# Experimental Inference of the Basic Level from Crowd Work

A method to annotate taxonomies with basic level information.

**Tom Humbert**

A thesis presented for the degree of
Master in Information Science

Faculty of Science
Vrije Universiteit Amsterdam
The Netherlands
xx/10/2022

# 1 Introduction (1.5p)

Basic Level categories are the very first categorizations made by humans and the most general categories that are specific enough to describe an entity[1]. These words are also the first learned by children and the most important to communication, as they grant a cognitive advantage. Lexical taxonomies that categorize entities into hierarchies start at the most inclusive category and end in the most specific category to describe an entity. Basic level categories have been proven to be an inflection point in these hierarchies; these categories carry more useful information in comparison to their hierarchical super- and subordinates. The measurable advantage gained from communicating on the basic level is called the *basic level effect* (BLE).

More recently, methodologies to classify categories as basic level have been developed[2, 3, 4, 5]. Despite sharing the same goal, the methodologies already diverge at the initial data collection step, when selected categories need to be classified. Hollink et al. established a Gold Standard data set of 518 threefoldly manually annotated categories, used to train a binary classifier[2]. This Gold Standard is an important artifact to this paper, it serves as the starting point of this work. In contrast, Mills used the Amazon Mechanical Turk service to hire crowd workers to annotate 11'221 categories. Then, a rule-based heuristic classification scheme was established to determine basic-levelness of categories[6]. In this work a third approach, also employing crowd work, is proposed to create basic level annotations. Instead of asking directly which category is thought to be basic level, a psychological experiment infers annotations by capturing the BLE through timed responses.

Inter-personal differences in regard to the BLE have been proven to exist, introducing the concern that a few annotators may not be able to correctly determine basic-levelness. A further concern regarding the annotation process is that annotators follow guidelines when in doubt, having time to research entities unknown to them. This could conflict with Rosch's theory, as basic-levelness is partly defined in accordance to the exhibition of the BLE[1]. Even if a category could logically be described as basic level, it still might not trigger people's BLE. Furthermore, the act of researching and contemplating on a category during an annotation task could result in a form of specialization. Specialists in a domain often exhibit the BLE for subordinate categories[7, 8].

The motivation driving this work is that a classifier cannot carry the bias of a few expert annotators if its output should activate the BLE for many. The aim of this work is to establish a methodology to discern the basic level – while limiting bias influencing the trained model – through the inherent knowledge people have about the basic level. A selection was made from five category branches of the WordNet taxonomy, previously annotated by the teams of Hollink and Henry[2, 4]. This selection was used to create visual stimuli for one of Rosch's original experiments, originally meant to prove the existence of the basic level effect[1]. In an iterative approach, the experiment was adapted to be ran online without researcher interaction, in order to measure participants'

BLE through reaction times. The experiment was built with PsyToolkit[1] and conducted on the crowd-sourcing platform Prolific[2]. The hypothesis is that an experiment using known basic level categories to determine the existence of the basic level effect in humans, should inversely also be suitable to determine the basic-levelness of a word from the measured basic level effect on humans. Two evaluation schemes, a simple and a more advanced, are subsequently suggested to evaluate the recorded reaction time and establish a new set of annotations. The research questions that are answered by this work are;

1. How would a psychology experiment – meant for the inverse purpose of proving the existence of the basic level – be reconstructed for crowd-sourcing annotations of a large data set with respect to the basic level?

    1.1 How are the crowd workers' accuracies influenced by the experiment setup and how can those influences be limited?

2. How is the performance of the crowd-sourcing annotators evaluated w.r.t. the basic level when large inter-annotator differences can be expected?

    2.1 To what extend do the experiment results differ between native and non-native speakers?

For the sake of reproducibility, the processes of selecting candidate categories, building the corresponding experiment and evaluating the responses have been captured in code. The code and the results of the experiments can be accessed via GitHub[3].

# 2 Related Work (3.5p)

Basic level studies represent a convergence of multiple fields of research. The basic level was defined by descriptive and exploratory experiments in the field of psychology[1, 9]. In linguistics, categories were identified as basic level through heuristic rules based on lexical features[6]. Computer Sciences showed that the basic level can be automatically classified[2, 5]. Here, Princeton University's WordNet often acts as the backbone, providing a hierarchy of 117'000 categories([10]). Lastly, a focus on crowd-sourced annotations warrants an insight into best practices within these online work environments.

