Non-directive corpus annotation to reveal individual perspectives with underspecified guidelines: the case of mental workload

Iuliia Arsenteva[1,2], Caroline Dubois[1], Philippe Le Goff[1], Sylvie Plantin[1], and Ludovic Tanguy[2]

[1]Orange Innovation
[2]Université Toulouse - Jean Jaurès
{iuliia.arsenteva, caroline.dubois, philippe2.legoff, sylvie.plantin}@orange.com
ludovic.tanguy@univ-tlse2.fr

Abstract

This paper investigates personal perceptions of mental workload through an innovative, non-directive corpus annotation method, allowing individuals of diverse profiles to define their own dimensions of annotation based on their personal perception. It contrasts with traditional approaches guided by explicit objectives and strict guidelines. Mental workload, a multifaceted concept in psychology, is characterized through various academic definitions and models. Our research, aligned with the principles of the perspectivist approach, aims to examine the degree to which individuals share a common understanding of this concept when reading the same texts. It seeks to compare the corpus produced by this non-directive annotation method. The participants, mainly employees of a large French enterprise and some academic experts on mental workload, were given the freedom to propose labels and annotate a set of texts. The experimental protocol revealed notable similarities in labels, segments, and overall annotation behavior, despite the absence of predefined guidelines. These findings suggest that individuals, given the freedom, tend to develop overlapping representations of mental workload. Furthermore, they demonstrate how non-directive annotation can uncover shared and diverse perceptions of complex concepts like mental workload, contributing to a richer understanding of how such perceptions are constructed across different individuals.

1 Introduction

Defining the scheme and guidebook is a crucial step in any text annotation process. While uniform guidelines facilitate achieving consensus and homogeneity, they can also mask variability in annotators' perspectives. However, when a target phenomenon refers to complex concepts, it becomes particularly important to consider and analyze the multiplicity of viewpoints. Therefore, adaptations to this methodology are necessary to avoid framing annotators and to account for the multiplicity of perspectives. Perspectivism (Cabitza et al., 2023) offers a reflection on this question and discusses the consequences on annotated data used for machine learning in NLP.

Our research aims to assist in designing a language processing tool that can detect elements related to mental workload in employee messages within companies.

Mental workload is a critical phenomenon, as it represents a major concern that affects many aspects of workplace wellness. In the workplace, mental workload directly influences employee well-being, productivity, and overall job satisfaction. As a result of various media efforts to popularize the concept, mental workload has become a common notion frequently referenced in the workplace. Our goal is to develop a more grounded mental workload model that can be compared to current academic ones. Subsequently, this model will be used in the analysis of text messages. As demonstrated by Le Gonidec (2022), individual perception plays a crucial role in mental workload. This concept relates to how individuals perceive tasks and their environment. Therefore, exploring methods to account for the diversity in the perception of this phenomenon during annotation is particularly important. By incorporating diverse points of view, we can ensure that no important aspects are overlooked in developing a more grounded model of mental workload.

Our aim is to explore the possibility of using a personal and individual text annotation method at an early stage. To investigate this, we asked people working in different positions within the same company, as well as academic experts on mental workload, to freely express their points of view on a common multifaceted object by annotating a shared set of texts. Therefore, we propose a kind of *hyperperspectivist* and non-directive approach, which is achieved without providing a specific definition

Proceedings of the The 4th Workshop on Perspectivist Approaches to NLP (NLPerspectivists 2025), pages 142–152

of the target phenomenon or annotation guidelines. As a result, we constituted a collection of annotated data from participants with no prior annotation experience.

This study investigates the hypothesis that there is a common representation, even partial, of mental workload existing among individuals. It can be observed and made explicit through text annotation. Additionally, we hypothesize that significant variations also exist and, if formalized, they could contribute to a richer and more comprehensive view of the concept. Based on this hypothesis, the study aims to address the following research questions:

1. Is text annotation a suitable method for collecting and analyzing the points of view of untrained professionals on a complex concept such as mental workload?

2. Do individuals share a common (or at least partial) representation of mental workload? This hypothesis aims to determine whether common elements emerge in the way people perceive and evaluate mental workload, which would suggest a collective understanding of the concept.

3. Does the representation of mental workload differ depending on professional profiles? This hypothesis explores whether occupational differences influence how each individual perceives and annotates mental workload.

We seek to address these questions with the perspective of designing an annotation scheme. This scheme will be applied in machine-learning solutions aimed at automatic processing.

In this paper, we first introduce the concept of mental workload and the multiplicity of viewpoints in text annotation in Section 2. In Section 3, we describe the research methods, including the used data, the participants, and our experimental protocol. We then present the analysis of the collected data and the main results in Section 4, before drawing conclusions in Section 5.

