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
24/02/2023

Abstract

The basic level is a psycho-linguistic concept, describing words that are readily understood by people. Words in lexical taxonomies have been labeled manually and automatically with classifiers. We propose a novel method, inspired by psychological reaction time experiments. We built two experiments and conducted them online with crowd-workers. Results are evaluated to assess our approach and to label words as basic level according to measurements.

Contents

1 Introduction

The understanding and generation of human speech by AI has become important to many software applications (e.g. speech-recognition, text-to-speech or computer vision). At the core of these software applications are knowledge organisation systems (KOS) that organize human vocabularies in hierarchies of class inclusion. The hierarchies are constructed such that very broad categories, such as *'edible fruit'*, contain finer-grained categories such as *'apple'* that again contain more specific categories (e.g. Granny Smith). One such hierarchy of categories is WordNet, Princeton University's lexical taxonomy [1]. See figure 3 in section 4.1 to see a small part of the WordNet hierarchy. Software applications must choose words from these KOS that optimize the human-computer interaction. As an example, take Microsoft OneDrive's automated categorisation of photographs. If a high-level, broad category such as 'building' is chosen, then the collection will contain for most people a rather big and random collection of photos. If the category was more precise, such as 'library' or 'museum', the collection size shrinks and the content may become useful (e.g. to people that enjoy museums). The choice of the category at the right level decides in this case directly over the size of the resulting collection and its usefulness to users.

As humans, we usually know intuitively what to call things. This phenomenon has been studied by Eleanor Rosch in 1976 [2]. She demonstrated that in lexical taxonomies, there exists a level of hierarchy at which words (categories) reside that are most important to human understanding of speech. She called it the Basic Level (BL), and further named the induced improvements on cognitive processing, when using BL categories, Basic Level Effects (BLE). In-between sub-taxonomies, the depth at which the BL can be found varies, yet within sub-taxonomies, for example that of *fruit* or that of *vehicles*, the BL seems to reside at the same depth. In psychological experiments, she showed that people, when asked to, are most likely to categorize objects at the BL. BL categories (BLC) are also the first words children learn to name things [2].

Labelling categories in lexical taxonomies as BLC or not is the endeavor of multiple AI research projects, as it promises to further improve human-computer interaction. This goal has been approached by training automated classifiers [3, 4, 5], by establishing heuristic rules [6] or by hiring crowd-workers to pick and choose the BLC from taxonomy branches according to guidelines [7]. The works of Hollink et al. [3] and Henry [4] particularly stand out because of the large number of manually labeled, openly available WordNet categories, further described in section 2.2.

The previously described methodologies can all be considered top-down approaches. This means that the properties of BLC have been determined a long time ago and are now used to build methods to classify categories as BL or not. We believe that, especially when data is labeled manually according to guidelines, there is a possibility for bias to influence the results. Research shows that some people show BLE for categories subordinate to the BL [8, 9], e.g. a bird specialist correctly identifies a specific bird species within the time needed

by a non-specialist to identify the creature as a bird. This specialization could influence annotators to forego guidelines. Furthermore, the labels that are generated by automated classifiers must be controlled for correctness. Eliminating possible bias and incorrect labels requires many more annotators to manually check data sets; a laborious, lengthy, and therefore expensive, task.

In this work, we present an alternative, novel bottom-up approach to identify BLC. We consider it bottom-up because, similar to how the existence of the BL was proven, we identify BLC from reaction time measurements in a psychological experiment. The scope of this approach, the hypothesis it is based on and the research questions we want to answer are described in the following section.

1.1 Scope and Research Questions

We base our work on the hypothesis that **an experiment meant to determine the existence of BLE in humans may inversely be suitable to determine BLC in lexical taxonomies**. The research questions that guide our work are:

RQ 1: To what extent can Rosch's results be reproduced with crowd-workers?

RQ 2: What are possible causes for inaccurate answers?

RQ 3: What is the effect of a training phase on the experiment?

RQ 4: How do results differ between experiments that had varying guidelines and study populations?

RQ 5: To what extent can BLC be inferred from BLE?

Our main contribution is a methodology (i.e. pipeline) to label BLC within lexical taxonomies, such as WordNet. Yet, to test our methodology we made further contributions; 1) We (re-)created a psychological experiment in PsyToolkit to probe BLE, that can be run with Prolific crowd-workers (see section 3.1). 2) We developed a software tool to interactively select WordNet categories and subsequently generate the appropriate experiment code (see section 4.2). 3) We gathered a data set of reaction times toward WordNet categories from 37 individuals (see section 5). 4) We created a R Notebook to easily evaluate said reaction times and the accuracy of answers from the experiment (It can be re-used for any subsequent experiments). 5) We devised a computation scheme to infer BL labels from BLE detected in reaction times (see sections 3.3 and 5.4). 6) We created a collection of 66 (legally reusable) images that clearly depict elements of the categories we used in our own experiments. The entirety of our contributions is available online on GitHub[1] under the GNU GPL v3.

[1] https://github.com/tomhumbert/BL_CrowdGauge

2 Related Work

2.1 Basic Level theory

The experiment we conduct online is based on research Eleanor Rosch conducted to prove the existence of the Basic Level (BL), marking the inception of this psycho-linguistic concept [2]. Historically, the first categories with which humans tried to organize the world around them were conceived because humans form a mental 'average' image of things by visual or actionable likeness. The name of such an 'average' image is considered a basic level category (BLC), such as *apple*, a category that describes all fruits that look and taste like apples. Categories can be described as a grouping of things that share co-occurring visual and actionable attributes, serving as cognitive cues that make people recall a mental image. The BL is the point in a lexical taxonomy that separates it in two parts; the superordinate categories above the BL are very general (e.g. *edible fruit*) and the subordinate categories below the BL that describe more specific instances of a BLC (e.g. a *Granny Smith* apple). The increased speed and accuracy with which people performed in Rosch's experiments, where BLC were compared to their sub- and superordinates, were considered basic level effects (BLE) [2]. We further elaborate on Rosch's implementation of the experiment in section 2.3, along with similar experiments that probed BLE.

2.2 Annotations for automated Basic Level prediction

What sparked this project is the prospect of speeding up the process of labeling BLC, a necessary step in every BL research. Rosch initially created and used six sub-taxonomies in her experiments; she created hierarchies for musical instruments, fruits, tools, clothing, furniture and vehicles [2]. Through an experiment, and the help of multiple judges, images of entities were categorized at the subordinate, the basic and the superordinate level. Each image is thus described by a triplet of categories (see figure 1); images that share the same superordinate form a sub-taxonomy. The 18 images that were selected resulted in a total of 42 labeled categories. Markman expanded Rosch's six sub-taxonomies to 20, creating a total of 140 labeled categories [10]. Later BL studies either reused the Rosch-Markman labels, aligned them to other lexical taxonomies or created entirely new labels.

The extent to which we can correctly infer BLC from BLE is judged by comparison to the following works. We compare - as no comparable results exist yet - the performance of our method to that of automated classifiers (e.g. in terms of accuracy) and inter-annotator agreements. (A measured BLE could be seen as a vote for a BLC from within a triple of categories.) C. Mills' heuristics-based filtering and voting system identifies basic level categories in his test data set with 77% accuracy by applying lexical or structural rules such as "Filter capitalized words" and "The synset has hyponyms" [6]. Mills' test, development and training data sets consisted in total of 184 categories, a combination of the manually curated Rosch and Markman data sets. Mills further investigated

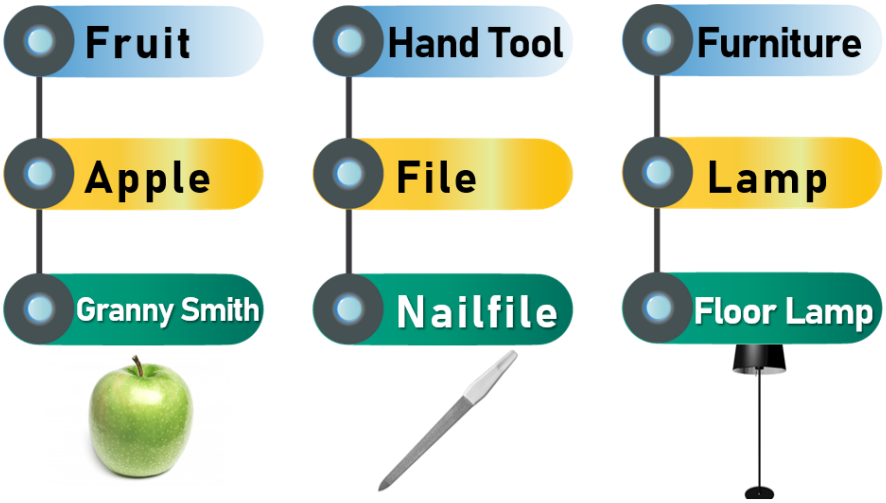

Figure 1: Three examples of triples, consisting of a superordinate category, a basic level category and a subordinate category (from top to bottom) including their respective depictions.

the possibility of crowd-sourcing annotations of categories in Princeton's WordNet[2] [7]. Crowd-workers would be given multiple lists of categories, each list forming a path from a very general to a very precise category in a taxonomy branch. Given a set of guidelines inspired by Rosch's BL theory, MTurk[3] workers selected the BL category from the lists. Only the selection of the BL category was required because the remaining labels could be inferred from the WordNet structure. Each worker (number unknown) labelled 11'221 of the 82'115 nouns contained in WordNet. They achieved an inter-annotator agreement of 92% and a Cohen's kappa of 0.61.

