Learning the basic level from text: Studying different corpus characteristics in predicting the basic level

Investigating discourse type, intended audience, and size of corpora in predicting the basic level in a hierarchy of concepts

SUBMITTED IN PARTIAL FULFILLMENT FOR THE DEGREE OF MASTER OF SCIENCE

Niamh Henry
13245236

MASTER INFORMATION STUDIES
INFORMATION SYSTEMS
FACULTY OF SCIENCE
UNIVERSITY OF AMSTERDAM

2021-06-30

	Internal Supervisor	**External Supervisor**
Name	Dr Frank Nack	Laura Hollink
Affiliation	UvA	CWI
Email	f.m.nack@uva.nl	l.hollink@cwi.nl

Universiteit van Amsterdam

CWI

Learning the basic level from text: Studying different corpus characteristics in predicting the basic level

ABSTRACT

The basic level theory proposes that there is a level of abstraction within a taxonomy that carries the most information for people. Concepts in this level are the quickest to be recognised by humans and we tend to use them most in natural language, especially when speaking to children [25] [16]. Knowledge organisation systems could greatly benefit from "knowing" this level and having the ability to predict it, to enable better human interaction with information systems. This thesis collaborates with the Centrum Wiskunde & Informatica (CWI) to build upon previous approaches to predicting the basic level through investigating differences in the use of the basic level in written and spoken language, and when addressing children in a data driven manner. This is through learning the basic level with varying corpora of different discourse types, audience ages and sizes in words. The results from this thesis highlights that larger corpora are more reliable in predicting the basic level. However, when comparing smaller sample sizes more specified corpora are able to learn the basic level just as well as, if not better than, written texts for a general audience.

1 INTRODUCTION

Knowledge organisation systems (KOS) are widely utilised on the web by a myriad of applications to organise an abundance of information [2]. These systems require categorisation of concepts, and explicit semantic relationships between them to effectively provide users with useful representations of the information that is readily available. Typically, this organisation is taxonomic whereby the concepts are organised as instances from specific to more general categories, or classes, which provide to be crucial for information systems to function for effective human-computer interactions. However, users cognitive structures may not match the semantic structure of the KOS, making it difficult for users to navigate and explore. To further understand and improve the interactions between users cognitive space and the information space, it is useful to turn to cognitive models and theories [24] to replicate our human thought process on the decisions we make to categorise concepts, and how we name them. The basic level theory from cognitive psychology, stipulates that there is a level of abstraction in a hierarchy of concepts that we use most often in naming things as specifically and distinctly as we can, in as few utterances as possible [25]. KOS could greatly benefit from being aware of this level if they knew what the correct level of abstraction is required to interact effectively with users.

Research into the basic level has shown that it is this level at which we recognise concepts the quickest and most accurately, and use most in natural language [25]. The basic level conveys more information than other levels of abstraction, hence, we usually prefer to communicate using this level unless we have a need to be more specific [25]. Studies have shown that it is the level children learn first, and adults use them frequently when communicating with children, to teach them "the name of a thing" [18] [4]. This level holds perceptual universality features by holding the most inclusive categories for which a concrete image of the category can be formed - they tend to have similar shapes and can be identified from average shapes of members of the class [25]. The universality of basic level concepts allows for a crosswalk between KOS [12], and the potential to provide data anchors to cluster concepts by the basic level, to allow users to identify clearly what the information really represents [2]. Many methods have been applied to reap the benefits of this level of abstraction including image recognition, labelling and categorisation, ontology generation and matching, brand awareness in advertising and word sense disambiguation [24] [23] [28] [12] [20]. Therefore, the use of the basic level theory to understand human categorisation is appropriate, and essential to user centred design of taxonomies, ontologies, browsing interfaces and other indexing tools and systems [24].

Despite the vast benefits of the basic level in KOS, there lacks a common, robust system that is aware of the basic level to automatically detect it, and to predict if new concepts are in the basic level. No matter the use of the basic level in the semantic web, a robust text classification system that can identify the basic level in a hierarchy of concepts is required. Hollink et al. [15] developed a classifier to predict if synsets within the WordNet hierarchy are basic or not basic level concepts, and it signalled textual data, in the form of word frequency data (in addition to structural and lexical data) is useful for learning the basic level. The data used for this was Google ngrams which is a huge corpus of books that are not specific to a single domain, and can be computationally expensive to work with. The prominent use of the basic level in daily communication [25], and specifically, in communication with children [4] indicates other textual sources may be useful for learning the basic level. Therefore, to contribute to the requirement of a robust system for classification of the basic level, this thesis addresses the following research questions to identify what characteristics of textual data are useful to predict the basic level in a hierarchy of concepts. The findings from these investigations are incorporated into a classifier to predict all nouns in WordNet as the basic level or not, to result in a dataset of basic level concepts in WordNet for future reuse.

