a controlled experiment carried out during the SPARK project. The aim was to compare the differences between the effects that different ICT tools

https://doi.org/10.1017/dsd.2020.141 Published online by Cambridge University Press

(Augmented Reality and Spatial Augmented Reality - i.e. AR rendered through projections) have on design creativity and refer those differences to a control group that was not supported by any ICT tool. All three co-design teams were composed of 2 designers and 2 clients. All the protocols dealt with the same design task: the development of the packaging for a fresh soup held into a single serving plastic bowl with film lid and cardboard sleeve. All the protocols followed the same general structure, sharing the same objective: further develop three pre-prepared alternative designs for the cardboard sleeve graphics and layout by combining graphical elements (colours, logos, text, images etc) in order to propose a complete packaging design. Despite the shared general structure, the co-designers participating to the session with no ICT support managed to produce just one valuable idea at the end of the session, despite they got inspired by the three concepts to further develop. The ICT supported sessions, in turn generated one good idea for the three concepts. Co-designers participating in the AR-supported protocol, differently from SAR protocol, also noted a few additional intermediate promising concepts before defining the final concepts for the three proposals.

3.2. Semantic analysis metrics

3.2.1. Indexes and metrics for their calculation

Semantic networks are employed to provide a structural representation of knowledge in the form of graphs. After exteriorizing and representing knowledge in the form of a semantic network, a number of graph-theoretic measures could be employed for quantitative analysis. The following is a sample of four graph-theoretic (network-theoretic) measures that were computed with WordNet 3.1 is-a hierarchy of nouns. These four measures employ network composed of word nodes (connected in is-a network hierarchy), meaning nodes (terminal nodes called leaves that represent all the meanings of a word node), and links between the nodes (Georgiev and Georgiev, 2018):

- Polysemy is the number of direct links between a word node A and its meaning nodes, reckoning for the number of meanings of the word node (Georgiev and Taura, 2014). For example, ‘mind’ node has seven noun meaning nodes of ‘head’, ‘recall’, ‘judgment‘, ‘thinker’, ‘attention’, ‘idea’, and intellect’.
- Abstraction of word node A is the normalized fraction of the shortest path distance from the root word node to a word node A, and the maximal shortest path from the root in the network. Abstraction accounts for how generalized is the word node compared to the most specific instance (Georgiev and Georgiev, 2018).
- Information Content (IC) is the amount of information carried by a word node inside the graph. The IC is measured as a normalized fraction of the number of leaves of the word node, and the maximal number of leaves in the network (Blanchard, 2008; Georgiev and Georgiev, 2018).
- Semantic Similarity of two word nodes, A and B, is measured by the IC of the least common subsumer (LCS) of the two words (Resnik, 1995), essentially quantifying how alike are the two word nodes. The LCS of A and B is the most specific word node which is an ancestor of both A and B in the is-a hierarchy (e.g., the LCS of ‘boat’ and ‘car’ is ‘vehicle’).

3.2.2. System implementation

In the implemented system, a Natural Language Processing (NLP) pipeline module is responsible for the calculation of the four semantic variables. It loads three graphs, then takes the text from speech-to-text process and list of nouns. In the semantic variables module, for each of the four semantic measures, a function was created to calculate their values.
In the implemented system (see Figure 1), an NLP pipeline module transforms the input text data, through a series of steps, to a calculation value for each of the four semantic variables.
The NLP pipeline module imports several other modules, including the Utilities module, the Semantic Variable module and the spaCy tagging model.
The Utilities module is responsible for loading graphs. The NetworkX package (https://networkx.github.io/) is used for re-constructing, manipulating, and utilizing functions of networks on the

https://doi.org/10.1017/dsd.2020.141 Published online by Cambridge University Press

graphs from Georgiev and Georgiev (2018). The Semantic Variable module contains functions for calculating semantic variables and imports a Graph Helper module, which includes functions for helping in graph calculations, such as the lowest common ancestor for two words.