Chen et al. also used Rosch's and Markman's total of 184 categories, extended it and further created an equivalent data set of Chinese categories using the Chinese WordNet counterpart, resulting in a combined total of 433 categories [5]. Their most promising classifier was directly compared to Mills' system. Both approaches achieved an accuracy of 86% using the same test data, Chen's classifier having higher recall and f-score.

Hollink et al. chose to manually annotate 518 categories from 3 WordNet sub-taxonomies, representing three of the domains originally used by Rosch. An inter-annotator agreement of Krippendorf's $\alpha = 0.72$ was achieved between the three expert annotators. 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 a high accuracy of 82% and a Cohen's κ of 0.61 [3]. N. Henry and two other annotators further

[2]https://wordnet.princeton.edu
[3]https://www.mturk.com/

extended Hollink's labels by 468 categories from two additional WordNet subtaxonomies, also domains originating in Rosch's work [4]. The total of 986 three-fold manually annotated categories are available online and were reused in this work. We use these expert annotations, forthwith called Hollink-Henry labels, to compare our results to. Part of this labeled taxonomy can be seen in figure 3 under section 4.1.

2.3 Experiments on Basic Level effects

E. Rosch et al. expanded the definition of the basic level and proved that it has a measurable effect on humans in a series of twelve experiments [2]. They asked participants to list attributes of objects, calculated within-category and between-category object outline similarity, conducted timed trials with visual stimuli and analyzed the language patterns of children and of the American Sign Language. In a same-different matching task, participants reacted significantly ($p < .05$) faster when the BL category of the stimuli picture was known in advance. A mean of 554 ms was recorded for these trials. The object recognition task - which we adapted - tested 45 participants on whether entities are recognized first as members of their BL category, i.e. if the BL grants a cognitive advantage in the form of faster reaction times. Participants were separated into three groups, one per category level, and were prepared by running twice through all stimuli in a training round before reaction times were finally measured, the category name was given 500 ms before the picture was shown. A significant difference ($p < .05$) between RT for different category levels was found. A Tukey showed that BL categories are recognized faster, followed by superordinates, then subordinates, but no significant difference was found between RTs of false type stimuli of subordinates and superordinates. True and false type BL stimuli had respectively means of 535 ms and 578 ms. As a comparison, subordinate mean RTs were 659 and 642 ms, superordinates were 591 and 630 ms. The mentioned experiments both had very low error rates, 2.7% and 2% respectively [2]. We think that we can reuse this experiment to measure potential BLE in order to determine BLC.

The same experiment has been adapted e.g. by Tanaka et al., Johnson & Mervis and Rogers et al. to study the effects of domain specialisation [8, 9, 11]. Johnson & Mervis' adaptation compared the performance of bird experts with that of novices in the domains of birds, dogs and vehicles [9]. They reduced training to 8 stimuli, but gave participants the list of subordinate categories beforehand and gave 2500 ms to register the category name during the experiment. The mean RTs for BL stimuli ranged from 1300 to 1600 ms depending on the participant group, mean RTs of super- and subordinates were substantially different between domains (e.g. novice group mean RTs for; birds: 2100ms, vehicles: 1400ms). They recorded (insignificantly) faster reaction times for subordinates over BL in the vehicle domain. Their error rate was in all cases very low (max of 4%). Their data shows that experts still show BLEs, but also have a cognitive advantage for subordinate categories. We also learn that RTs differ strongly depending on the category domain. To our best knowledge, no research

has yet been conducted on the creation of basic level annotations based to the exhibition of BLEs in such timed trials.

2.4 Reaction time experiments with visual stimuli

Reaction times can be influenced by many factors, some pertaining to the stimuli, but most to the individual. Kemp found that visual stimuli are registered slower than auditory ones, the information needs 20-40 ms to reach the brain [12]. Jain et al. found in simple visual RT tasks that reaction times of young, healthy individuals range between 180 to 260 ms, averaging at a mean of 239 ms for physically fit people (248 ms for sedentary lifestyle) [13]. A significant difference ($p < 0.001$) was discovered between male and female genders, except when both groups were regularly physically active. Additionally, Otaki et al. found differences between RTs of different age groups. In simple RT tasks, young and middle-aged groups of people show no significant difference, but in choice RT tasks, young subjects (20-21 y.o.) were significantly faster ($p < 0.01$) than middle-aged people (40-55 y.o.) [14]. They also recorded slower RTs for more complicated tasks. In trials conducted with bilingual children, Tse et al. found that RTs in a Simon task rise depending on the language proficiency of the child [15]. The experiment we reconstruct is a visual choice RT task, we thus must keep in mind that reaction times can be slightly different depending on gender, age, physical fitness and language proficiency. Differences between the means of the RTs based on demographics may be significant, yet they only range within 10-40 ms. Differences in mean RT based on category levels in Rosch's experiments range between 50-130ms for true type stimuli.

2.5 Studies on/with crowd work

Martin et al. studied the population and behaviour of crowd workers in relation to scientific research. They found that crowd work on popular platforms such as MTurk[4] or Prolific [5] bears no significant difference to in-person trials in terms of quality, yet Prolific workers tend to deliver better quality work than MTurk workers [16]. Crowd-working tasks are successful when they offer well-planned research conduct, good task design and adequate pay (following platform-specific suggestions). Quality work is fostered by keeping workers engaged during the task, for example by providing intermediate feedback on their performance. Uninterested workers may do the task not seriously to quickly finish and receive payment, resulting in invalid data.

Gagné et al. found that unserious, random guessing can be found in the results as fast reaction times paired with a low accuracy [17]. They argue that an error rate of 51-70% can be expected from workers, but that some researchers set the expected accuracy threshold to as high as 85%. Workers usually stay focused well for 30 minutes, after which the quality of their work diminishes.

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

Gagné also provides insight into the crowd-working platform populations; workers show little diversity in terms of age, gender, socio-economical background or education. The majority population is characterized as young males with good command over technology [17]. We conduct our experiments with a relatively small population sample, wherefore a largely homogeneous population might prove advantageous, because we have seen that demographics influence RTs.

3 Methodology

We adapted the Rosch experiment for an online environment; our main changes are that participants are primed visually (instead of audibly) and that all participants see all stimuli instead of having three groups, one per category level. Conducting the experiment required the creation of stimuli, the coding of the experiment and the hiring of crowd-workers that participate in the experiment. To create stimuli, we manually selected 66 BLC from the Hollink-Henry labels and 66 photos that represent the BLC (see section 3.1.3 to learn about how choices were made). This step is facilitated by an in-house programmed software, mainly to properly display the WordNet sub-taxonomies. The experiment is coded and hosted on the PsyToolkit platform. Participants are hired on Prolific.

Before we launched the experiment and paid participants, we conducted a pilot with 6 (familiar) participants, to ensure that the introduction and explanations of the experiment were clear and that the experiment code on PsyToolkit did not have undetected bugs. The first experiment was launched through Prolific with 25 participants. A follow-up experiment with 12 people was launched with the intention to compare the results of the previous group of largely non-native English speakers to those of a group of native English speakers.

The second experiment introduced a training round, removed the suggestion that participants need to be as fast as possible and removed feedback within the experiment about participants' speed. Because we have changed multiple variables between the first and second experiment, we cannot directly compare their respective resulting measurements. We declare the second experiment as an alternative to the first and only analyze within-experiment measurements.

The following subsections describe how exactly we adapted the Rosch experiment, how we evaluate our measurements and lastly, we describe a simple computation scheme that is applied on the experiment measurements in an attempt to label BLC.

3.1 Adapting Rosch's Experiment

We created 396 different stimuli that each of the 37 participants encountered in the same order, in two slightly different versions of the experiment. The premise of our experiment is that participants discern stimuli as True or False while their reaction time to perform that action and their accuracy are recorded. It was chosen from Rosch's work because it can be conducted without direct

interaction of an experimenter, it collects many responses (measurements) in a short time span, it is structured around measuring BLEs and it is - of three similar Rosch experiments - the simplest to construct in terms of visual stimuli.

3.1.1 Experiment Stimuli and Task.

The visual stimuli consist of an image and a title (see fig.2). The set of titles are categories at three different depths of a taxonomy branch. We used the Hollink-Henry labels as our Gold Standard to determine our candidate BLC, see section 3.1.3 for an explanation of how we chose candidate categories. Per BLC, we select one subordinate category and an image representing it; superordinate categories are given by the Hollink-Henry labels. The combination of one category and an image forms a stimulus. Stimuli are created such that they are either True or False. Each image will appear in six distinct stimuli, three True type stimuli and three False stimuli for each of the three hierarchy levels. False type stimuli are constructed by pairing images with wrong categories chosen from a different sub-taxonomy.