2 Theoretical backgrounds and related work

2.1 Mental Workload

Despite a marked interest in the topic over the past 40 years, there is no clear and universally accepted definition of mental workload (Cain, 2007).

As per the IWA (Individual – Workload – Activity) model proposed by Galy (2016), mental workload is defined as the cognitive demand of a task (Sweller, 1988). It includes the mental effort required to perform a task and can be assessed through various indirect measures, such as subjective measures (self-reported assessments of mental effort and perceived tension), performance measures (behavioral indicators such as response accuracy and latency) and psychophysiological measures (physiological responses, for example, heart rate variability reflecting cognitive load).

To enhance generalizability, Longo et al. (2022) presented a more operational and modellable definition. Mental workload (MWL), according to them, represents the degree of activation of a finite pool of resources, which are limited in capacity, while cognitively processing a primary task over time. This process is mediated by external stochastic environmental and situational factors, as well as affected by definite internal characteristics of a human operator, for coping with static task demands, by devoted effort and attention.

Regarding the workplace environment, the use of technologies is increasing every day, and academic research has shown a relationship between mental workload and 'technostress' in the professional context (Castillo et al., 2023). This work highlighted that studying these two concepts together can offer advantages, such as the development of new strategies to help workers and managers deal with technostress. Understanding these concepts is essential, as new technologies have become an integral part of work, affecting both performance and well-being.

In terms of applicability of mental workload research, findings in aeronautics Martin et al. (2013), demonstrated that while modeling MWL is a valuable approach to synthesize existing literature and to develop assessment methods, it cannot replace empirical studies that further refine and clarify its boundaries.

2.2 Dealing with subjectivity in annotation

Since we aim to collect annotations of subjective interpretations of data, we are interested in a perspectivist approach. The perspectivism is a recent movement in the field of Natural Language Processing and is increasingly utilized in the annotation of subjective topics. One of the main concerns in annotation is the bias introduced by the cultural

context of annotators, making it crucial to consider all points of view in non-objective topics.

The perspectivist approach, proposed by Cabitza et al. (2023), aims to address the representativeness and reliability to fundamental truth in machine learning systems and has been adopted by several researchers. Plank (2022) highlights the importance of human label variations in machine learning pipeline, emphasizing that such variation is often mistakenly treated as noise but should be treated as an opportunity to make systems more trustworthy. Basile (2020) also critiques the gold standard approach, arguing that it is inadequate for subjective tasks such as detecting irony, sarcasm, or abusive language, where the perspectives of the annotator can vary significantly.

Chulvi et al. (2023) suggest that the disagreement in the annotations of sexist texts is rooted in social factors rather than individual differences. The related work on sexism annotation was published a year later by Tahaei and Bergler (2024) who studied the effect of demographic characteristics of annotators in sexism detection. Their experiments showed that including annotators from different demographic groups can improve performance in classifying sexist tweets. Goyal et al. (2022) demonstrated that self-identified backgrounds (e.g., African American, LGBTQ, or neither) influence the toxicity assessments in online comments.

In this study, we seek to maximize the application of perspectivism by not only including annotators of different profiles, but also giving them a wide range of freedom in their actions. To achieve this, we mobilize annotators from the outset, even before developing an initial annotation model (Pustejovsky et al., 2017). This early involvement allows us to capture a wide range of interpretations and insights, ensuring that the annotation process is informed by diverse viewpoints from the very beginning. In doing so, we aim to create a more robust and inclusive annotation framework that can better accommodate the complexity and variability of linguistic data.

3 Methods

3.1 Data and participants

Usually, annotation task guidelines are designed to be as precise and objective as possible. However, in this study, our objective was to explore subjectivity and to capture a wide range of perspectives on the same topic. Therefore, we intentionally limited the specific assignments provided to the annotators.

For this study, we selected eight short messages (ranging from 78 to 245 words) from various campaigns of "Micro Ouvert". "Micro Ouvert" is an internal tool developed and used by a large French company to collect employees' spontaneous opinions on a number of topics while ensuring complete anonymity. These eight texts reflect the insights of employees on topics such as their experiences with changes in the workspace, attending conferences, and their general motivation to work at the company. An example of such a message is presented in A.1 in its original French version, along with the English translation. All the language data used and collected are in French, and the experiments were conducted solely in French with our English translations provided in the paper.