## 2.1 Basic Level Theory

Humans categorise entities by perceived likeness. Categories go from general to specific, such as *flora* and *fauna*, *fruits* and *furniture* or *kitchen chairs* and *garden chairs*. Grouped entities often share co-occurring visual and actionable attributes, serving as cognitive cues that make people recall a mental image. A

[1] https://www.psytoolkit.org
[2] https://www.prolific.co/
[3] https://github.com/tomhumbert/BL_exp_creator

category is thus defined by a set of such attributes. The basic level resides within a hierarchy of ordered entities; its superordinates are more general categories, and its subordinates are very specific. Basic level categories have in comparison the greatest amount of cognitive cues, conjuring a mental image the quickest[1].

The basic level can differ among groups of people from different backgrounds. Basic level categories are those first learned by children, necessarily taught by their care-takers and their direct environment. It has been observed that certain communities experience some categories as basic level, that others would consider subordinate. They specialize in a domain essential to their livelihood[1, 11]. I has been observed that communities that are dependent on, and very close to, nature and wildlife can easily recall specific names of plants and animals. To a certain degree, everyone necessarily becomes an expert in their specific domain, i.e. work environment[7, 8].

## 2.2 Basic Level Experiments

Experiments have shown that people, when asked to name visual stimuli, will most likely name entities at the basic level[1]. The basic level effect is thus measurable in timed categorization tasks. Rosch devised twelve experiments to expand the definition of the basic level and prove that it has a measurable effect on humans. Four 'economical' experiments capture and align visual and actionable attributes of basic level categories, establishing the basis on which further psychological experiments were built. Participants listed attributes of objects, listing the most for basic level categories. Object outlines were shown to be similar when they belong to the same category. Other experiments probed the basic level effect [BLE] in humans. Most experiments required direct interaction with participants.

Due to their participatory nature, three experiments on the BLE are especially relevant to this work. They require adult participants and are timed or they record accuracy as a measure. One experiment presents two images side by side, masked by colored screens and obfuscating shape cut-outs. Participants were primed with a category (i.e. the category was named out loud), then decided whether the named object was located on the left or right side. Answers were more accurate when participants were primed with a basic level category. Another experiment presents two images side-by-side; either the same object is shown twice or two different objects are shown. All stimulus pairs are shown twice, once with priming, the other time without. Both accuracy and reaction times are recorded. It was found that the priming with basic level names of objects decreased reaction times, proving a cognitive advantage, namely the BLE. A subsequent experiment shows a single image at a time, priming with categories at different levels. Of 18 images, each is shown twice. Once, giving a correct category, another time an incorrect one. Participants decide quickly, using two buttons, if the image corresponds to the category they are told. It was shown by a two-way ANOVA of the reaction times, using the level and name of categories as factors, that reaction times for basic level priming were, in both the false and true cases, significantly faster[1].

## 2.3 The WordNet Lexical Database

The vast taxonomy of Princeton University's WordNet, seen as a network graph, comprises 117'000 nodes[10]. Each node represents a group of synonymous words, called a *synset*. Synsets have conceptual relations such as hyperonymy and hyponymy. A hypernym of a synset can be equated to a superordinate to a category, a hyponym to a subordinate. Starting from the least specific synset *'entity'*, the beginning of the taxonomy of all that is perceived or known, many paths can be taken to leaf nodes, representing the most specific categories. As such, this hierarchical structure is very compatible with, and useful to, basic level research. The basic level will reside somewhere within such a path. Moreover, WordNet provides software packages for the Python and R programming languages, allowing to integrate it easily into software projects[12].

## 2.4 Basic Level Prediction

The basic level theory shows great potential to improve communication. Knowing which words will foster understanding could be of great use, but manually determining the basic-levelness of every word – once per target group – seems unfeasible. A filtering system, based on heuristics that identify basic-levelness of words, e.g. using lexical features, was shown to reliably classify words[3]. Mills used eight filtering and four voting rules to identify categories as basic level within 24 domains (partial WordNet taxonomies), the 184 categories were taken from the manually labeled Rosch and Markman[9] data sets and were aligned to the WordNet taxonomy. Compared to the manual annotations, the system achieved an overall accuracy of 94%. Chen et al. extended the category selection by words from the Chinese WordNet counterpart[5]. 433 categories, in Chinese and English, taken from the Rosch and Markman collections were used to train classifiers. The most promising classifier was directly compared to Mills' system on the same data, both approaches achieving an accuracy of 86%, but with a higher recall and f-score than Mills' system. Hollink et al. classified synsets using features such as the depth in the WordNet graph, the length of the lexical glossary or the Google Ngram. A random forest algorithm achieved an accuracy of 82% and a Cohen's $\kappa$ of 0.61[2].