As seen in 2, a participant, after reading the instructions, will rate packages of 20 stimuli, then get a break with feedback on their performance, before moving on to the next package of 20. This design, of including breaks and feedback, is inspired from the literature on crowd-work [16, 17]. Each participant encountered all stimuli in the same random order. Originally the experiment primed participants vocally [2], another adaptation of this experiment from the literature showed the title a few milliseconds before the image [9]. We opted to display both the category and image at the same time. A 600x400 pixel view-port shows the stimulus. The time until the participant either presses the 'A' (false) or 'L' (true) key (reaction time) is measured from the moment the stimulus is displayed. Our button layout was adopted from a PsyToolkit tutorial [18]. In hindsight we would have liked to only use one (the strong) hand of the participant. After each button press, an indicator tells the participant if their answer was accurate, then the screen is emptied and the next stimulus is shown after a short delay.

3.1.2 Crowd-working task configurations.

The first experiment suggested that participants should answer as fast as possible, first, in the job description, and second, in the instruction slides preceding the experiment. The feedback screen showed the fastest and slowest RT and the number of correct answers. A random sample of 25 participants was hired from the entire Prolific worker population. The sample was largely homogeneous, with 88% being male, 88% being white and 80% being younger than 30, which is in line with the literature. In our additional questionnaire, only 2 participants indicated that they are native English speakers. The job was advertised to last 15 minutes at a rate of 9 GBP (the rate Prolific suggested).

The second experiment suggested that participants should answer as accurately as possible, every mention of 'being fast' was removed. The feedback

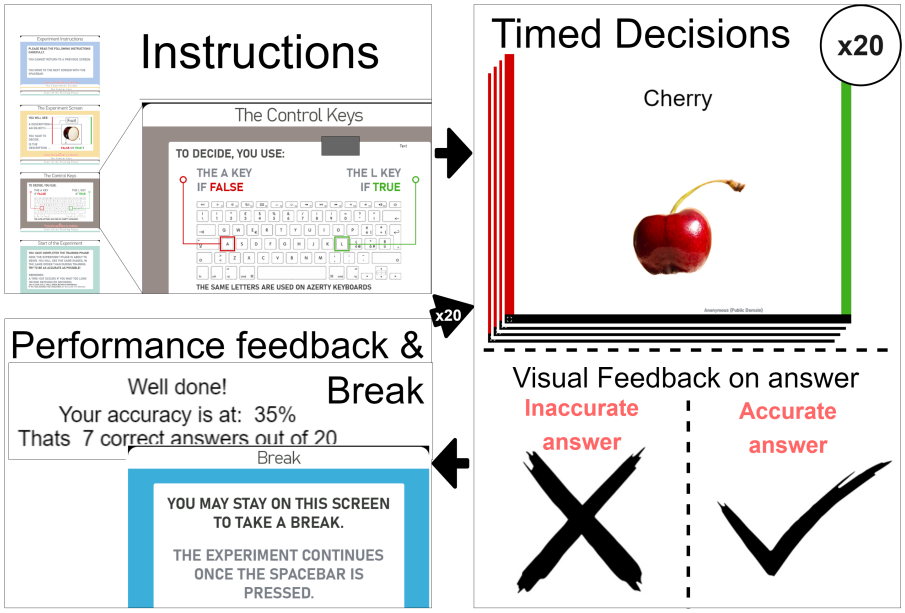

Figure 2: Experiment procedure; It starts with instruction slides, then 20 iterations of 20 timed decisions that are followed by feedback.

screen only provided the number and percentage of correct answers. This experiment had an additional training round, because the first experiment showed that participants need some time until their RT normalizes. The training round is an exact copy of the experiment round. This time a balanced sample of 12 participants was requested from the British Prolific population. The participants were ultimately balanced in terms of sex and age, yet were to 83% white. Two participants indicated that they are not native English speakers. The job was advertised to last 30 minutes at a rate of 9 GBP.

3.1.3 Category and Image Selection.

Our five superordinate categories are given by the Hollink-Henry labels (see section 2.2), namely *edible fruit*, *furniture*, *garments*, *hand tools* and *musical instruments*. The data set was filtered for categories that were annotated as BL by at least two of three annotators. We created 66 triplets, a triplet consisting of a BLC and its super- and subordinates, resulting in 396 distinct stimuli (see section 3.1.1 and figure 1). The amount of triplets per branch varies slightly, some branches only have little relevant cases.

Basic and subordinate level categories were selected according to the following self-imposed guidelines. Basic Level categories must have subordinates, should be commonly known and have simple names. For some sub-taxonomies,

e.g. hand tools, the notion of what is 'commonly known' was stretched in order to collect a similar amount of categories over all sub-taxonomies. Chosen subordinates may be lesser known, yet from all possible choices, it should still be the potentially best known object. Our intention is not to play a guessing game. Brand names and names made from several words were avoided if possible. Lastly, depths of hierarchy may be skipped to select a better (sub)subordinate category. The selected image must be a photography of a real entity. The photography should clearly and prominently show the entity. For clothing, only the item should be shown without a human model. Copyrights were respected and the license and owner of the image are mentioned in the bottom right corner of the stimuli.

3.2 Evaluating the experiment results

We evaluate participant accuracy visually through plotting accuracy per participant and over time and through a Tukey test to determine experiment components with a significant effect on accuracy. Reaction times are plotted by category level and over time. We reproduce Rosch's table of means, adding the median values, and ANOVA to have a direct comparison to her results. In experiment 2, we visually compare individual participant RT and accuracy means and the moving averages of their RT during training and the experiment round. We exclude data from workers whose mean accuracy is below 85% while their reaction times are faster than other's, because it signifies unserious work [16].

3.3 Processing results to create BL labels

According to theory, people show BLEs for BLC [2, 8, 9, 10, 11]. This effect was visible in Rosch's mean RT she recorded in her experiment. We devised a simple computation scheme that first calculates the mean (and median) RT averages for each true type stimulus. Per triple (a collection of three categories that each describe the same image), we select the category for which the fastest RT was recorded as our projected BLC. As an example, take the following triples with respective mean RT:

{edible fruit: 720 ms, apple: 480 ms, Granny Smith: 520 ms}
{furniture: 840 ms, bench: 640 ms, pew: 590ms}

The categories *apple* and *pew* would be chosen as BLC. Super and sub-ordinate labels are then inferred from their respective hierarchic positions around the chosen BLC. In the second triple, both the words *furniture* and *bench* will be considered superordinate. We compare the resulting labels to the Hollink-Henry labels in a confusion matrix to calculate the accuracy with which we correctly label categories.

4 Background on Tools

4.1 The Princeton University's WordNet Lexical Database

WordNet is a hierarchical taxonomy comprising 117'000 nodes, of which each represents a group of synonymous words called a *synset* [1]. Synsets are given hyperonymy and hyponymy relations. A hypernym of a synset is its superordinate category, a hyponym is its subordinate. This structure is very compatible with BL research, explaining its frequent use in studies [7, 5, 3, 4]. Our BLExplorer further uses WordNet's Python integration to query synsets and their hyponyms [19].

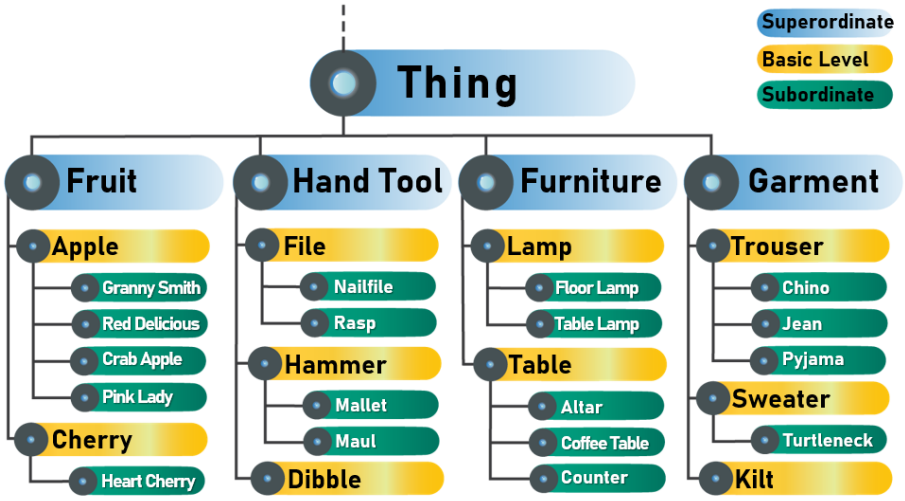

Figure 3: Partial WordNet Taxonomy, showing examples from 4 sub-taxonomies (fruit, hand tool, furniture and garment).