We recruited four experts in mental workload, all of whom are academics in psychology with a focus on this topic, and 23 employees from the same company, including managers, human resources specialists, and other executives, particularly concerned by MWL in their team management roles. Indeed, it was crucial that the participants were motivated to perform the annotation tasks. These participants will hereafter be presented as members of the following 4 categories: 4 mental workload Experts (E), 8 Human Resources Specialists (HR), 12 Managers (M), and 3 Other specialists (O). The participants had no relationship with the experimenters, which prevented bias that could compromise the results. It is important to note that none of the participants in our study had prior experience with annotation tasks, neither experts nor employees. All participants were given information on the context of each message to be annotated and a possible clarification by the experimenters in case of misunderstanding. However, a significant difference between the experts and the employees is that the former were external to the firm and had no prior knowledge of the working context. In contrast, the employees were more familiar with the company's culture and the general content of the messages.

Having outlined the diverse participant profiles and message selection in the previous section, we now turn to the experimental protocol, which describes the methodology employed to capture the subjective interpretations of MWL.

3.2 Protocol

The participants were not informed in advance about the details of the experimentation but were only made aware of the subjects of our study and the estimated duration of the session (approximately 75 minutes). They were informed about the data collection process and they gave their written consent to participate in the study. The experimentation was conducted in the form of recorded interviews with two experimentalists (co-authors of this paper) and included two parts. The sessions lasted between 1 and 2.5 hours per participant. The final set of 8 texts was chosen after pre-testing with 5 participants, based on time demands and the observation of fatigue among the participants by the end of the study.

In the first part (which took an average of 40% of the session), the participants were assisted in defining the mental workload from their perception, during a semi-structured interview, using the explicitation techniques (Vermersch, 1994). The participants were encouraged to explain their representation of MWL without being directed towards a particular definition. Then, the experimentalists extracted keywords from their actual speech and submitted them to the participant for validation. We will refer to these labels as *prior labels*. It provides us with the initial set of data: the recorded definition of MWL as well as the keywords associated with this concept. The main purpose of this step is to define the labels for the next task: annotation.

In the second part of the task, the participants were asked to annotate eight short messages using the open-source text annotation tool *Doccano* (Nakayama et al., 2018). The tool was specifically configured to propose only their own prior labels. The participants had to annotate text parts illustrating an expression of mental workload, according to their judgment, without any constraint on the segments. The participants were also allowed to freely introduce new labels, which were then added into the annotation interface by the experimentalist. The labels used to annotate at least one segment are referred to as *used labels*. Additionally, the participants were allowed to assign multiple overlapping labels to their annotations. All annotators received the same sequence of texts, but the sequence was rotated so that each annotator began with a different text. This standard counterbalancing was made in order to minimize the learning effect and balance the performance among all the users for all the texts.

After each interview with each participant, we were able to collect the following data: 1) the list of prior and used labels 2) the annotated segments from the eight messages (start offset, end offset, and associated label).

At this stage, we did not rename or unify the labels that had similar meanings, and instead kept the exact formulation expressed by the participant. For example, we treated *priority* and *prioritization* as two distinct labels. However, the experimenters took care to consistently put down and unify the formatting of all repeated keywords and phrases across the sessions. We also ignored the polysemy in this study. The results of this methodology are presented in subsequent sections.

4 Analysis and results

To address the first research question concerning the similarity of individual representations of mental workload, we conducted a comparative analysis of the following components: labels, segments, relations between labels, segments, and user categories.

4.1 Labels

First, to identify patterns that indicate a shared understanding of MWL, we began with an examination of the prior labels (defined by the annotators during the first phase of the interview) and the used labels (actually employed in the annotation process). The maximum number of prior labels in the annotation task was 12, with a minimum of 5. On average, each participant defined 9 labels. In contrast, the maximum number of labels utilized during the annotation task was 15, while the minimum was 5, with an average of 10 labels per each participant. The average intersection between prior and used labels is 7.

We identified the labels that were commonly defined and/or used by different participants. At this stage of our analysis, we only consider strictly identical labels. The most frequently used labels are presented in Table 1, along with the number of annotators who proposed (center column) and used them (rightmost column). Labels in boldface correspond to those for which the frequency of use is at least 3 more that the frequency as prior labels.

We can see that the most frequent labels refer to individuality (*individual, personal*), demand (*temporality*, *pressure*), and task (*complexity, meaning*).

Labels	Annotators proposed	Annotators used
individual	7	7
temporality	7	6
meaning	1	5
complexity	4	5
stress	4	5
objectives	4	4
pro	4	4
professional/personal	4	4
uncertainty	1	4
context	4	4
ill-being	1	3
task	3	3
personal	4	3
permanence	4	3
environment	2	3
prioritization	3	3
time	2	3
pressure	2	3
powerlessness	0	3
recognition	0	3
volume	4	3

Table 1: The most frequently used labels (by at least 3 distinct annotators)

These three dimensions align with academic models of mental workload (Longo et al., 2022; Galy, 2016; Le Gonidec, 2022).