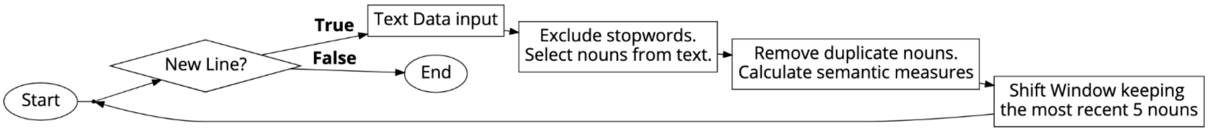

Figure 1. Pipeline of the implemented system

The spaCy model (https://spacy.io/models/en) features neural models for tagging, parsing and entity recognition. It utilizes available pre-trained statistical models for English. The model allows for part-of-speech (POS) tagging and dependency parsing.

3.3. Approach for the comparison of sessions

Due to the commonalities and differences the three protocols have, it is expected to notice akin similarities and difference with the results produced by the semantic analysis of the transcripts collecting verbal interactions among co-designers. As the values computed for the four metrics vary in time, depending on the considered time frame or on the number of protocol segments considered at once, the comparison between the values computed for the four metrics and the characteristics of sessions will be mainly carried out through estimators (moments) of descriptive statistics (e.g. mean and standard deviation). In terms of targets for the comparison, three different evaluations appear to be meaningful (Figure 2).

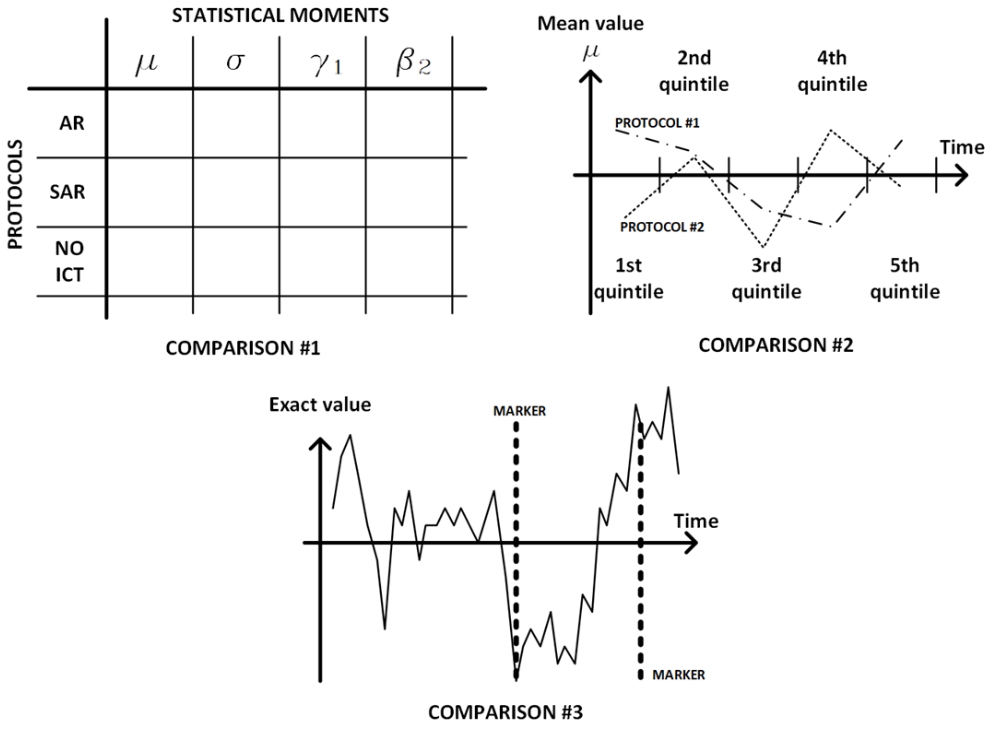

Figure 2. Summary of the experimental comparisons run on the protocols for the semantic metrics

First, as the overall goal of the three protocols is the same, the differences between the mean values computed for the 4 metrics presented in section 3.2 on the whole protocols should be small if not negligible.