4.2 Our in-house developed BLExplorer

BLExplorer is a Python software application with a GUI that simplifies the exploration of WordNet categories and their sub-ordinates in tree-like structures[6]. Moreover, BLExplorer creates a project folder that gathers all the resources that are needed by the experiment code. The experiment code itself is also generated by BLExplorer. Just like Rosch, we need to create triplets consisting of a superordinate, basic level and subordinate category to be used in the experiment stimuli (see figure 1. We extracted a list of all BLC from the Hollink-Henry labels and used it as initial input. BLC from the same sub-taxonomy all share the same superordinate category. BLExplorer then displays all subordinates (and sub-subordinates) in a tree-like structure, together with their descriptions

[6]This software can be found in our GitHub repository.

(see fig.4). The user can then a) select a superordinate, b) attach a picture to the selection or c) remove the BLC from the experiment data set. Attached images are automatically renamed and placed in the correct location of the project folder. The gathered information is processed and integrated into the experiment code, which can be directly uploaded to PsyToolkit.

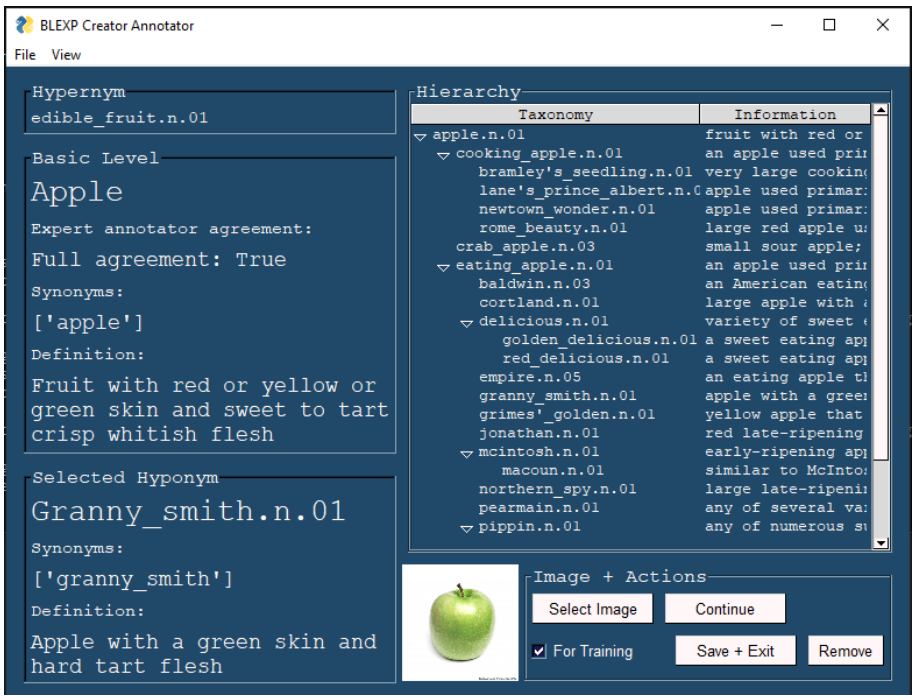

Figure 4: The BLExplorer category selection view.

4.3 The experiment programming platform PsyToolkit

PsyToolkit allows scientists to program, host and conduct experiments directly on the service's website [20, 21]. It was chosen from a list of tools compatible with Prolific, because it is free of cost and is the only to allow programming of the experiment from scratch (instead of simpler, yet less versatile approaches). PsyToolkit claims that reaction times recorded in experiments on their platform can be expected to be precise with a maximum deviation of 50 ms, due to keyboard input lag and screen refresh rates.

5 Results and Analysis

5.1 Pilot

A 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 scenario. Of 6 participants, 2 spent about 10 minutes on the task, three needed about 20 minutes and one participant needed 35 minutes to complete the task. They needed a mean average of 20 minutes. The overall accuracy of all participants was 86%. Their mean reaction time was 878 ms, ranging between 500 and 1500 ms. We found that superordinate and basic level stimuli are reacted to faster, mostly with higher accuracy. The fastest mean reaction time of 0.791 seconds and highest accuracy of 95% were recorded for the class of false-type BL stimuli. The slowest RTs were found for both true and false type subordinates with the respective means of 1012 and 942 ms. Subordinates also received the most inaccurate answers in both stimulus type groups, the false type being the lowest at 75%. The results align with Rosch's results and therefore supported the conduct of the experiment with crowd-workers. In the survey part of the pilot, no problems in regard to understanding instructions were reported. Most described the task as being enjoyable, comparing it to a game. One participant initially did not register the feedback indicator for incorrect answers during the experiment. Indicators and instructions were graphically overhauled for the two experiments.

5.2 E1 - Experiment One

In E1, the 25 participants spent a mean average of 16 minutes on the task (median of 13 min), times ranged from 10 to 50 minutes. These are the times recorded by Prolific and include the reading of instructions and breaks taken during the job.

Inaccurate answers and what may have caused them. Participants tended to give accurate answers, we recorded a mean accuracy of 88.5%. Self-reported native English speakers (participants 9 and 16) did not stand out in terms of accuracy.

Although not significant, participants visibly make more errors during the first 100 trials. As the experiment progresses the rate at which errors are made appears stable but slightly increases near the end, see fig.5. We decided to remove participant 1 from the final data set on the basis that they made the most errors while being the fastest to submit, which is a sign of unserious work. Statistics and computations following this paragraph will not include participant 1. It is possible to be both fast and reliable, as participant 7 shows, but it is not the norm.

Of 396 stimuli, 26% did not receive any wrong answers at all. Stimuli tend to receive correct answers, only a small fraction of stimuli receives a lot of mistakes as is shown by fig.6. Overall, it is easier for participants to distinguish false type stimuli (92% accuracy), true type stimuli have an accuracy of 84%.

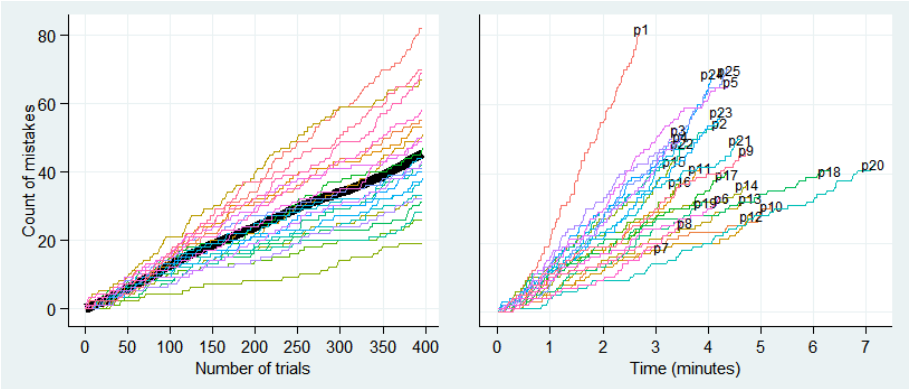

Figure 5: Left: Count of mistakes per participant over all trials. Horizontal gray line represents expected maximum mistakes. Thick black line shows trend. Right: Count of mistakes per participant over real time they spent answering (not considering breaks). The participant number is indicated at the end of each line.

Subordinates receive most mistakes for true and false type stimuli, 62% and 89% mean accuracy respectively. BLC receive more mistakes than superordinates, although not significantly. The branches with the highest and lowest accuracy are *clothing* at 92% and *hand tools* at 83%.

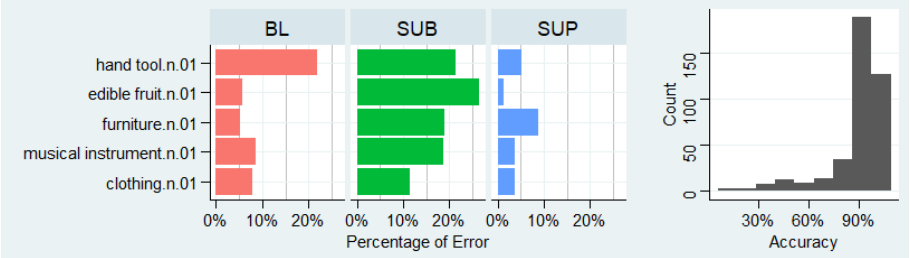

Figure 6: Left: Percentage of error per branch and category level, Right: Distribution of stimuli by recorded accuracy.

Apart from the stimulus type, the hierarchy level and the taxonomy branch of the category have been found to influence accuracy. One-way ANOVA were performed to compare the effect of experiment components on the accuracy recorded per stimulus, seen in table 1. A Tukey HSD test found that only the difference between superordinate and basic level accuracy is not significant (table2).