Variations in the selection of categories and annotation processes can lead to differences in classification performances. In many cases, data scientists will personally annotate the training data with the correct class, such that the classifier learns accordingly. Mills and Chen used previously established annotations of Rosch and Markman[3, 5]. Mills additionally hired crowd-workers on the MTurk platform to annotate categories[6]. This crowd-working task asked workers to select the basic level category within synset paths of varying lengths from superordinate to subordinates, following annotation guidelines. Hollink et al. manually annotated 518 categories from 3 partial WordNet taxonomies (domains). An inter-annotator agreement between the three annotators of Krippendorf's $\alpha = 0.72$ was achieved. These annotations were further extended with 468 categories from two additional domains in Henry's work, also annotated by

three annotators[4]. The total of 986 three-fold manually annotated categories, available for reuse, presents an excellent basis for further research.

## 2.5 Timed experiments on crowd work platforms

This section highlights findings about the conduct of experiments in an online setting and the quality of crowd work. First, it should be noted that the worker populations on popular crowd-working platforms, such as MTurk[4] or Prolific [5], were found to have a low diversity of workers in terms of age, gender, socio-economical background or education[13, 14]. The majority population of most platforms is described as being young, white males with good command over technology[14].

Successful crowd-working tasks are interesting to workers because they offer well-planned research conduct, good task design and adequate pay. The potential personal interest of a worker, their engagement and creativity, are secondary. Nonetheless, raising the pay does not always equate in higher quality work. Suggestions by the crowd-work platform for worker compensation should be followed[13]. To keep a worker engaged, they can be given feedback about their accuracy during tasks, adding a certain learning component[13], improving the final results. The suggested maximum time for an online task lies around an hour, but 30 minutes are ideal for worker motivation retention[14].

The quality of work is influenced by multiple factors. The online environment gives workers a certain sense of autonomy and privacy. Unlike the laboratory setting, social pressure is removed from experiment participants. This is a positive feature with respect to well-being, but it can also lead to variations in the quality of delivered work. In terms of quality, Prolific's workers tend to deliver better work than Amazon's Turkers[13]. Moreover, workers will always try to do the given tasks as fast as possible, while delivering the minimum required quality of work. Still, it was shown that the overall quality of in-person and online work bears no statistically significant difference[13].

In timed classification tasks, an error rate of 51-70% can be expected from workers. Some researchers set the accuracy threshold, defining a successful task execution, to as high as 85%[14]. Mistakes are not only due to workers trying to be fast, they may also be interrupted by their environment or notifications on their electronic devices (e.g. the device they are currently running the experiment on)[14]. Additionally, Participants may lose motivation or interest during the experiment, answering randomly to finish the task quickly[13]. This random guessing can be found in the results as fast reaction times paired with a low accuracy[14]. It can also be detected through comprehension checks, so-called honeypots, dispersed throughout the study.

[4]https://www.mturk.com/
[5]https://www.prolific.co/

# 3 Methodology (3p)

In this section, the process to create annotations through measurements of a psychological experiment is presented. First, an experiment was selected and adapted for an online environment. A data set of categories and images was assembled through a personally developed software tool. Then, the experiment was run on a crowd-worker platform. Then, the experiment's validity is established. Lastly, the results, i.e. the reaction times of the workers, are turned into annotations.

The research is conducted with a focus on social and technical sustainability. The social sustainability dimension is characterized by the responsibility of conducting psychological experiments. The conduct of the experiment is aligned with a set of ethical guidelines, partially stemming from the EU's GDPR. The well-being of the experiment participants must be ensured through-out the research, they must be aware of factors that could be cause for discomfort. Crowd workers are adequately paid and are well informed about the task and the handling of the collected data. Finally, the participants enjoy complete anonymity at all times.

The technical sustainability dimension of this project is concerned with verifiability and reproducibility. This is motivated by the 'replication crisis' in which cognitive psychology researchers have found themselves in recent years[15, 14]. Results should not be under-reported, results are shared even if they do not support the hypothesis. In addition, to facilitate the reproduction of this research, software tools are developed that partly automate, or at least simplify, the workflow. The software tool to select category candidates and generate the experiment is described in section two. The evaluation of the experiment results is entirely scripted, it can be used for further trials without the need for modification. The entire code (including experiment results) is available on GitHub and is adequately documented.

## 3.1 Experiment Adaptation

The selection of the experiment was made from Rosch's original experiments[1]. Although there have been more recent, similar experiments, they appear to also have used Rosch's experiments as a base[7, 8]. Considering the main hypothesis of this research, that an experiment that can prove the existence of the BLE could conversely be used to measure the BLE to find basic level categories, some selection criteria were established. The experiment must; have direct participation of adults to be suitable for a crowd-working environment; be reproducible in an online-only setting without direct interaction with the researcher to generate enough responses in a short time frame; use timed responses to evaluate the BLE introduced by the use of a specific category; be simple, no text input should be required, reducing time spent on the task, allowing for a higher throughput and making the crowd-working task more attractive. Only experiments six and seven (see section 2.2) comply with these conditions. Experiment seven was chosen as it only uses a single image as stimulus, further simplifying

the process. The main premise of the experiment is that participants need to distinguish true from false stimuli.