Second, the overall dynamics of the three sessions was observed to be different, as mentioned in section 3.1. Therefore, a more fine-grained description of the protocols in the time should raise

https://doi.org/10.1017/dsd.2020.141 Published online by Cambridge University Press

similarities and differences that reflect the process co-designers followed during the protocol. For this reason, protocols get split into segments of uniform duration to identify potential similar patterns, which is also consistent with previous applications of the same metrics for the investigation of parts of the conversation (Georgiev and Georgiev, 2018). The segmentation of protocols or part of them into three parts (tertiles), as for the previous applications of the semantic metrics, appear to be a not particularly suitable choice because of the nature of the considered sessions. Indeed, the two sessions that received ICT support allowed participants to work on three different concepts along with the protocol. Yet, these sessions focused on the three concepts with non-uniform durations for the three concepts, so that two developed concepts can fall into a tertile, while one gets free of them. For this reason, the protocols get split into five parts of uniform duration (quintiles) for the comparison of trends of average values in time.
Third, the two sessions that received ICT support are also characterized by the presence of markers that co-designers introduced in the AR/SAR system to spot the emergence of good ideas. The comparison between sessions can be carried out to check if these markers in the protocol correspond to values, for one or more metrics, are uniformly high and/or low (i.e. if markers in the sessions correspond to peaks or valleys for the distribution of metrics values, in time, along with the protocol).

4. Analysis and discussion of the experimental data

4.1. First comparison based on statistical moments

The four semantic measures of polysemy, abstraction, IC and semantic similarity were calculated as moving window of five nouns. Calculations are based on existing information-theoretic and graph-theoretic formulas (Resnik, 1997; Blanchard, 2008). Table 2 collects the figures for the four metrics by means of statistical moments used as estimators.

Table 2. Summary of the first four statistical moments (Mean, Standard Deviation, Skewness and Kurtosis) for the distribution of values generated by the computation of the four metrics; Negative values in italics

	Mean	St.Dev	Skew.	Kurt.		Mean	St.Dev	Skew.	Kurt.	
Polysemy	0,504	0,090	*0,011*	*0,177*	**AR**	0,672	0,075	0,072	*0,028*	**Information content**
	0,480	0,078	*0,138*	0,132	**SAR**	0,654	0,082	*0,029*	*0,116*	
	0,480	0,083	*0,030*	*0,048*	**NO ICT**	0,663	0,076	*0,023*	*0,105*	
Abstraction	0,663	0,041	*0,381*	*0,183*	**AR**	0,261	0,076	0,361	*0,023*	**Similarity**
	0,670	0,042	*0,879*	1,651	**SAR**	0,270	0,073	0,420	0,004	
	0,661	0,038	*0,013*	*0,137*	**NO ICT**	0,278	0,077	0,623	0,593	

The figures of Table 2 confirm the expectation mentioned in Section 3.3. As the three protocols (AR, SAR, NO ICT) focused on the same design task and followed the same overall plot for the session (i.e. further development of concepts and ideas for the packaging of a fresh pumpkin soup), the values should not present particular differences between sessions.
By looking at the mean values it is possible to notice that they are different across the 4 metrics (Polysemy varies around 0,5; Abstraction and information content around 0,66; Similarity around 0,27). Metrics by metrics, figures about mean values are also quite similar across the three protocols: they typically vary in the range of few hundredths (max variation is for polysemy: 0,024). The standard deviation shows a similar behaviour across the three protocols, as it ranges in thousandths, except for polysemy (max variation 0,012). This means that the distribution of values for the four metrics across the session is globally the same in terms of average value and dispersion (despite dispersion is much more significant for similarity, as its mean values are lower) but the data are in any case distributed slightly differently. As the third comparison will focus on peaks and valleys, this should appear more meaningful for protocols that have a positive kurtosis (few values around the mean are more frequent than the

others) and a marked skewness (presence of long tails in the distribution), being it negative or positive. Indeed, compared to the first two moments, skewness and kurtosis have particularly different values within and between the metrics. Most of the distributions by metrics have a longer left tail except for similarity (it is the only metrics presenting positive values for skewness). Almost all the distributions also present negative values for the kurtosis, meaning that these distributions across the metrics are platykurtic (flat): average values are not particularly more frequent than the others recorded, the distribution of data is similar to a rectangle. A few exceptions deserve being mentioned, as SAR have positive kurtosis for polysemy (0,132), abstraction (1,651) and similarity (0,004), while the other only positive value for kurtosis is for the non-ICT supported protocol (0,593).

4.2. Second comparison of protocols segmented into quintiles

The segmentation into quintiles allows for the comparison of the different protocols in time, as the expectation is that sessions sharing similar elements (e.g. being supported by ICT technology) also show a more marked similarity in time. Figure 3 presents four graphs, one per each of the considered metrics.