Participant Reaction Times. Similar to the accuracy, participants need about 120 trials until the mean reaction time stabilizes, seen in fig.7. Overall, reaction times seem to keep decreasing over the course of the experiment. Our

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.

measurements partially align with Rosch's results, participants react the slowest to subordinate stimuli, as can be seen in fig.8. ANOVA for both true and false type stimuli show that the differences between subordinate and superordinate ($p < 0.001$) and subordinate and basic level ($p < 0.001$) are significant. Our results deviate in that the true superordinates are significantly faster than true BL stimuli and there is no significant difference for the false type stimuli of these category levels. We further notice that the median is in this case a better measure for the average because of multiple extreme outliers. Table 3 presents the mean and median averages per level and stimulus type. Those averages show the same pattern even if stimuli are removed that have received many errors ($> 15\%$)

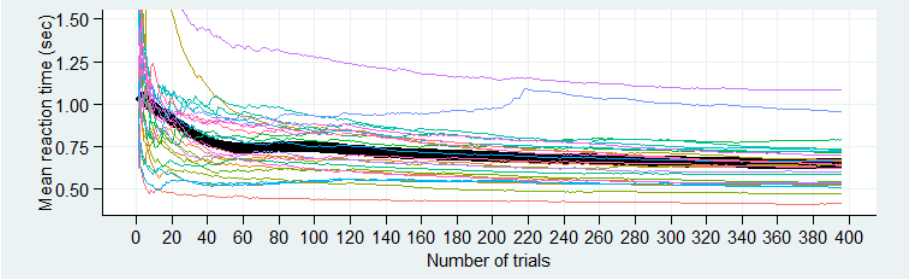

Figure 7: Rolling mean reaction times over all trials. The thick black line indicates trend, colored are individual participants.

5.3 E2 - Experiment Two

Experiment two (E2) was conducted after the results of E1 had been received and evaluated. Because E1 almost aligned with Rosch's results, we opted for multiple changes in E2 meant to improve accuracy and reaction time readings. We added a training round because both accuracy and RT needed some time to

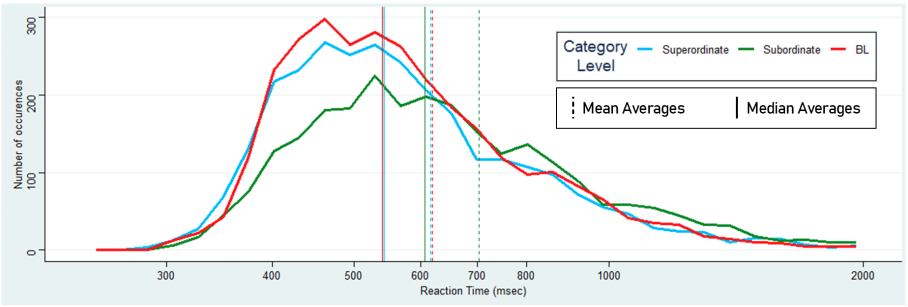

Figure 8: Distribution of reaction times per category level on a log scale, including median and mean averages for each level.

	Stimulus type	Superordinate	Basic Level	Subordinate
Mean	True	593.7607	631.2980	735.5922
Average	False	644.9381	633.1402	707.6932
Median	True	524.5	549.0	633
Average	False	567.0	553.5	613
High Error	True	563.82	582.99	667.64
Cases Removed	False	605.07	634.15	688.05

Table 3: Reaction times in terms of stimulus types. Top: All measurements. Bottom: Only stimuli with $> 85\%$ accuracy.

stabilize in E1; participants encounter the same stimuli than in E1 but twice in the same order. Participants are told to be as accurate as possible, references to being fast were removed. We further removed the information about RT in the feedback during the experiment and now only show the mean accuracy and number of correct answers. The 12 participants required a mean average total time of 31.5 minutes (median = 30.5 min), with times ranging from 21 minutes to 48.

Improvements through training. Participants in E2 were very accurate in their answers, both in the training round (TR) with 94.42% mean accuracy and in the experiment round (ER) with 97.10%. The training round significantly improved accuracy ($t = -2.78$, $p = 0.014$). Reaction times also became faster for every participant. The improvements are depicted in fig.9.

Figure 10 further shows the rolling averages of the reaction times. Reaction times strongly decrease within the first 50 trials of the training round, after which the averages keep decreasing linearly during the rest of the round. The mean average of reaction times stays almost constant during the experiment round, with only one participant showing a sudden increase in reaction times shortly before the 150st trial.

(Few) Mistakes were made. Figure 11 visualizes the errors made over trials and time during the experiment round. It shows that the rate at which mistakes are made is fairly linear, slowing down insignificantly for a few trials

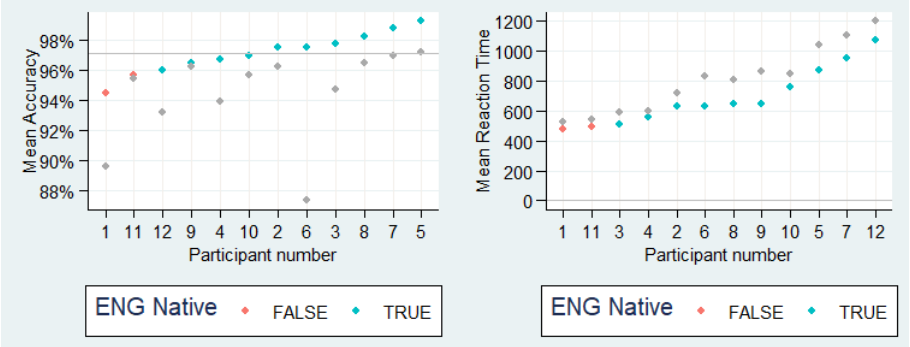

Figure 9: Accuracy and reaction time means per participant during training (grey points) and the ER. Blue points are native English speakers, red are not.

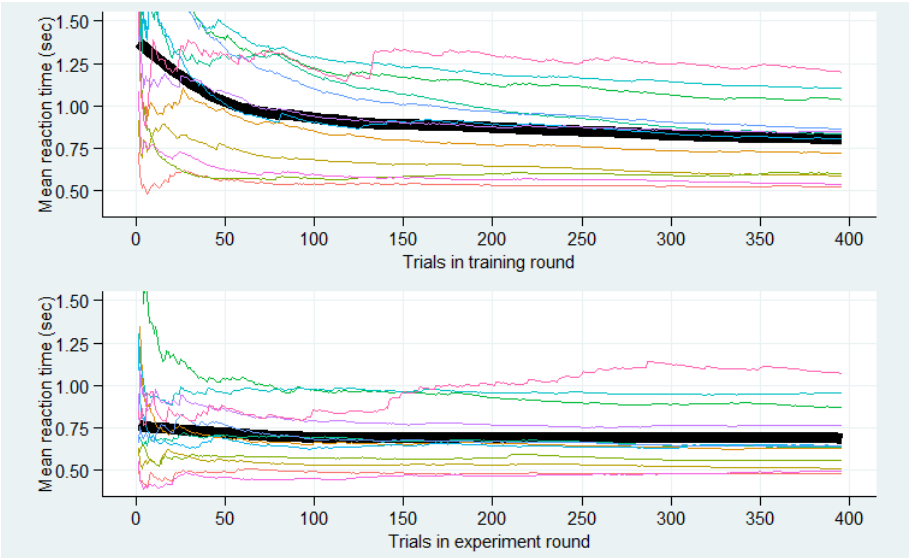

Figure 10: Mean reaction times over trials, in training phase and after. Black is the overall trend, colored are individual participants.

around the middle of the experiment. Every participant appears to have their own specific rate at which they make mistakes.

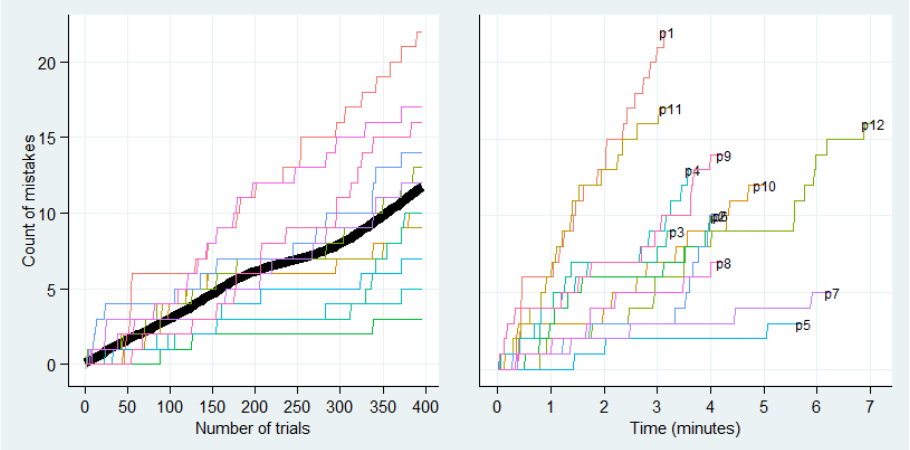

Figure 11: Mistake accumulation over trials and over time.

A series of one-way ANOVA determined that stimuli only showed significant differences when grouped by category level ($p < 0.01$) or by the category name ($p < 0.001$). A follow-up Tukey test on the category level shows that only the difference between superordinate and BL categories is insignificant. Overall, most mistakes are made for the group of subordinate categories. The group of subordinate edible fruit categories has the highest percentage of errors at 10%.