The experiment was originally conducted with 45 participants, divided into groups of 15, one group per category level (superordinates, basic level, subordinates). No participant would see the same stimulus at its three different levels. But, because the objective became to nominate a category having the greatest BLE on a participant as a basic level category, all participants are shown all stimuli in this adaptation. Rosch originally selected three objects from each of six non-biological taxonomies. As this work uses the annotated data sets of Hollink and Henry, only five taxonomies (edible fruit, furniture, garments, hand tools, musical instruments) are available to choose from. Nonetheless, this adaption should ultimately be used to annotate large data sets, wherefore at least ten objects are chosen from each taxonomy. The exact number changed between the different stages of the iterative adaptation process. Lastly, the original experiment primed participants vocally. The use of sound was avoided for the online setting as it seemed impractical and unnecessary. Furthermore, the population of potential participants would have been limited to those that have a set-up allowing them to listen carefully to sound. Participants with auditory impairments would also have been excluded. The vocal instructions were exchanged with text.

The experiment was programmed using the PsyToolkit framework. Most other experiment creation frameworks do not use code, experiments are compiled in a more visual manner. This is great for programming lay-people, but the options are limited and the platform is often demanding payment. PsyToolkit is free of use, highly adaptable and experiments can directly be conducted on their service. PsyToolkit is also compatible with Prolific. The basis of the experiment code was inspired by their example of a 'Simon Says' task, showing true and false type stimuli and timing the responses of the participants[16, 17].

## 3.2 Stimuli

At its base, the experiment requires a set of category triplets. A triplet contains a suspected basic level category, a superordinate and a subordinate of that category. As a result from using the WordNet taxonomy, a triplet also forms a path through its network. A picture is assigned to each triplet, representing an object of the subordinate category of the triplet. As an example, if the triplet is *fruit* > *grape* > *Thompson Seedless*, then a picture of the Thompson Seedless Grape is chosen. Each stimulus shown during the experiment execution shows one of the words from the triple as a title, with the according image underneath. This generates three true type stimuli. These are complemented with three false type stimuli, showing the same titles from the triple, with an image of an object from a different taxonomy. Thus, each triple generates six distinct stimuli.

### 3.2.1 Category Selection

To verify the results of the experiment, the chosen triples must be comparable to the annotations of the aforementioned data sets. The data set was filtered for categories that were annotated as basic level by at least two of three annotators. From there, about ten categories were chosen from each taxonomy. At least one chosen category per taxonomy did not have perfect annotator agreement. This would allow for interesting comparisons in terms of accuracy and reaction time. In the pilot phase, 66 triples were used, resulting in 396 distinct stimuli. In the final run, the triples were reduced to 50, exactly ten per taxonomy, resulting in 300 stimuli. This ultimately constitutes 150 annotated categories.

Subordinate Categories were selected according to self-imposed guidelines. Basic level annotated categories often had multiple subordinates at multiple depths in the hierarchy. Any of these subordinates could be chosen, one subordinate level could be skipped over to its next subordinate if that allowed for a clearer choice of object images. For example, the basic level category *orange* has an intermediate level of *sweet* and *bitter oranges*, in which case *Valencia orange*, the subordinate of *sweet orange*, could be chosen, making the image selection more precise. If a basic level category had no subordinates, it was removed from the list of candidates. How to deal with these cases can be discussed in later future work. Category names are often composed of multiple words, such as the *Muscat grape*, and WordNet synsets often list multiple versions of names, such as simply *Muscat*. The shorter term was used in most cases to reduce the time participants have to spend reading the title of a stimulus during the experiment. Furthermore, the experiment is not meant as a guessing game, testing participants on their knowledge. On the contrary, the selection from the subordinates should always fall on the (suspected) most well-known category. To avoid confusion, the chosen categories must undeniably belong to a single taxonomy. For example, WordNet categorizes *wheel chair* as *furniture*, although it would also fit the *vehicle* category. Lastly, brand names were avoided as subordinates, except for fruits, where strain and brand are often equivalent.

No selection needed to be made in terms of superordinates, they are given by the five annotated taxonomies, namely *edible fruit, furniture, garments, hand tools* and *musical instruments*. Although *garments* was changed to its more general superordinate *clothing*.