By visually analysing the four graphs, it emerges that the ICT-supported protocols (continuous and dotted lines) are not so similar, in time, for both polysemy and abstraction. However, their behaviour is particularly more correlated for the two semantic metrics concerning the information content and similarity, as Table 3 witnesses. For information content (bottom left diagram), the behaviour is M-shaped for the AR- and SAR-supported sessions. The diagram about the similarity metrics (bottom right in Figure 3) shows that AR- and SAR-supported sessions are both characterized by W-shaped curves.

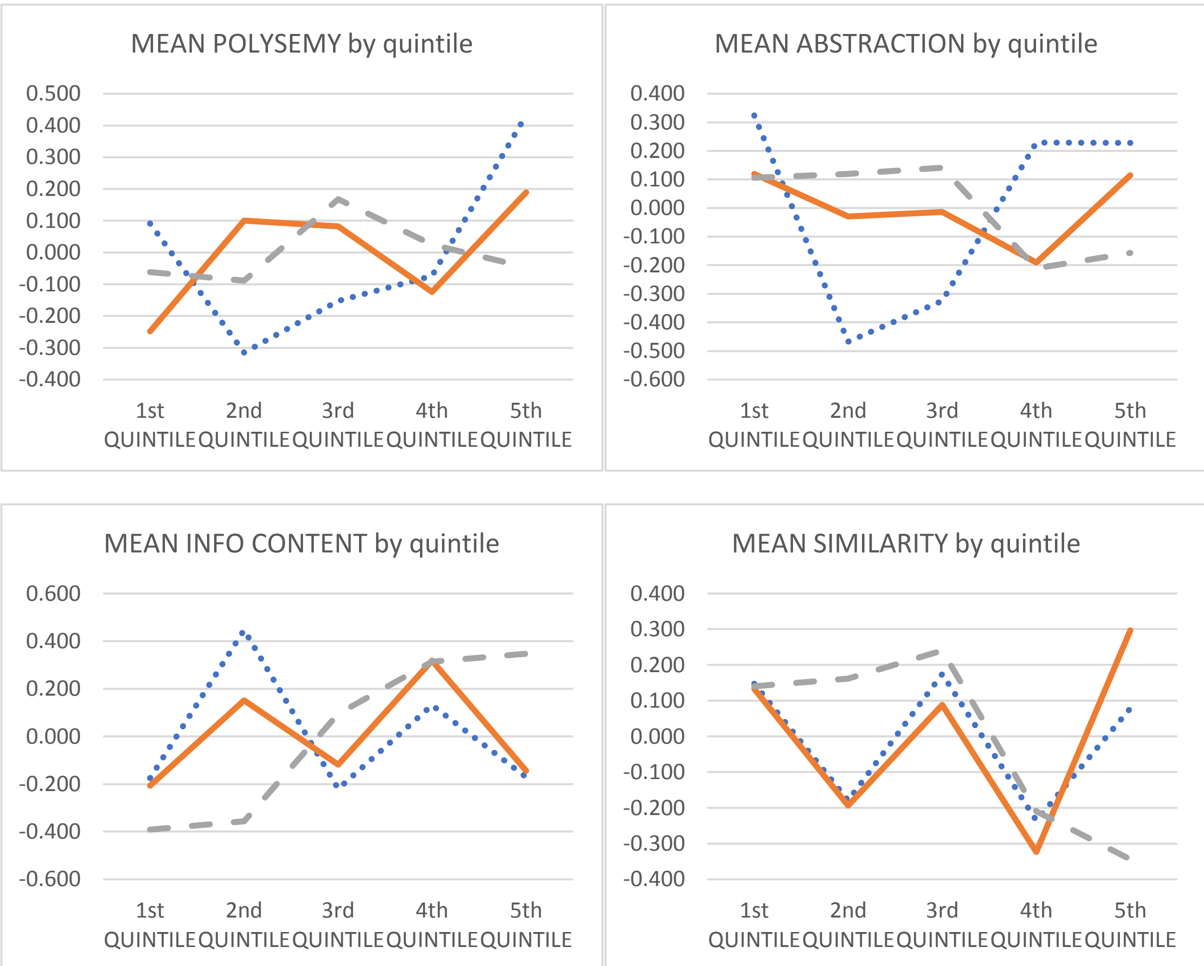

Figure 3. Mean values organized by quintile for the four semantic metrics (dotted line: AR protocol; continuous line: SAR protocol; dashed line: NO ICT protocol)

https://doi.org/10.1017/dsd.2020.141 Published online by Cambridge University Press