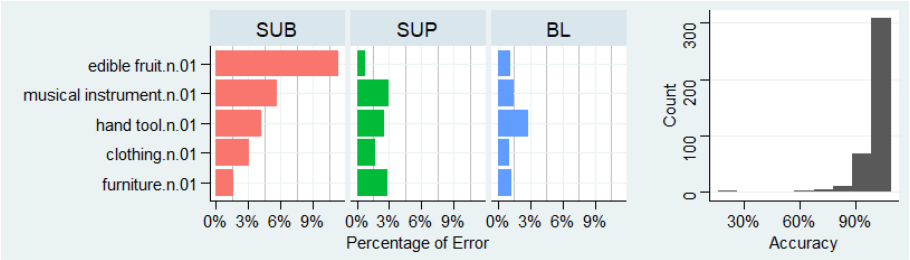

Figure 12: Left: Percentage of error per branch and category level, Right: Distribution of stimuli by recorded accuracy.

Reaction Times. The results do not reflect Rosch's results as reaction time averages for the three category levels are nearly indistinguishable, seen in fig.13. The median average appears to be a better measure in this case. The mean average of BL category reaction times is influenced by a high number of outliers, making it appear slower than subordinates, while its median is clearly the fastest (table 4). A BLE cannot be detected; the ANOVA for groups of category levels shows no significant differences.

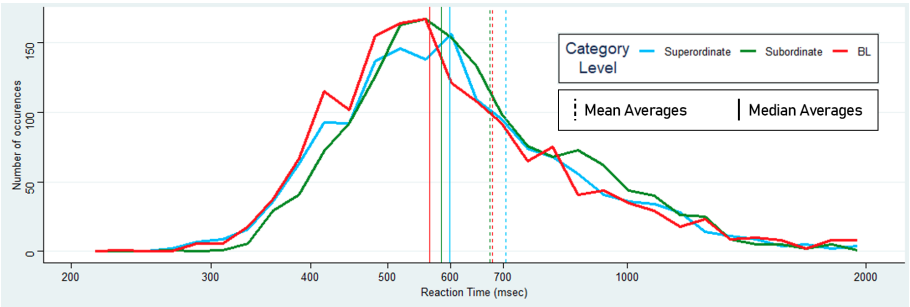

Figure 13: Distribution of reaction times per category level, including median and mean averages for each level.

	Stimulus type	Superordinate	Basic Level	Subordinate
Mean	True	653.63	637.70	647.97
Average	False	794.47	706.97	713.74
Median	True	575.0	541	554
Average	False	625.5	588	617

Table 4: Reaction time averages in ms in terms of stimulus types.

5.4 Inferring BL labels

In section 3.3 we describe a simple computation scheme; we calculate the average RT measured for individual stimuli, we only use RT recorded for true type stimuli. We choose from all triplets (described in section 3.1.3) the fastest category as our BLC. These generated labels are compared to the Hollink-Henry Gold Standard in a confusion matrix, see table 5. We present the results of using both the mean and median averages.

For RT from E1, we generate the correct labels with an overall accuracy of 61.6% using mean averages (see 5). Still, the labels we generate are not entirely random, we record a high sensitivity for subordinates (0.95) and specificity for superordinates (0.98). By design of our computation and the nature of the sub-taxonomies, superordinates and subordinates can never be confused. The confusion mostly happens between BLC and superordinates. Incidentally, we observed the same effect in ANOVA of RT and accuracy, both of which could not find significant differences between BL and superordinates. Our participants clearly show that their personal BL lies above the group of categories we chose as subordinates. Labels generated using the median are less accurate at 55% and there are even more superordinates that are labeled as BL.

For RT from E2, the correct labels are found with an accuracy of 58% using mean averages and 63% using the median. Again, no confusion between sub- and superordinates only shows that no mistakes were made during computation. This time, we record relatively high values for sensitivity ($>= 0.7$) and specificity (> 0.8) for sub- and superordinates. BLC are confused slightly less

often with superordinates than with subordinates. The results of this evaluation are inconclusive and would require a much closer inspection of those categories that have been confused frequently.

We did not (yet) compare labels from E1 and E2 to see if the same labels were confused. Although we consider our Gold Standard the absolute truth for the purpose of this experiment, the confused categories may also point at labels that could be revised. Presumably, the BL is a concept that varies from person to person and population to population. Nonetheless, the results of E1 from a largely homogeneous group of participants shows promise that our method could work, if we would consider more factors during the computation, e.g. selecting BLC on a participant basis to perform a majority vote or including accuracy in the computation as a further BLE.

Mean				Median			
E1	Ref.				Ref.		
Generated	bl	sub	super	Generated	bl	sub	super
bl	28	3	35	bl	21	7	37
sub	35	63	0	sub	37	58	0
super	3	0	31	super	7	0	28
E2	Ref.				Ref.		
Generated	bl	sub	super	Generated	bl	sub	super
bl	24	21	21	bl	29	21	16
sub	21	45	0	sub	16	45	0
super	21	0	45	super	21	0	50

Table 5: Confusion matrix of Hollink-Henry Gold Standard (Ref.) and hypothetical annotations from RT (Generated) using means and median averages for E1 and E2.

6 Discussion

We have conducted two psychological experiments on a crowd-working platform with differing configurations and compared our results to those of Rosch (and those from adaptions of the same experiment). We subsequently interpreted the RT measurements to find BLE in order to infer knowledge about BLC. We found that Rosch's description of the experiment configuration is missing details required to confidently compare our results. For example, Rosch only indicates that participants were students from the same introductory psychology class. It is unknown, yet highly unlikely, that this group represents a random and balanced sample from the general population. It is more likely that the study sample was very homogeneous. Only the results of our first experiment, whose participants belonged largely to the same demographic group (young males), partially aligned with Rosch's results with respect to BLE detection. In our second experiment, we hired a truly random and balanced, yet small,

group of participants, leading to results in which no BLE could be detected. Additionally, research shows that BLE may be different for different groups of people (e.g. domain specialists) and the additional knowledge that base RT are significantly different between groups of different age, gender and physical fitness. We therefore believe that the experiment must either be conducted with a small but homogeneous group or very large, truly random and balanced group of people.

Our experiments differ in terms of accuracy and the rates at which mistakes are made. E1 has faster RT and higher error compared to E2. Yet, they also share many similarities; error rates (mistakes made per trial over time) are largely linear, also on an participant specific level. The overall trend shows for both experiments - albeit weak - indicators for certain stages, the initial error rate slows down during the experiment (between 250-300 for E1 and 200 for E2), but picks up again around the end of the experiment. From the literature, we know that crowd-worker concentration should be expected to drop after 30 minutes of work, which is also the median average time our participants needed in E2. This leads us to believe that this may be the maximum time participants should spend on this experiment, to ensure worker concentration and therewith, valid RT.

The experiments also show similarities in terms of how accuracy is affected by groups of stimuli; especially subordinate category stimuli and stimuli from the *edible fruit* sub-taxonomy attract mistakes. Moreover, participants make twice as many mistakes for true type stimuli; it is easier to correctly point out a false type stimulus. Then, there are some categories that always attract many mistakes irrespective to the group they belong to. One such example is the *bosc pear*.

Only in E2, we recorded an accuracy close to that of Rosch ($< 2\%$) or that of Johnson & Mervis ($< 4\%$). We cannot claim that we achieved this high level of accuracy due to the additional training round, because we additionally suggested that participants should be as accurate as possible, instead of the suggestion we gave in E1 to be as fast as possible. Rosch does not declare if she asked participants to be as fast or as accurate as possible. We are of the opinion that the former must have been the case, due to the similarity of Rosch's RT mean averages to those recorded in E1. It may be possible that BLE are harder to measure when participants pay attention to being very accurate.

Nonetheless, we could observe a substantial increase in accuracy and decrease of RT moving from the training round in E2 to the subsequent experiment round. Considering that Rosch let participants train all stimuli twice before recording her results, it is likely that participants can give very accurate answers while trying to be as fast as possible. We believe it is a requirement to this experiment that participants try to be as fast as possible. In addition, the accuracy could even be used as a further factor (it is described as BLE by Rosch) in our computational scheme to infer BLC.

Using the results of E1 and E2, we have partial successes in generating BL labels. We cannot reliably detect BLC, but we often correctly find super- and subordinates. We believe that this shows that this method could be successfully

used, if the computation scheme was more refined. If the mean or median are to be preferred, or if no averages should be used at all by devising a majority vote method needs to be investigated in future work. Now, in this case the experiment was constructed knowing the Bl, but the goal is to label categories for which the BL is unknown (or uncertain). Future work could further look for which categories the BL label is uncertain, and why this would be the case. The question, at which depths of a taxonomy categories should be picked to construct stimuli must also be considered if this method was reproduced.

In conclusion, we believe that exhibitions of BLE for specific categories is strongly tied to the demographics of the population. It is for that reason that we think that when taxonomies are given BL labels, it needs to be done for several regions of the planet, for many languages, but also for global languages such as English multiple times. English speakers from different parts of the world may experience different BLE. It is also for this reason, that these labels must be generated automatically, by classifiers, by experiments such as this or by a combination of both.