### 3.2.2 Image Selection

The images in the original experiment were all made by the researchers, each object was photographed from multiple angles. Then, the images representing the object best were selected by a vote. This process was not needed with the vast amount of images found online nowadays. Still, images had to comply with a few rules. The image must be a photography of a real entity, digitally unaltered. The photography should prominently show only the object (exceptions were made for e.g. small decor on a table, a leaf on an orange). For clothing, only the clothing item should be shown, without a human model. Fi-

nally, all images needed to be licensed such that they could be legally re-used non-commercially. The license and owner of the image were then added in the bottom right corner of the stimulus.

#### 3.2.3 BLExplorer Tool

The Basic Level Experiment Explorer Tool (short BLExplorer) is meant to simplify the process of searching through all the possible subordinate categories of a potential basic level category. It takes as input a list of basic level categories, formatted as WordNet synsets (e.g. grape.n.01) and displays all subordinates in a tree-like structure. To further improve usability, it is operated through a GUI and saves the progress a researcher has made in compiling triplets. Collected images can be added to the triple, which are then named correctly and saved in the project folder. When the selection of the experiment data set is made, the researcher can automatically generate the final experiment. This means that the tool combines all the stimuli, creates randomized false type stimuli and finalizes the experiment by generating the code and folder structure that can directly be uploaded to the PsyToolkit website.

### 3.3 Execution

The experiment is conducted three times, two pilots and one final run. The first pilot is conducted in a very early stage of development to eliminate eventual bugs and other short-comings, such as missing information for participants. The second pilot is conducted formally on Prolific. The experiment contains 396 stimuli at that point. 25 workers are hired on Prolific at an hourly rate of 9 GBP an hour, according to Prolific recommendations. The workers were paid for 15 minutes of active work (breaks in-between rating were not accounted for), resulting in a payment of 2.25 GBP. The 15 minutes were estimated by attributing 2 seconds per stimulus, 1 second to react and another second for the delay until the next stimulus is shown. The final run is conducted with 300 stimuli, but it includes a training round of 36 stimuli, one triple per taxonomy. From the results of the second pilot (see section 4.2) it was found that participants needed more time than expected. A further second per stimulus is added, resulting in an expected time of about 17 minutes per participant, leading to a payment of 2.55 GBP.

A worker is given a full introduction of the experiment both on Prolific and on PsyToolkit, such that they are aware of what they can expect and what is expected of them. Workers must finish the work seriously and in its entirety in order to be paid. Before the experiment, they have to complete a short survey asking them if English is their native language and if they live in an English speaking country. Then they move on to the experiment. The experiment starts with a short slide show, explaining the experiment and the controls. Then they start reacting to the stimuli, 20 at a time. After each package of 20 stimuli, they see their overall accuracy to motivate them. In the pilot they were also shown their fastest reaction times, but it was suspected that this leads to less

serious work. After the accuracy feedback screen, they move to a break screen on which they can rest until they continue. At the end they are sent back to Prolific through a personal link that lets the researcher know that the experiment has been concluded successfully. A visual representation of the process can be seen in Fig.1.

### 3.4 Validation

The validation of the results is done in two steps. The first step is concerned with worker accuracy. During the pilot, the accuracies were used to argue the validity of the experiment. The origins of the error rates of stimuli are evaluated by the means of several ANOVA, testing each factor (e.g. the image, the title, the according taxonomy) for negative influence in order to answer R.Q. 1.1. Furthermore, error rates of individual workers are analyzed to determine seriousness. A serious worker has either high accuracy (above 85%), or if their accuracy is below 85%, their total time to completion is within the time range of the other workers. The specific measurements of participants are also used to answer R.Q. 2.1.

Subsequently, the reaction times are used to determine basic levelness of categories. According to Rosch's theory, a participant should react the quickest in accordance to their personal basic level. Two evaluation schemes to infer the basic level from timed responses are proposed. The first simply selects the fastest mean reaction time, over all participants, of categories within a triple. The level of the other two categories are inferred from their position in the WordNet hierarchy with respect to the selected basic level. The second scheme treats the reaction time of each participant as an annotation, again, selecting their fastest response within a triple as the basic level. Then, an inter-annotator agreement is calculated using Cohen's Kappa and Fleiss' Kappa to group participants by their native language.