Table 3. Pairwise Pearson's correlation between the protocols, by semantic metrics

	Polysemy	Abstraction	Information Content	Similarity
AR-SAR	0,12	0,21	0,77	0,88
AR-NO ICT	-0,22	-0,61	-0,32	0,28
SAR-NO ICT	0,11	0,34	0,21	-0,10

This is a promising result due to the application of semantic metrics on transcribed protocols. In fact, both the ICT-supported sessions already demonstrated to facilitate communication (SPARK consortium, 2018), as the presence of a flexibly changeable shared design representation (rendered through AR or SAR), reduced the need of iterating unproductive design moves. A simpler communication, indeed, appears to be more related to the latter two metrics, for which the correlations are higher than for the former two.

4.3. Comparison of trends and correlations between sessions (on the same metrics)

Despite the second comparison showed that some metrics present fewer similarities than expected (e.g. abstraction and polysemy) and the results of the descriptive statistics for the first comparison show that the distribution of data is not particularly different among the three protocols for what concerns mean values, the exploration also considered the opportunity to rely on semantic metrics values to potentially spot relevant moments in the protocol. As the values computed for the four semantic metrics vary in different ranges, the identification of peaks shared by more than one metric at a time required the standardization of values. Each of the computed values has been therefore standardized according to the following formula: z=(x-mean)/standard deviation. This enables the adoption of a single threshold to identify values in both the left and right tail of these distributions and spot peaks and valleys. Figure 4 shows an excerpt of a spreadsheet with the transformed values for the four semantic metrics on the left-hand side, segments from the transcribed protocol on the right and some columns collecting YES/NO cells in the centre.

Time	z-polysemy	z-abstraction	z-info_content	z-similarity	CANDIDATE	POLYSEMY	ABSTRACTION	INFO_CONTENT	SIMILARITY
00.48.24	0.533	0.356	-0.484	0.361	NO	NO	NO	NO	NO
00.48.34	0.533	0.356	-0.484	0.361	NO	NO	NO	NO	NO
00.48.35	0.533	0.356	-0.484	0.361	NO	NO	NO	NO	NO
00.48.35	1.568	1.968	-1.390	2.357	YES	YES	YES	NO	YES
00.48.42	1.568	1.968	-1.390	2.357	YES	YES	YES	NO	YES
00.48.43	1.568	1.968	-1.390	2.357	YES	YES	YES	NO	YES
00.48.45	1.568	1.968	-1.390	2.357	YES	YES	YES	NO	YES
00.48.46	1.568	1.968	-1.390	2.357	YES	YES	YES	NO	YES
00.48.47	1.568	1.968	-1.390	2.357	YES	YES	YES	NO	YES
00.48.50	0.043	0.625	-1.004	-0.047	NO	NO	NO	NO	NO
00.48.54	-0.836	-0.743	-0.031	-0.296	NO	NO	NO	NO	NO

Time	z-polyser	Stop words	Nouns in the line	Line number	Line content as for the original doc
00.48.24		[]	[]	800	Yes send it to disappearance ... under under under under under
00.48.34		[]	[]	801	That is right ...
00.48.35		[]	[]	802	There ... you can see ... there is there is there is
00.48.35		['part']	['thing' 'picture' 'picture']	803	A timid shy thing is part of a picture ... it is no longer framed in the picture but there it
00.48.42	1	[]	[]	804	Ok
00.48.43	1	[]	[]	805	I like it like that enough this is so beautiful
00.48.45	1.	[]	['thing' 'thing']	806	Ok .. One thing ... one thing ...
00.48.46	1.	[]	['screenshot']	807	I take the screenshot ...
00.48.47	1	[]	['thing']	808	One thing I noticed now is missing ...
00.48.50	0	['top']	['parsley' 'leaf']	809	that is if you could put a parsley leaf on top of it
00.48.54	-	[]	['soup']	810	But on the soup or on ...?