7 Conclusion

We built two experiments, conducted them online with crowd-workers and evaluated their accuracy and reaction times. We were able to partially reproduce Rosch's results, we found significant differences in RT between subordinate category stimuli and the other stimuli in E1, yet we did not find a significant difference between BLC and superordinate categories. We failed to find any significant differences between levels in E2. Here, it should be noted that the median is a better measure for the average than the mean (that was used by Rosch).

Inaccurate answers were found to have many causes. Every participant has a personal rate at which they make mistakes, irrespective of the stimuli they encounter. This rate slightly increases at the end of the experiment, concentration may be influenced by the prospect of finishing soon or simply through fatigue after 30 minutes of high focus. Other mistakes are made because of the nature of stimuli. Most mistakes are made for subordinate categories, which is in line with BL research. We also found that some sub-taxonomies attract more mistakes than others (e.g. *edible fruit*) and that its easier to discern false type stimuli. Some words introduce exceptional amounts of mistakes across both experiments (e.g. the *bosc* pear).

A training phase significantly reduces the amount of mistakes participants make. The training round even appears to be necessary to record valid RT; while RT averages speed up my more than half a second over the course of the training, the rolling average RT appears to be almost constant during the main experiment phase.

E2 is, apart from the additional training round, also different from E1 in that it suggests participants to be as accurate as possible and that its participants represent a truly random and balanced sample from the British population. Be-

cause we record very high accuracy already during the training and because RT are generally slower than in E1, we believe that one of these changes between experiments must be the reason we cannot find significant differences between category levels. First, the literature shows that base RT are different for groups of people with differing demographics, and BL literature further points out that certain groups of people show different BLE because of their domain specialization. Second, we know from literature that crowd-working task requirements are taken seriously by most workers (out of fear of not being paid). The requirements of, one, being very fast, and two, of being very accurate, potentially influence RT and accuracy in the exact opposite directions. We believe that workers should be told to be as fast as possible to record strong BLE, while setting also a limit on the ratio of mistakes over time to catch cheating participants.

Inferring BLC from BLE appears to be possible in general. We have at least labeled 94% of subordinates correctly through our method. We believe that significant differences between levels of hierarchy must be found to reliably label categories, this is supported by the fact that we could not find significant differences between superordinates and BLC and as a result only labeled half of the BLC correctly. We believe that some taxonomy branches would require closer inspection and deliberation during the choice of stimuli in multiple, successive experiments. To correctly infer BLC, it is necessary to greatly improve on the computation scheme, by including accuracy for example. Future work could analyze RT as a kind of vote by participants and compute inter-annotator agreements and stimuli quality, for example by using the Crowd Truth method [22].

In conclusion, we believe that our approach should be used in conjunction with automated classifiers. The classifier would estimate at which depth of the hierarchy BLC are potentially found for a specific sub-taxonomy, then the experiment would be used to corroborate or correct the findings. An experiment only makes sense when there are multiple sub-taxonomies; the number of 400 stimuli seems to be perfect amount to keep the experiment with one training round at exactly 30 minutes, the time to retain optimal concentration. It should be noted that our stimuli were constructed manually in an image editing software, but it could be automated in code using the WordNet aligned ImageNet[7]. Ultimately, the experiment could easily be turned into a game that participants would interact with on a voluntary basis; many participants expressed that they experienced joy during our experiment and kept trying to improve their scores (values on the intermittent feedback screens).

8 Acknowledgements

I first and foremost thank Laura Hollink, my external supervisor, who supported this work in every step of its long-winded path. I appreciate the time you spent towards this project. I also thank Jacco Ossenbruggen, who agreed to act as my

[7] https://www.image-net.org/

second supervisor, and later as my primary supervisor, and who shared valuable insights into the topic of BL with me.

Furthermore, I thank my partner who is sure to find words of constructive criticism about my writing and graphical design. I also thank family and friends that volunteered to test early versions of the experiment and provided me with precious feedback.

References

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

[2] 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.

[3] 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.

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

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

[8] 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.

[9] 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.

[10] 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.

[11] T. T. Rogers and K. E. Patterson, "Object categorization: reversals and explanations of the basic-level advantage.," *Journal of experimental psychology. General*, vol. 136 3, pp. 451–69, 2007.

[12] B. J. Kemp, "Reaction time of young and elderly subjects in relation to perceptual deprivation and signal-on versus signal-off conditions.," *Developmental Psychology*, vol. 8, pp. 268–272, 1973.

[13] A. Jain, R. Bansal, A. Kumar, and K. Singh, "A comparative study of visual and auditory reaction times on the basis of gender and physical activity levels of medical first year students," *International Journal of Applied and Basic Medical Research*, vol. 5, pp. 124 – 127, 2015.

[14] M. Otaki and K. Shibata, "The effect of different visual stimuli on reaction times: a performance comparison of young and middle-aged people," *Journal of Physical Therapy Science*, vol. 31, pp. 250 – 254, 2019.

[15] C.-S. Tse and J. Altarriba, "The relationship between language proficiency and attentional control in cantonese-english bilingual children: evidence from simon, simon switching, and working memory tasks," *Frontiers in Psychology*, vol. 5, 2014.

[16] 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.

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

[18] "Simon task (as programmed by psytoolkit)." `https://www.psytoolkit.org/experiment-library/simon.html`. Accessed on 22-02-23.

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

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

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

[22] A. Dumitrache, O. Inel, L. Aroyo, B. Timmermans, and C. Welty, "Crowdtruth 2.0: Quality metrics for crowdsourcing with disagreement (short paper)," *ArXiv*, vol. abs/1808.06080, 2018.

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

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

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

9 Appendix

9.1 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. [2]

2. Classifier : "An algorithm that sorts data into labeled classes, or categories of information" [23]

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 [7]. 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." [24], "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. [25]

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]. [19] 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." [2]

9.2 Stimuli Image Attributions and Accuracy

Figure 14: All stimuli images with attribution and accuracy achieved during E1 and E2 (as True type stimuli). Page 1.

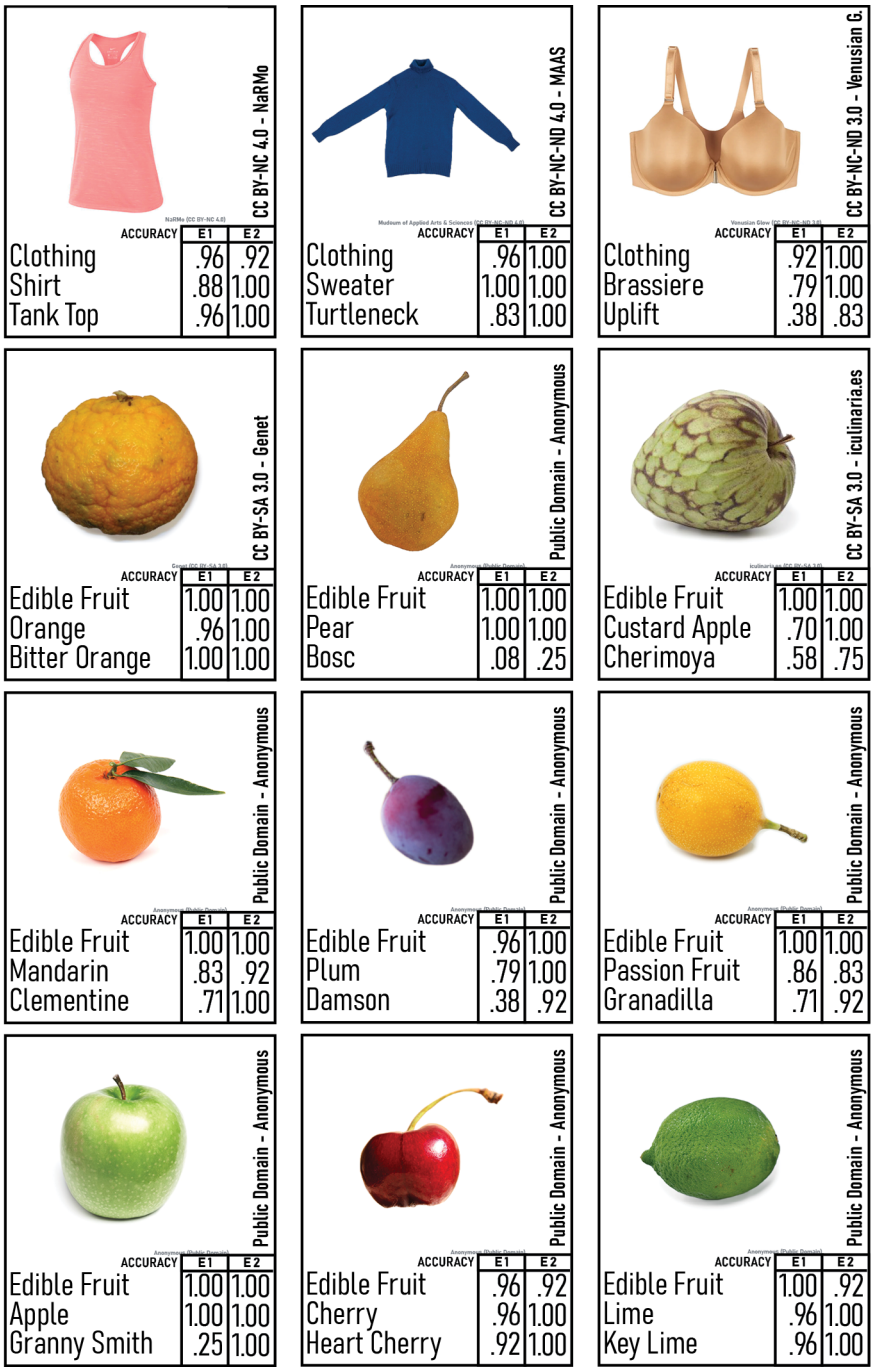

Figure 15: All stimuli images with attribution and accuracy achieved during E1 and E2 (as True type stimuli). Page 2.