It is noteworthy that the individual feature of mental workload was among the first labels proposed and used by participants. While models of mental workload by Sweller, Galy, and Le Gonidec acknowledge this individual aspect, it is usually less emphasized in discussions between non-experts, especially when compared to more common aspects such as time management and task multiplicity. Indeed, the second most frequently used label corresponded to the temporal dimension of mental workload.

As indicated in bold in Table 1, some labels were frequently added by the participants during the annotation step. They reflect the fact that some aspects of mental workload were identified (or remembered) by the participants and could be considered as contingent to the topics of the selected messages. However, some of them *meaning*, *uncertainty* were also suggested as prior labels by other participants, and in any case, they were considered relevant for the annotation.

4.2 Label clusters

As mentioned earlier, many of the labels collected were semantically or formally close. To obtain a more global view of the dimensions considered by the participants, we employed a large language model (LLM) to propose a clustering of all labels.

We selected Claude Sonnet 3.5 (Anthropic, 2024) among other mainstream LLMs by prompting it to propose some categories and to regroup the 197 distinct used labels. The prompt is provided in A.2. We did not ask for a precise number of clusters but ended up with the 14 listed in Table 2, along with the number of labels and the number of participants (Annotat.) who used at least one of them for annotating. An LLM was chosen over a human annotator to ensure objectivity. The authors reviewed the associations and categories and were generally satisfied with the results. The choice of LLM over other clustering methods is defined by the fact that our approach focused on grouping labels that made sense together and, more importantly, on giving those groups clear names.

Cluster	Annotat.	Labels
Emotional and Psy. Aspects (I)	18	24
Relationships and Interactions (I)	18	19
Workload (W)	17	15
Environment and Context (A)	17	17
Temporality (W)	17	10
Organization and Management (A)	14	16
Cognitive and Mental Aspects (I)	13	16
Constraints and Difficulties (W)	13	15
Balance and Well-being (I)	13	14
Abilities and Skills (I)	11	11
Impact and Consequences (W)	11	12
Processes and Actions (A)	9	11
Adaptation and Change (A)	8	9
Management and Recognition (I)	5	8

Table 2: Label clusters and description according to Claude. IWA refers to the components of Galy's model

The resulting clusters align to the three main components of the IWA (Individual, Workload, Activity) model of mental workload Galy (2016). Therefore, we can consider that these induced categories confirm that the labels do not diverge from the models of mental workload, and that they can serve as a coarser-grained way to identify the qualitative behaviors of the participants.

4.3 Annotated segments

Having analyzed the labels, we now turn our focus to the annotated text segments. The average number of segments per annotator is 69, with an average length of 39 characters. The minimum number of segments recorded for a user was 13, while the maximum reached 248. The shortest segment had a length of 1 character (a question mark), compared to a maximum length of 1144 characters (i.e. an entire message). This variation was expected due to the lack of guidelines and constraints.

At this stage, we aimed to find specific segments of text that attracted the most attention from the participants, without considering the associated labels.

We extracted the most frequently annotated segments along with the number of users who annotated each segment. Although annotators had the option to overlap their annotations and assign multiple labels to the same text segment, we decided to consider only the number of users to avoid distorting the interpretation of markers across all users. We calculated, for each character position in each text, the number of different users who included it in at least one annotated segment, and then identified the contiguous characters that exceed a given threshold. Based on the inflection point in the curve displaying the number of segments, we selected a threshold of 14 different annotators (50% of them) and identified 52 different text segments. These segments consist of a variety of text units, ranging from single words (e.g., stress, meaning) to entire phrases (e.g., Reconnect with nature, with our environment, with humans). The complete list of these segments is provided in French in A.3.

After analyzing these segments, we noticed that the vocabulary containing words with negative connotations attracted the most attention. The annotators associated the text elements reflecting discomfort, overwhelm, overload, and disconnection with an increased mental workload.

Following a separate analysis of labels and segments, the next section explores the relationships between them.

4.4 Labels and Segments Similarity

Variability of annotation across annotators comes from both the labels used (intentional similarity) and the segments delimited (extensional). The latter similarity has been considered at the dataset level, as we have identified the main zones of interest in the target texts. On a finer grain, we aimed to identify the extent to which specific labels are used by different participants to tag the same text segments.