Figure 4. Spreadsheet for the identification of markers in the protocol (right side, cells with a red background) and the identification of peaks and valleys in the set of standardized values for the four metrics (YES cells in the centre shows if the z-values overcome a threshold)

https://doi.org/10.1017/dsd.2020.141 Published online by Cambridge University Press

The excerpt of the spreadsheet shows that there is preliminary evidence that the markers mentioned in the protocol, which correspond to moments where co-designers took screenshots of ideas/concepts that they considered worthy of further development, could be recognized by means of the outliers in the z-distributions of the four semantic metrics. It should not be surprising that the metrics on Information Content (column titled IC) does not overcome the threshold, as this result is consistent with the general conclusions that emerged with the statistical analysis done for the first comparison. This approach, however, requires to arbitrarily set a value for the threshold, so that it is possible to spot "convergence" of peaks or valleys by more than one of the metrics. The concurrent presence of peaks/valleys (outliers in the distributions) can be considered as a potential, semi-automatic, trigger to spot "relevant moments" in the protocol so that the analysts can focus on the moments before and after them. Nevertheless, it is also worth mentioning that this approach also highlights potential "false positive" moments, as the selection of the threshold is not univocal and there is not yet any rule or guideline for its selection.

5. Summary and outlook

It was possible to differentiate sessions based on semantic metrics. The results show that metrics about the Information Content and the Similarity maps with sufficient precision the differences between ICT- and non-ICT-supported sessions so that it is possible to envision future refinement of the approach.

The strengths of the approach can be summarized in terms of speed and automation of analysis. Moreover, it is an objective and replicable approach that can be applied nearly real-time.

As for weaknesses of the approach, we rely on a general-purpose thesaurus. The approach used in this study utilizes a non-specialized lexical database WordNet 3.1. Consequently, there are limitations in terms of lack of representation of domain-specific words and meanings. Our intent is to replicate the same study after customization of the thesaurus so as to check whether the metrics become more effective and the overall approach more robust. Should this be confirmed, then we should determine under what conditions the results can be considered valuable enough to justify the investment in such customisation.

Furthermore, there are some weaknesses of the approach attributable to NLP in general, for example, parsing of verbal data is not always fully accurate. It should be noted that we use one of the best performing parsers currently available.

In summary, in this study, we used three different and previously analysed design protocols of co-creative sessions in the field of packaging design, carried out with different supporting tools, to test-bench a semantic analysis approach for discriminating the different design sessions. We were able to map differences with some of the metrics and to outline further refinements on the approach that may provide further insights into design activities.

Acknowledgements

The work reported in this paper is part of the SPARK project, funded by the European Union's Horizon 2020 research and innovation programme (GA688417). This paper reflects only the authors' views and the EU Commission is not responsible for any use that may be made of the information it contains. The authors would like to acknowledge the valuable support of all SPARK partners in this research activity. This research also has been partially financially supported by the Academy of Finland 6Genesis Flagship (grant 318927).

References

Balters, S. and Steinert, M. (2017), "Capturing emotion reactivity through physiology measurement as a foundation for affective engineering in engineering design science and engineering practices", *Journal of Intelligent Manufacturing*, Vol. 28 No. 7, pp. 1585-1607.

Becattini, N. et al. (2017), "Characterisation of a co-creative design session through the analysis of multi-modal interactions", *International Conference on Engineering Design*, Vol. 2017, pp. 479-488.

Blanchard, E., Harzallah, M. and Kuntz, P. (2008), "A generic framework for comparing semantic similarities on a subsumption hierarchy", In: Ghallab, M., Spyropoulos, C.D., Fakotakis, N. and Avouris, N. (Eds.), *ECAI 2008: 18th European Conference on Artificial Intelligence including Prestigious Applications of Intelligent Systems (PAIS 2008)*, IOS Press, Patras, Greece, pp. 20-24.

https://doi.org/10.1017/dsd.2020.141 Published online by Cambridge University Press