## 4 Results and Analysis (9.5p)

### 4.1 P1 - First Pilot (0.5p)

The first pilot was conducted in an early stage of the experiment's development to test the efficacy of the experiment instructions and the overall procedure while mimicking a crowd-work situation. Two participants of 6 spent around 10 minutes on the task, three needed about 20 minutes and one participant took 35 minutes to complete the task. They needed on average 20 minutes. The overall accuracy of all participants was 86%. Their mean reaction time was 878 milliseconds, measurements lie between 500 ms 1500ms. Keeping in mind that the sample size is not representative, all findings align with the basic level theory, superordinate and basic level stimuli are reacted to faster, mostly at higher accuracies. The fastest mean reaction time of 0.791 seconds was recorded for false-type basic level stimuli, also having the highest accuracy at

95%.The slowest were both true and false type subordinates at respectively 1.012 and 0.942 seconds. Subordinates also received the most inaccurate answers in both stimulus type groups, the false type being the lowest at 75%. As these measurements reflect what was to be expected, there is good reason to believe that further measurements with a representative sample can answer the research questions. In the survey part of the experiment, no problems in regard to understanding instructions were reported. Most described the task as being enjoyable, taking pride in fast reaction times. One participant initially did not register the indicator reporting incorrect answers during the experiment. Indicators were improved in a further iteration.

## 4.2 P2 - Second Pilot (3.5p)

Pilot two was formally conducted through Prolific with 25 paid workers without setting a restriction on worker residence. Two workers of the 25 indicated that they were native English speakers, one of which lived in an English speaking country, making the results only representative for a population to whom English is not their first language. Participants spent an average of 16 minutes on the task, the fastest spent 10, the slowest 50 minutes, the median being 13 minutes, all including possible breaks taken during the experiment.

### 4.2.1 P2 - Accuracy

For pilot 2, the accuracy is reviewed to discuss validity and to highlight why changes were made for the final experiment. Participants were fairly accurate with a mean accuracy of 88.5%. All participants stay within a 0.08 distance of the mean and they tend to be more correct as can be seen in the distribution in fig.1. Both self-reported native English speakers, participants 9 and 16, have a moderate accuracy, close to the mean.

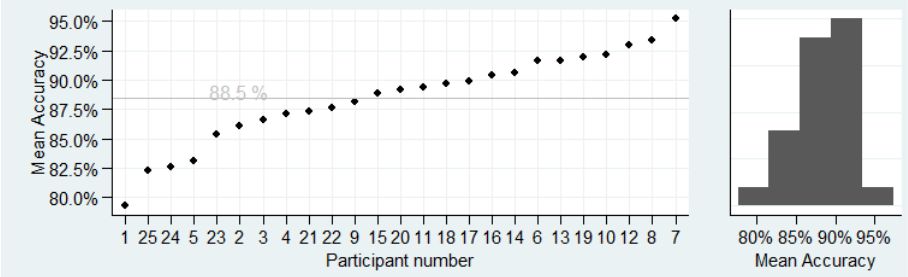


Figure 1: Left: Mean accuracies of all participants. Horizontal line represents overall mean accuracy. Right: Distribution of participant accuracies.

Participants make more errors in the beginning of the experiment, but as the experiment progresses, the error rate becomes linear for every participant as can be seen in fig.2. Participant 1 has made most errors and was also the

fastest to submit. Participant 7 has the lowest number of mistakes and was the second fastest to finish the experiment. Per participant, the relation between error count and time appears to be linear.

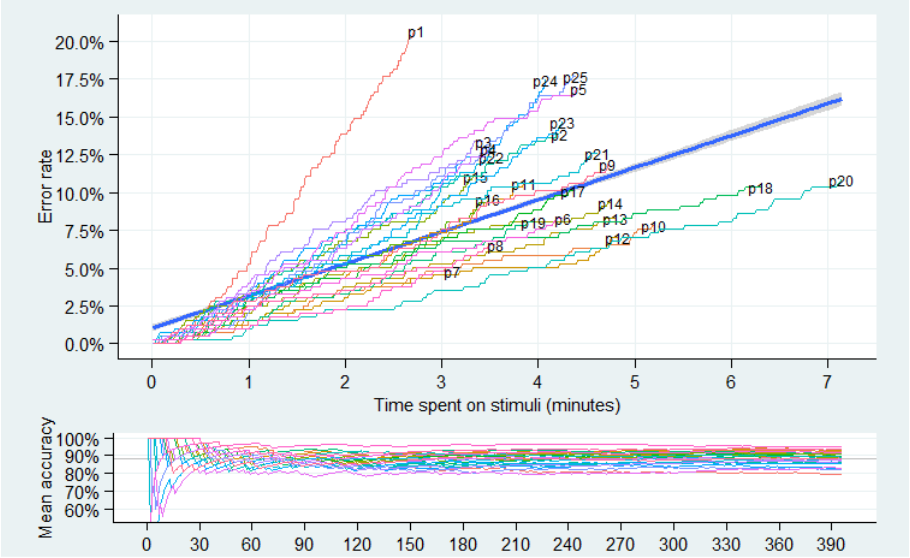


Figure 2: Top: Participant accuracy over time (considering the cumulative reaction times) and trend (in thick blue). Bottom: Participant mean accuracies over all trials.