If we consider the set of text segments labeled L by participant P across the corpus, we can define the extensional similarity the set of segments labelled L' by another participant P' as the amount of overlapping. More precisely, we used the Jaccard index to measure the ratio between the number of characters (defined by their offsets) that the two sets have in common, to the union.

If $ext(L_P)$ is the set of characters (offsets) labeled as L by participant P (as one or several segments), we can define the extensional similarity between two labels from two participants (L_P and $L'_{P'}$ as the Jaccard index between the labeled text segments:

$$simext(L_P, L'_{P'}) = \frac{|ext(L_P) \cap ext(L'_{P'})|}{|ext(L_P) \cup ext(L'_{P'})|}$$

In other words, if the two participants used two labels (either different or identical) to tag the exact same parts of the target texts, $simext$ will be 1, while it will be 0 if there is no overlap.

We computed the $simext$ values for every pair of labels from two different users. We then focused on two specific subsets of pairs.

First, we considered the pairs of different labels with a high level of similarity ($simext > 0.4$, for a total of 169 pairs). We observed 15 cases where different annotators applied the exact same labels to the same segments. For example, on the same segment *leaving us in the dark* three annotators applied the label *uncertainty* independently. This is understandable, as the selected messages influenced the choice of labels.

Second, as expected, we found a number of synonyms used to tag the same text parts, for example *pause* and *respite*.

Third, and more interestingly, we found cases where the same segments were labeled with semantically related words, although not synonyms. For example, the labels could describe either the *causes* or the *consequences* of the same phenomenon. In this instance, the same segment of the text *a month before the show, and not 3 days before* was annotated by different persons using the label *temporality* for one and *to juggle* for the other. This means that one person describes increased MWL as a result of time constraints, while another perceives it as a consequence involving the need to juggle and manage multiple tasks simultaneously. Another example is when different users applied the labels *uncertainty* and *ability to reason* to the same text segment *leaving us in the dark*. In the first case, the annotator used a paraphrase of the segment itself to express the *cause*, while the other used a label for a *consequence*, indicating that this impacts the ability to reason. These cause-and-consequence designations are supported by participants' verbal expressions during the test. The first annotator

stated: "This uncertainty is caused by the lack of answers that the person expected, as before, and that they could receive." In contrast, the second annotator said: "Leaving us in the dark means that there are no indicators or elements to be able to reason with, to face something."

This clearly indicates that although we found a number of indications that the annotators exhibit similar behavior, there are also variations in their points of view on this complex concept.

It is important to note that we didn't observe any contradiction in annotators' behavior: while there were complementary variations, there were no opposing opinions, similar to what has been observed in other related studies on perspectivism.

4.5 Overview of annotator behavior

In this last section, we examine variations among individual annotators to identify the main profiles and assess whether these variations are linked to the annotators' position and status.

We conducted a multidimensional analysis at the annotator level by computing the following variables:

- *NbDistinctLabels*: the number of distinct labels initially proposed by the participants during the first stage of the interview (before the annotation). Min:5, Max:15, Avg: 10.1.
- *UsedLabels*: the ratio of these prior labels that were actually used for annotation. Min: 41%, Max: 100%, Avg: 81%.
- *AddedLabels*: the ratio of used labels that were added by the participant during the annotation stage. Min: 0, Max: 67%, Avg: 30%.
- *NbSegments*: the total number of text segments produced by the participant during the annotation. Min: 13, Max: 248, Avg: 69.1.
- *TextCoverage*: the total amount of text (number of characters) included in at least one annotated segment. Min: 427, Max: 5065, Avg: 1987.
- *AverageOverlap*: the average number of segments in which an annotated text part is included (at the character level). Min: 1, Max: 2.7, Avg: 1.3.

These differences between the participant groups for each variable are illustrated in the boxplots in Figure 1. While there is no clear difference for

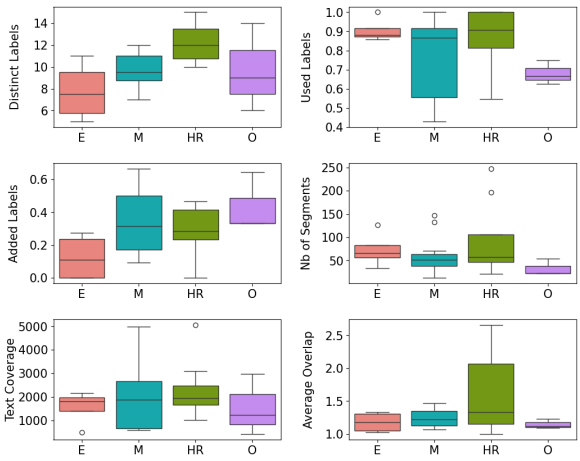

Figure 1: Boxplots of behavior variables across participants' categories (E=Mental workload experts, M=Managers, HR=Human Resources specialists, O=Others)

variables associated with text segment selection (number of segments, overlap, and text coverage), significant variations are observed in the choice of labels. Experts tend to have fewer labels, use all their initial proposals, and do not need any additional ones. HR specialists and managers exhibit roughly similar profiles, but the former use a larger number of labels. We note that the *Other* group is in the middle ground and remains inconclusive without additional participants.