RQ: *What corpora properties are useful in predicting the basic level?*

- SQ1: *Can we learn the basic level from both written and spoken discourse, such as the written and spoken sections in the British National Corpus (BNC)?*
- SQ2: *Can we learn the basic level from discourse that is intended for children to be the audience, such as children's books in the BNC, and child directed speech in the CHILDES corpus?*
- SQ3:*Can we learn the basic level from corpora samples smaller than Google ngrams?*

2 LITERATURE REVIEW

2.1 The Basic Level Theory

The semantic category system is like a hierarchical taxonomy, whereby concepts are organised from highly generalised concepts that are relatively abstract, to detailed, concrete concepts [26]. One level in this organisation has a salient status in human cognition that is gestalt and has appropriate concreteness - the basic level [26]. This level was confirmed in a series of cognitive experiments by Rosch et al. [25], who found we recognise this level quicker and more accurately than at other levels. It is the most general and inclusive level at which categories can be recognised quickly, and accurately through correlational attributes common to all, or most members of the category [25]. Basic level categories are the first categorisations made during perception of the environment, they have similar shapes and they can be identified from average shapes of members of the class [25]. For example, in a taxonomy of furniture, *table* is the basic level. We can mentally picture the general shape of a table, which will represent most of the subordinates of the table class, and infer the purpose and use of a table. We cannot have such a concrete image to represent the superordinate concept of concept *furniture.*

Most of human knowledge, including language, is organised at this level [14]. The basic level is neither too abstract, nor too detailed - it allows us to be as distinctive and specific as we need to be in as few utterances as possible [25]. However the *need to be* in this definition highlights the contextual limitations to the universality of use of the basic level, as a domain expert will likely prefer more technical, subordinate terms rather than the basic level [25]. Jolicoeur et al. [16] elaborate, stating that the level at which objects are first identified depends on typicality not only associated with the basic level, but the background phenomena that govern the individuals environment, as this context can change the entry level point dependent on the domain, and the people. Al-Tawil et al.'s [1] research into using basic level concepts in a linked data graph to detect users domain familiarity provides a strong start to addressing this issue in using basic level categories, as they spark a move towards providing the correct entry-level for users [1]. This would allow for the correct implementation of the basic level across KOS.

2.2 Application of the Basic Level theory

Green's [11] early findings showed that the basic level is universal, and thus, provides useful for crosspaths between different information system,s due to the high probability of occurrence of the concepts across them. KOS such as ontologies, enable information sharing at manipulation, but most data users are disconnected from their design. Knowing the basic level will allow for a more human-like decision for what information a system should display to its user that is more consistent with human-thinking. Cai et al. [5] approach this problem in a context aware model to build ontologies under the theory of basic level concepts, extracted from collaborative tags that users freely create and apply. They propose many potential future directions of this to assist search engines and other KOS. Others [2] have utilised basic level concepts as "knowledge anchors" in a data graph for meaningful learning - for new knowledge to develop starting with familiar entities (anchors) and expanding to new and unfamiliar entities.

In applications of the basic level in information systems, focus has been heavily placed on the advantages they provide to categorise images [23] and to provide image descriptions [24] and captions [20]. This focus may be due to claims that only perceptual features are critical for the basic level advantage [21]. However, this was disproved in experiments that found perceptual features are not necessary for the basic level advantage [6]. Albeit, the potential applicability within this field is vast, as studies have shown promise for more human-like labels for images by categorisation systems that are aware of the human naming choices in deciding between the terms *grampus griseus*, or simply, *dolphin* [23]. Almost all studies with perceptual features utilise or rely on textual descriptions of images to classify them, or to generate textual descriptions, yet, fully investigating different sources of text has been overlooked in basic level research. Interestingly, Ordonez [23] developed two models (a language and visual based) for predicting the entry-level category for images, with the linguistic model providing better results. Hence, more research is required into utilising the advantages from basic level concepts that can be found in natural language - no matter the potential uses of the basic level on the semantic web, text needs to be prioritised to allow for these applications to develop and grow coherently and effectively.