Stimuli tend to receive correct answers, only a small fraction of stimuli receives a lot of mistakes as is shown by fig.3. Subordinates have the lowest accuracy within their respective groups of true and false type stimuli, 62% and 89% respectively. Similar to P1, it is easier for participants to distinguish false type stimuli (92% accuracy), true type stimuli have an accuracy of 84%.

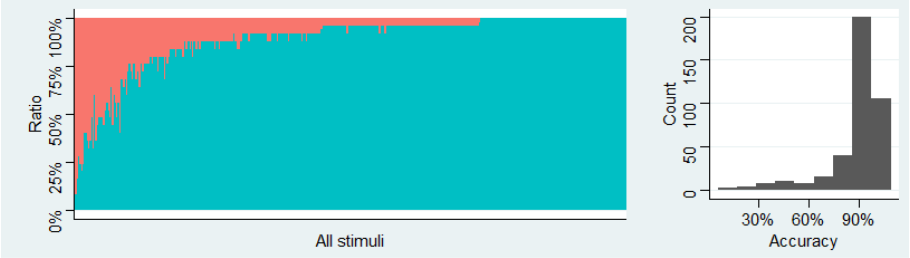


Figure 3: Left: All stimuli in terms of correct to incorrect answer ratio (blue is correct), Right: Distribution of stimuli by accuracy.

Apart from the stimulus type, the level and the taxonomic branch of the category have been found to significantly influence participant mistakes, as is

found by multiple one-way ANOVA of all experiment variables, seen in table 1. Only the difference between superordinate and basic level accuracies is not significant, shows a follow up Tukey test (table2). Branches can not be distinguished in terms of accuracy, no significant differences in terms of accuracy can be found. Except for the two branches with the highest and lowest accuracies; *hand tools* at 83% and *clothing* at 92% ($p < 0.05$).

| Component | DF | f | p |
|---|---|---|---|
| WordNet branch of the category name | 4 | 2.44 | 0.047 |
| Stimulus type (True or False) | 1 | 6.41 | 0.01 |
| Level of the category name | 2 | 12.62 | $5.98 * e^{-6}$ |

Table 1: Results of one-way ANOVA on experiment components w.r.t. accuracy.

| Pairing | diff | lwr | upr | p adj |
|---|---|---|---|---|
| superordinate - bl | −0.03700026 | −0.08126786 | 0.007267334 | 0.1216232 |
| subordinate - bl | 0.06067416 | 0.01542013 | 0.105928184 | 0.0050084 |
| subord. - superord. | 0.09767442 | 0.05205074 | 0.143298101 | 0.0000026 |

Table 2: Results of Tukey test on category level w.r.t. accuracy of answers.

#### 4.2.2 P2 - Reaction Times

The main objective of this work is to show that this experiment can gather annotations from crowd-work. Despite several errors, not present in Rosch's data, the reaction times are similar to those recorded by Rosch. But, the participants react about as fast to superordinates than they do to the basic level categories. Otherwise, the differences between subordinate and superordinate ($p < 0.5^{-4}$) or subordinate and basic level ($p < 0.005$) are significant. Removing the entries of stimuli with an accuracy below 90% does not change this fact, although it reduces the reaction times overall (see table 3).

| Stimulus type | Superordinate | Basic Level | Subordinate |
|---|---|---|---|
| True | 585.98 | 621.81 | 723.80 |
| False | 635.75 | 623.78 | 696.96 |
| True | 563.82 | 582.99 | 667.64 |
| False | 605.07 | 634.15 | 688.05 |

Table 3: Reaction times in terms of stimulus types. Top: All measurements. Bottom: Only stimuli with $> 90\%$ accuracy.

### 4.3 Final Experiment (5p)

### 4.4 A new Gold Standard (0.5p)

## 5 Discussion (1.5p)

## 6 Conclusion (0.5p)

## References


[1] E. Rosch, C. B. Mervis, W. D. Gray, D. M. Johnson, and P. Boyes-Braem, "Basic objects in natural categories," *Cognitive Psychology*, vol. 8, no. 3, pp. 382–439, 1976.

[2] L. Hollink, A. Bilgin, and J. van Ossenbruggen, "Predicting the basic level in a hierarchy of concepts," in *Proceedings of the Metadata and Semantics Research Conference*, Mar. 2021.

[3] C. Mills, F. Bond, and G.-A. Levow, "Automatic identification of basic-level categories," in *GWC*, 2018.

[4] N. Henry, "Learning the basic level from text: Studying different corpus characteristics in predicting the basic level," Master's thesis, University of Amsterdam - FNWI, Amsterdam, NL, July 2021.