To obtain a more global picture, we performed a principal component analysis (PCA) on the matrix representing each of the 27 participants across the 6 quantitative variables. The main factor map is shown in Figure 2. The variables are represented as blue arrows, and the participants are depicted as colored dots according to their group. The confidence ellipses (at the 95% level) are shown around the barycenter for each group. The first factor map is sufficient as it captures 70% of the total variance.

The first principal component (horizontal axis) is positively correlated with all variables except *AddedLabels*. On the right, there are the participants who produced a large number of segments, with a significant overlap and a high number of labels. The HR specialists are predominant, while the managers are located on the left side of the map. In other words, HR specialists exhibit a more dispersed annotation behavior with more segments and labels, in a more cumulative manner than the managers (who are less productive).

The second component (vertical axis) opposes participants who used a large amount of their ini-

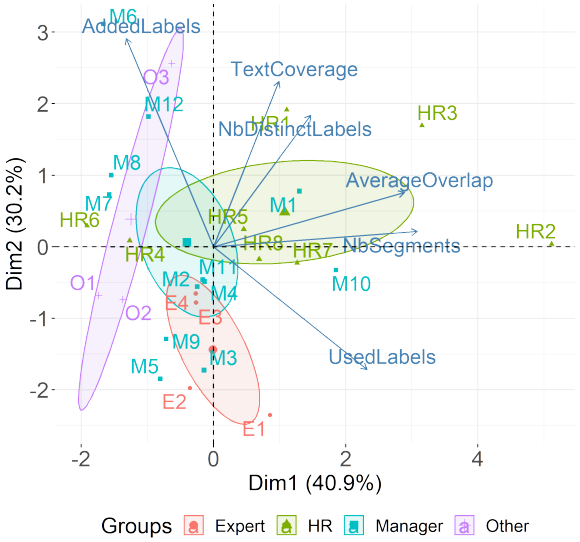

Figure 2: Principal Component Analysis of the participants

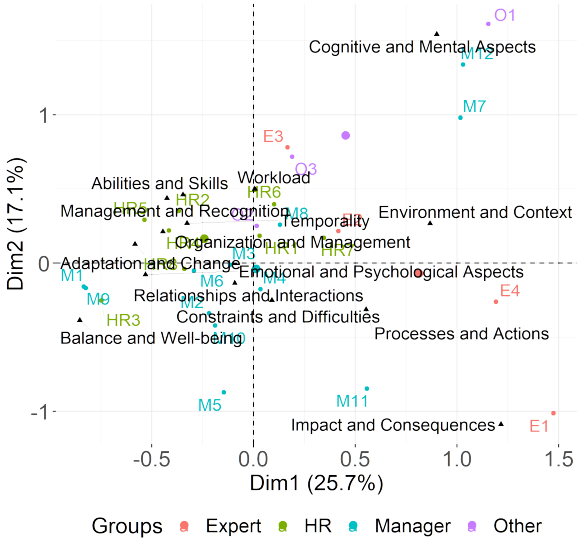

Figure 3: Correspondence Analysis: participants and label clusters

tially proposed labels (bottom) to those who had to provide additional labels in the course of the annotation process. It also appears that this distinction is correlated with the text coverage.

It is interesting to observe that the experts are very homogeneously located at the bottom of the factor map, even though none of them had prior experience with annotation and they come from diverse academic backgrounds. This seems to indicate that their knowledge allowed them to correctly anticipate the dimensions of the phenomenon, and that the labels they initially provided at the beginning of the interviews were both relevant and sufficient. Additionally, it appears that they produced a more focused set of segments, with less coverage and overlap than naive participants from other groups.

Our final analysis considers the clusters of labels that we requested an LLM to identify (see Table 2). We considered the number of labels from each cluster used by each participant (without considering the number or size of the corresponding segments) and performed a correspondence analysis. The first factor map is shown in Figure 3.