Boden, M.A. (2004), *The creative mind: Myths and mechanisms*, Routledge.
Cash, P., Štanković, T. and Štorga, M. (2014), "Using visual information analysis to explore complex patterns in the activity of designers", *Design Studies*, Vol. 35 No. 1, pp. 1-28. http://dx.doi.org/10.1016/j.destud.2013.06.001
Dong, A. (2009), *The Language of Design: Theory and Computation*, Springer, London.
Fauconnier, G. and Turner, M. (2003), "Polysemy and Conceptual Blending", In: Nerlich, B., Herman, V., Todd, Z. and Clarke, D. (Eds.), *Polysemy: Flexible Patterns of Meaning in Mind and Language*, Mouton de Gruyter, Berlin and New York, pp. 79-94.
Georgiev, G.V. and Casakin, H. (2019), "Semantic Measures for Enhancing Creativity in Design Education", *Proceedings of the Design Society: International Conference on Engineering Design*, Vol. 1 No. 1, pp. 369-378. https://doi.org/10.1017/dsi.2019.40
Georgiev, G.V. and Georgiev, D.D. (2018), "Enhancing User Creativity: Semantic Measures for Idea Generation", *Knowledge-Based Systems*, Vol. 151, pp. 1-15. http://dx.doi.org/10.1016/j.knosys.2018.03.016
Georgiev, G.V. and Taura, T. (2014), "Polysemy in design review conversations", *10th Design Thinking Research Symposium, Purdue University, West Lafayette*, Purdue University, Indiana.
Hay, L. et al. (2017), "A systematic review of protocol studies on conceptual design cognition", *Design Computing and Cognition'16*, Springer, Cham, pp. 135-153.
Hu, W.L., Booth, J. and Reid, T. (2015, August), "Reducing sketch inhibition during concept generation: psychophysiological evidence of the effect of interventions", *ASME 2015 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference.* American Society of Mechanical Engineers Digital Collection.
Kan, J.W. and Gero, J.S. (2017), *Quantitative methods for studying design protocols*, Springer, Dordrecht.
Liu, Y. et al. (2014), "A fuzzy psycho-physiological approach to enable the understanding of an engineer's affect status during CAD activities", *Computer-Aided Design*, Vol. 54, pp. 19-38.
Nguyen, P., Nguyen, T.A. and Zeng, Y. (2018), "Empirical approaches to quantifying effort, fatigue and concentration in the conceptual design process", *Research in Engineering Design*, Vol. 29 No. 3, pp. 393-409.
Resnik, P. (1995), "Using information content to evaluate semantic similarity in a taxonomy", *IJCAI'95 Proceedings of the 14th international joint conference on Artificial intelligence - Volume 1: Morgan Kaufmann Publishers*, pp. 448-453.
Ruckpaul, A., Nelius, T. and Matthiesen, S. (2015), *Differences in analysis and interpretation of technical systems by expert and novice engineering designers*.
SPARK Project consortium (2018), *Deliverable Results of the experiments benchmarking the platform*.
Steinert, M. and Jablokow, K. (2013), "Triangulating front end engineering design activities with physiology data and psychological preferences", *DS 75-7: Proceedings of the 19th International Conference on Engineering Design (ICED13), Design for Harmonies, Vol. 7: Human Behaviour in Design*, Seoul, Korea, 19-22 August 2013, pp. 109-118.
Taura, T. et al. (2012), "Constructive simulation of creative concept generation process in design: a research method for difficult-to-observe design-thinking processes", *Journal of Engineering Design*, Vol. 23 No. 4, pp. 297-321. http://dx.doi.org/10.1080/09544828.2011.637191
Ward, T.B., Patterson, M.J. and Sifonis, C.M. (2004), "The Role of Specificity and Abstraction in Creative Idea Generation", *Creativity Research Journal*, Vol. 16 No. 1, pp. 1-9. http://dx.doi.org/10.1207/s15326934crj1601_1
Wilkenfeld, M.J. and Ward, T.B. (2001), "Similarity and emergence in conceptual combination", *Journal of Memory and Language*, Vol. 45 No. 1, pp. 21-38. http://doi.org/10.1006/jmla.2000.2772
Wulvik, A., Menning, A. and Steinert, M. (2017), "A computational approach to expose conversation dynamics in engineering design activities", *DS 87-2 Proceedings of the 21st International Conference on Engineering Design (ICED 17) Vol 2: Design Processes, Design Organisation and Management*, Vancouver, Canada, 21-25 August 2017, pp. 101-110.