[5] Y. Chen and S. Teufel, "Synthetic textual features for the large-scale detection of basic-level categories in english and mandarin," in *EMNLP*, 2021.

[6] C. Mills, *Labeling and Automatically Identifying Basic-Level Categories*. PhD thesis, University of Washington, 2018.

[7] J. W. Tanaka and M. Taylor, "Object categories and expertise: Is the basic level in the eye of the beholder?," *Cognitive Psychology*, vol. 23, no. 3, pp. 457–482, 1991.

[8] K. E. Johnson and C. B. Mervis, "Effects of varying levels of expertise on the basic level of categorization.," *Journal of experimental psychology. General*, vol. 126 3, pp. 248–77, 1997.

[9] A. B. Markman and E. J. Wisniewski, "Similar and different: The differentiation of basic-level categories," *Journal of Experimental Psychology: Learning, Memory and Cognition*, vol. 23, pp. 54–70, 1997.

[10] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, pp. 39–41, 1992.

[11] E. E. Smith, "Effects of familiarity on stimulus recognition and categorization.," *Journal of experimental psychology*, vol. 74 3, pp. 324–32, 1967.

[12] P. University, "About wordnet." https://wordnet.princeton.edu/, 2010.

[13] D. B. Martin, M. S. T. Carpendale, N. Gupta, T. Hossfeld, B. Naderi, J. Redi, E. Siahaan, and I. Wechsung, "Understanding the crowd: Ethical and practical matters in the academic use of crowdsourcing," in *Crowdsourcing and Human-Centered Experiments*, 2015.

[14] N. Gagné and L. Franzen, "How to run behavioural experiments online: best practice suggestions for cognitive psychology and neuroscience," 2021.

[15] A. Franco, N. Malhotra, and G. Simonovits, "Underreporting in psychology experiments," *Social Psychological and Personality Science*, vol. 7, pp. 12 – 8, 2016.

[16] G. Stoet, "Psytoolkit: A software package for programming psychological experiments using linux," *Behavior research methods*, vol. 42, pp. 1096–104, 11 2010.

[17] G. Stoet, "Psytoolkit: A novel web-based method for running online questionnaires and reaction-time experiments," *Teaching of Psychology*, vol. 44, 11 2016.

[18] OpenAI. https://deepai.org/machine-learning-glossary-and-terms/classifier.

[19] E. Versi, "Gold standard is an appropriate term," *BMJ*, vol. 308, p. 187, 1992.

[20] L. Ehrlinger and W. Wöß, "Towards a definition of knowledge graphs," in *SEMANTiCS*, 2016.

# 7 Acknowledgements

I first and foremost thank Laura Hollink, my external supervisor, who supported this work in every step of this long-winded path. I appreciate every second you spend towards this project. I also thank Jacco Ossenbruggen, who agreed to act as my second supervisor and who shared valuable insights into the topic with me.

Furthermore, I thank my partner, who is sure to find words of constructive criticism about my writing and graphical design, making sure that some monstrosities never leave the confines of my computer storage. I also thank my family and friends, that volunteered to test early versions of the experiment, providing me with precious feedback.

# 8 Dictionary

1. Basic Level : "The basic level is the level of abstraction in a hierarchy of concepts at which humans perform cognitive tasks quicker and with greater accuracy." This level can be found at different depths of the hierarchy.[1]

2. Classifier : "An algorithm that sorts data into labeled classes, or categories of information" [18]

3. Concept : "An abstract or generic idea generalized from particular instances." , "The notion of..." (Merriam-Webster Dictionary)

4. Category : Refers to the possible name an entity could be called. The term *concept* has been used synonymously in literature [6]. W.r.t. WordNet; 'category' refers to a group of synonyms, 'category name' refers to one specific word of the group.

5. Gold Standard : The Gold Standard is not an ultimate and perfect example, but "The best available [one]" and "It is constantly challenged and superseded when appropriate."[19], "Measure to which others conform or by which the accuracy of others is judged." (Oxford English Dictionary)

6. Knowledge Graph : A knowledge base using graph-structured models to integrate data. It is used for knowledge representation, reasoning and inference.[20]

7. Knowledge Organisation System (KOS): A collective term to describe different kinds of classification schemes, knowledge graphs and taxonomies.

8. Synset (Wordnet) : A synonym set; A set of words that are interchangeable in some context without changing [the initial meaning].[12] It is also the name of the according structure in Wordnet's Python library.

9. Taxonomy : "A system by which categories (concepts) are related to another by means of class inclusion."[1]

# 9 Appendix

## 9.1 Attributions

## 9.2 Figures

## 9.3 Tables

## 9.4 Code Examples