Figure 16: All stimuli images with attribution and accuracy achieved during E1 and E2 (as True type stimuli). Page 3.

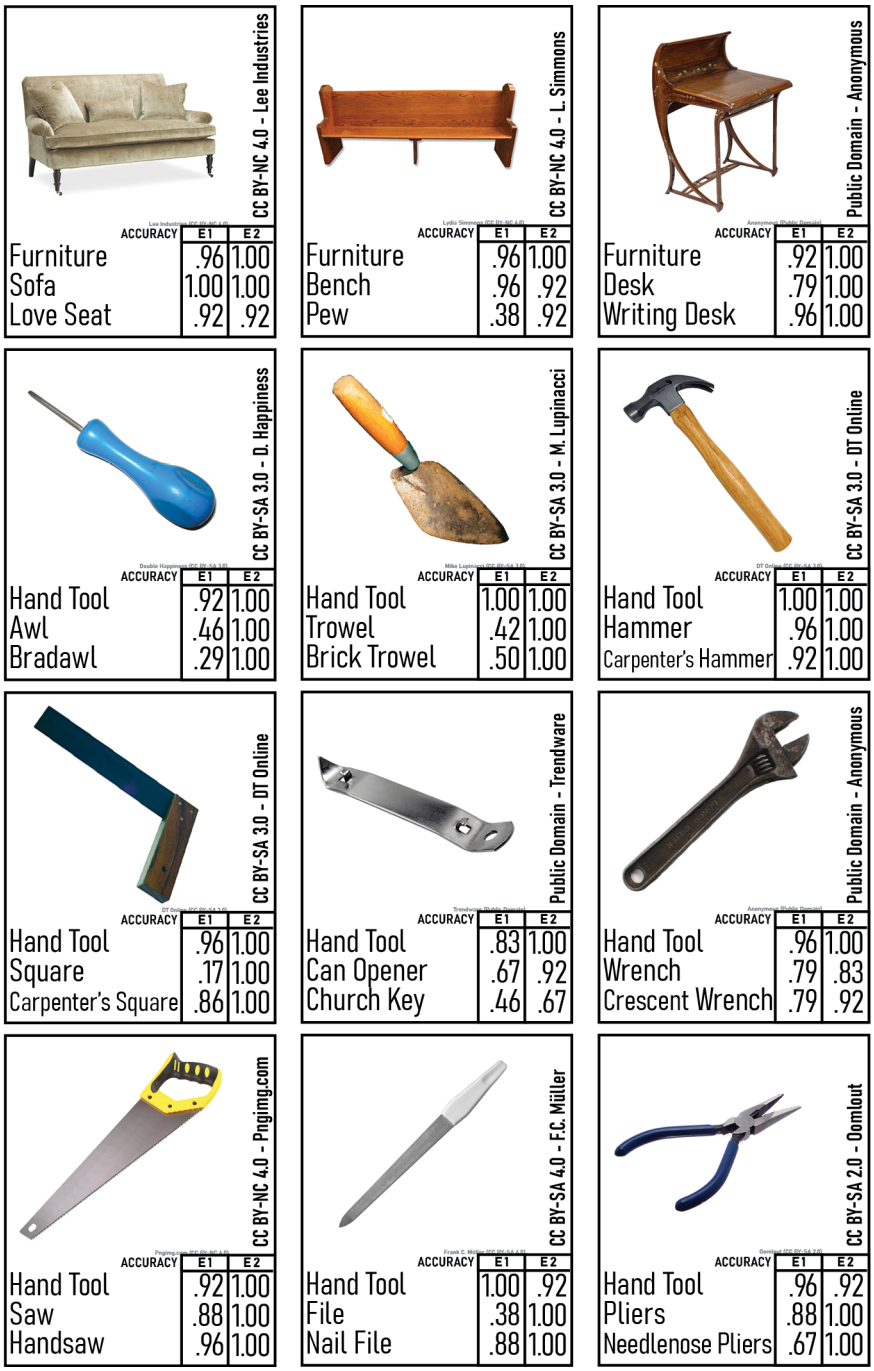

Figure 17: All stimuli images with attribution and accuracy achieved during E1 and E2 (as True type stimuli). Page 4.

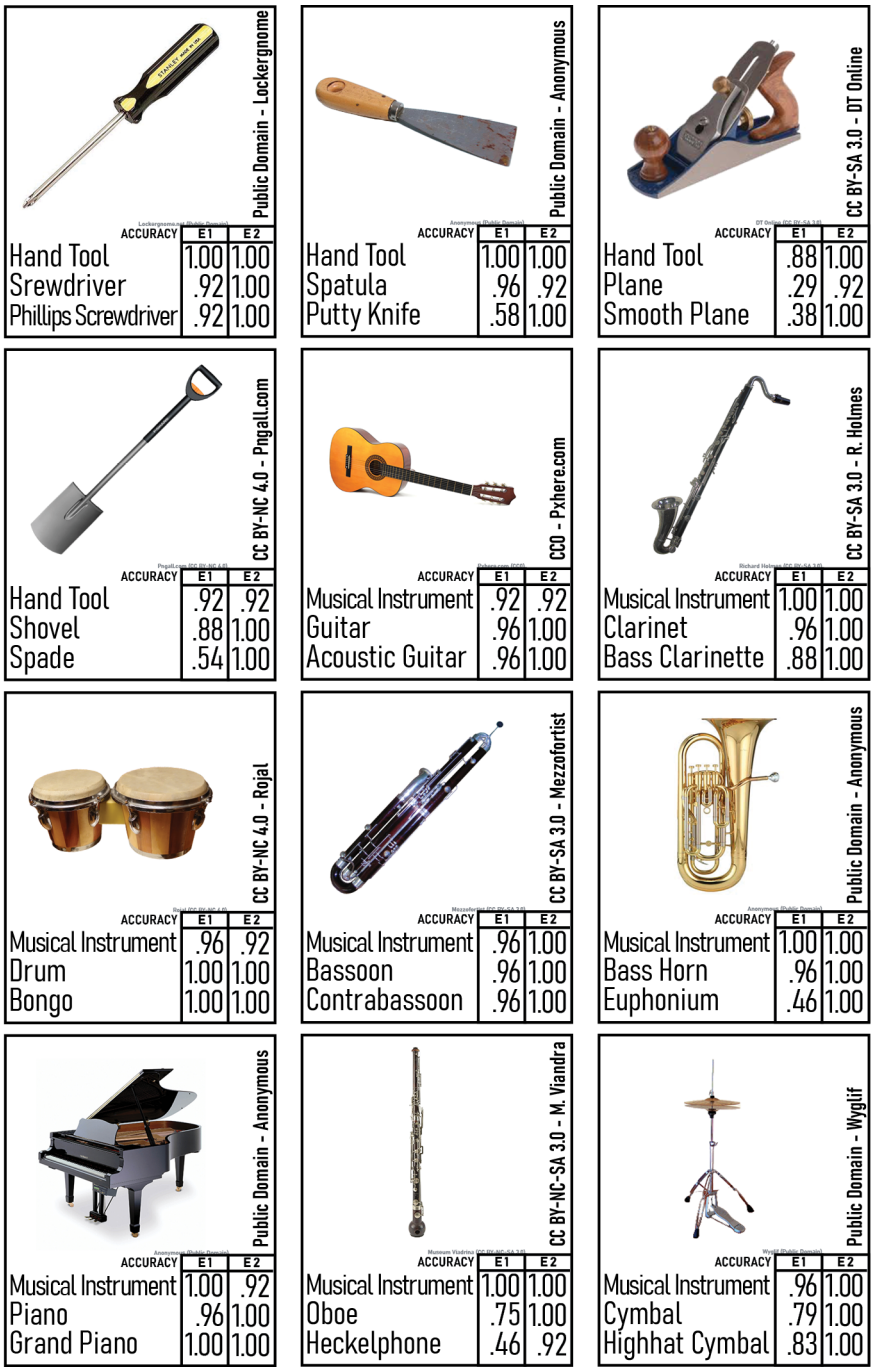

Figure 18: All stimuli images with attribution and accuracy achieved during E1 and E2 (as True type stimuli). Page 5.

Figure 19: All stimuli images with attribution and accuracy achieved during E1 and E2 (as True type stimuli). Page 6.