Here also, we can see that the experts exhibit a specific and coherent behavior. They all appear on the right side of the map, standing out with label clusters such as *Impact and consequences* and, to a lesser extent, *Processes and actions*. The second group that differs is the Other annotators, who focus on *cognitive and mental aspects*. On the left, Managers and HR professionals are positioned together, suggesting that they share similar preferences in the label categories used during annotation. The categories on this left side appear to be more individual-focused, encompassing aspects such as *Balance and Well-being*, *Relationships and Interaction*, *Adaptation and Change*. These clusters reflect individual experience and interpersonal dynamics within the context of mental workload.

These results seem consistent with the mission of managers and HR, more focused on individuals, while experts are more interested in mental workload processes.

Beyond this specific analysis, it appears that it remains possible to perform a qualitative analysis and to identify global tendencies in the annotations, even in the absence of specific guidelines.

5 Conclusion and future work

This study explored the subjective perceptions of mental workload through an innovative annotation method involving participants of diverse professional profiles, with different expertise related to the MWL concept. Regarding the formulated research questions, we can state that firstly, text annotation can serve various purposes, notably the analysis of different points of view on mental workload. Next, the analysis of the results showed that, even without clear instructions, people share a common representation of mental workload. Finally, despite this convergence in the representation, we observed

differences in user behavior based on their professional roles. By allowing annotators to define their own labels and freely annotate text segments, we captured a variety of perspectives on MWL. We observed similarities in the annotated segments, labels, and groups, indicating a common representation of MWL, as we anticipated. Additionally, differences emerged, highlighting the influence of individual professional profiles on the perception of this topic. The comparison between experts' and non-experts' approaches allowed us to see differences in the process of identifying MWL elements in the text. Therefore, we gained a better understanding of the non-expert analysis and potentially considered new candidate facets in the MWL concept by leveraging employees' knowledge of the workplace context.

This holistic approach promotes richer and more representative annotation and can thereby improve models for analyzing and interpreting textual data. The collected data can be viewed as explicit, structured, exemplified, and tested individual models of the MWL. The same annotation protocol could be applied to topics beyond mental workload, enabling a more inclusive approach.

Since we recorded all the interviews and have a transcript of the participants' comments on their own actions (such as definition, reformulating labels, choice of segments, etc.) we possess an even richer dataset than what we have presented here, which requires further analysis and effort.

As part of a larger project, we are now considering the development of an automatic annotation process for mental workload in the messages. However, we are currently at the stage of defining the annotation scheme. This innovative approach opens the way to a better understanding of the linguistic and cultural nuances that influence the assessment of mental workload, while emphasizing the importance of integrating different perspectives to enrich textual data analysis models.

Limitations

One of the limitations of this study is the small dataset of texts, which displayed limited variability, as the messages primarily focused on a small number of topics. This limitation is caused by the collection of texts within a specific working environment, and further investigations are needed to generalize the findings to other contexts. Nevertheless, we aimed to compile a dataset from various campaigns to present annotators with messages on different topics, all still related to work.

Another limitation is that the participants were recruited from a population that was very busy at work, allowing them to dedicate only 1 to 2 hours to the study. This time constraint may have impacted their level of involvement in the process. However, all participants were engaged in the task and performed well in annotation, despite having no prior experience. We assess their performance based on the annotations they produced as well as the received feedback. All participants expressed significant interest in the task and demonstrated high motivation and self-awareness; however, they also noted that it was complicated and mentally resource-consuming.

References

Anthropic. 2024. Claude 3.5 sonnet.

Valerio Basile. 2020. It's the end of the gold standard as we know it. on the impact of pre-aggregation on the evaluation of highly subjective tasks. In *CEUR Workshop Proceedings*, volume 2776, pages 31–40. CEUR-WS.

Federico Cabitza, Andrea Campagner, and Valerio Basile. 2023. Toward a perspectivist turn in ground truthing for predictive computing. In *Proceedings of the 37th AAAI Conference on Artificial Intelligence (AAAI 2023)*, pages 6860–6868.

Brad Cain. 2007. A review of the mental workload literature. Technical Report RTO-TR-HFM-121-Part-II, Defence Research and Development Canada Toronto, Toronto, Ontario, Canada.

Jose-Manuel Castillo, Édith Galy, and Pierre Thérouanne. 2023. Le technostress et sa relation avec la charge mentale en contexte professionnel. *Psychologie du Travail et des Organisations*, 29(4):197–213.

Berta Chulvi, Lara Fontanella, Roberto Labadie-Tamayo, and Paolo Rosso. 2023. Social or individual disagreement? Perspectivism in the annotation of sexist jokes. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics.

Edith Galy. 2016. Approche intégrative de la charge mentale de travail : une échelle d'évaluation basée sur le modèle ica (individu – charge – activité). In *Actes du 51e Congrès de la Société d'Ergonomie de Langue Française (SELF)*, Marseille, France.

Nitesh Goyal, Ian D. Kivlichan, Rachel Rosen, and Lucy Vasserman. 2022. Is your toxicity my toxicity? Exploring the impact of rater identity on toxicity

annotation. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–27.

Nolwenn Le Gonidec. 2022. *Conceptualiser et évaluer la charge mentale de salariés dans un contexte d'usage d'outils numériques : Le cas d'une entreprise de télécommunications*. Ph.D. thesis, Université Côte d'Azur, France.

Luca Longo, Christopher D. Wickens, Gabriella Hancock, and P. A. Hancock. 2022. Human mental workload: A survey and a novel inclusive definition. *Frontiers in Psychology*, 13:883321.

Caroline Martin, Sylvain Hourlier, and Julien Cegarra. 2013. La charge mentale de travail : un concept qui reste indispensable, l'exemple de l'aéronautique. *Le Travail Humain*, 76(4):285–308.

Hiroki Nakayama, Takahiro Kubo, Junya Kamura, Yasufumi Taniguchi, and Xu Liang. 2018. doccano: Text annotation tool for human. Software available from https://github.com/doccano/doccano.

Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 10671–10682. Association for Computational Linguistics.

James Pustejovsky, Harry Bunt, and Annie Zaenen. 2017. Designing annotation schemes: From theory to model. In *Handbook of Linguistic Annotation*, pages 21–72. Springer.

John Sweller. 1988. Cognitive load during problem solving: Effects on learning. *Cognitive Science*, 12(2):257–285.

Narjes Tahaei and Sabine Bergler. 2024. Analysis of annotator demographics in sexism detection. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 376–383. Association for Computational Linguistics.

Pierre Vermersch. 1994. *L'entretien d'explicitation en formation continue et initiale*. ESF, Paris, France.

A Appendix

A.1 Sample message

French Original:

> Points à améliorer: - Avant SDLR: Faire en sorte que la Brand revoie les slides un mois avant le salon, et pas 3 jours avant. Les modifications imposées sont très importantes et nécessitent une surcharge de travail tant pour etre en conformité que pour modifier le discours pour les visiteurs. Demander des slides en français et en anglais plutôt que de nous laisser dans le flou (sur la partie anglaise).

English translation:

> Areas for improvement: - Before Research and innovation fair: Have the Brand review the slides a month before the show, and not 3 days before. The changes imposed are very significant and require a lot of work both to comply and to modify the presentation for visitors. Ask for slides in French and English rather than leaving us in the dark (about the English part).

A.2 Prompt used for clustering labels

The prompt was designed and submitted in French. Below is the translation to English by the authors:

I've interviewed people and asked them what mental workload means to them, and they've given me a list of terms that they associate with the notion of mental workload. Can you cluster these terms and group them according to their meaning? Each term must only go into one cluster and must not be repeated. I want you to use all the terms from the list in this task.
{List of 197 distinct labels in random order}.

A.3 List of text segments in French annotated by at least 14 different participants

'lourdeur de la logistique et des règles',
'ne sont pas inclus',
'points de synchronisation',
'tissage des liens sociaux',
'collaboration et le partage d'informations et d'idées',
'en-dehors des temps de travail',
'on fait quoi maintenant',
'ne sais pas trop',

’peut-être dû réfléchir un peu’,
’sens’,
’un peu perdus’,
’on ne sait pas vers où, ni pourquoi’,
’malaise général’,
’pas promus’,
’absolument’,
’garder autant’,
’fallait-il pas simplifier’,
’donner du sens’,
’dans les hiérarchies supérieures il n’y en a pas assez’,
’numérique’,
’les écrans ont raison de notre bien-être’,
’Se reconnecter à la nature, à notre environnement, aux humains’,
’moi’,
’primordial’,
’essentiel’,
’revenir à des choses simples’,
’reconnecter’,
’non pas’,
’des robots’,
’nos émotions’,
’réponse différente’,
’tu as certainement un problème hormonal’,
’gêne occasionnée’,
’réponse différente’,
’gêne’,
’interrompre le travail’,
’nous sommes très heureux’,
’déshumanisé’,
’perso’,
’nauséabonde’,
’très froid’,
’heureuse’,
’retrouver mes collègues’,
’perdu une part de la bonne humeur, de l’ambiance’,
’échange convivial’,
’informel’,
’un mois avant le salon’,
’et pas 3 jours avant’,
’imposées’,
’très importantes’,
’surcharge de travail’,
’laisser dans le flou’