2.3 Textual based methods to identify the Basic Level

To exploit the advantageous characteristics of the basic level category in information systems, some research has delved into methods on how to identify these categories based on their characteristics from the vast amounts of text and data readily available. However, a robust method for identifying the basic level to further develop a range of multidisciplinary research has not yet been standardised. Some metrics have been utilised to identify the basic level through cue validity, category feature collocation and category utility however these purely focus on the features or attributes of concepts, which complicates the specification of features within domain if they indicate importance [13]. Significant starting attempts have been made with the use of Princeton's WordNet [9] - an open-domain dictionary scale resource, that organises concepts into a hierarchy of hypernyms and hyponymns [20]. According to Mills [20], WordNet is the best available resource to build upon identifying basic level and has been widely utilised with almost all studies utilising the platform.

2.3.1 Structural Features. There has been a range of methods applied to identify the basic level within WordNet, with Green [11] successfully highlighting the basic level by using structural features of the hierarchy, such as the number of relations, number of children, the depth of occurrence in the hierarchy and the length and structure of lexical units [11]. The relations of the basic level to other concepts allows for practical inference for how a thing needs to be constituted for most purposes [4]. However, this limits concepts within the basic level as purely structural, rather than additionally accounting for the semantic role the basic level plays in an individuals everyday activities within the physical world [14], due to the contextual entry-level nature of the basic level. Thus, investigating the occurrence of use of basic level concepts when we use them most - in natural language [16] - is required to identify all characteristics and uses of the basic level.

2.3.2 Discourse Features. As there are some contextual limitations to use of the basic level for those who need to be more specific, we need to look beyond the structural properties of the basic level within domains, to the use of basic level concepts in communication. Hajibayova et al. [14] discuss how the vocabulary in a language system is mapped to a specific semantic category in a specific hierarchy and reason that the vocabulary system follows the same hierarchical structure [14]. Basic level categories are the natural categories that people use in casual language [23] as referring to objects is a core function of human language [10]. Wisniewski et al. [29] performed a textual analysis on written casual discourse, and acknowledge their study is restricted to written language, submitting that speech should be investigated further hypothesising that speakers, as opposed to writers, would use fewer specific names for single objects, and may prefer the more general basic level term to describe it [29].

The main investigation into basic level concepts in speech has been into the occurrence of basic level terms when addressing children, as Brown's [4] early findings showed that children learn a middle level of concepts before those that are broader, or more specific. This is widely supported throughout the literature with reasoning pointing towards the representative qualities of the basic level [25] [17] [29]. Additionally, when addressing children adults tend to use the basic level, as it is how parents teach their children to name and categorise objects - the "name of a thing" that conveys what it really is [4]. This result has been upheld in studies of parent-child conversations [18] where preference for the basic level name is apparent in spontaneous speech to children within the CHILDES corpus [19] - a database of child directed speech that has been utilised for a small number of studies [18] [20] [8]. Thus, this corpus would provide to be a valuable source for identifying basic level concepts. This is shown through studies that utilise corpora to extract the frequency of a basic-level term, mostly using Google books ngrams to extract the number of occurrences of a word [10] [23] [15], suggesting corpora are a useful tool to find basic level terms.

2.4 Predicting the Basic Level

A robust system for identification of the basic level in a hierarchy of concepts within, and across, domains is required to fully reap the benefits of the basic level within KOS and the semantic web, using both structural information from the KOS, and considering the use of the basic level in varying contexts in everyday communication. Hollink et al. [15] provide a useful start to this, with their Random Forest classifier to predict if concepts are the basic level, or not in WordNet, trained using a gold standard manually labelled synsets as basic level or not from 3 branches (that are referred to as *domains*) that were confirmed by Rosch to have basic level effects - all synsets under the roots *edible_fruit.n.01*, *hand_tools.n.01* and *musical_instruments.n.01*. The input features included in this classifier include structural and lexical features extracted from WordNet (in line with Greens criteria [11]), in addition to a feature of Google ngrams frequency data. The ngrams feature provided to be an important indicator of the basic level, highlighting the prominence of the basic level in corpora. In addition, Mills [20] tested a range of features to find best identifiers of the basic level, and found one of his strongest indicators of a synsets being a basic level category was the word appearing in the CHILDES corpus.

Evidently through Hollink et al.'s research, structural and lexical features of concepts in a hierarchy, in addition to corpora frequency features, provides accurate results in identifying the basic level in WordNet [15]. However, it is unknown what type of corpora are required to learn the basic level. Ngrams is a huge resource of books and is an obvious choice - when learning in a data-driven manner, the bigger is usually the better. However, as research has previously pointed out, there is a preferred use of the basic level in natural language, with studies indicating we use it more in speech [29]. Despite this, no studies have enquired into purely spoken corpora to learn the basic level, despite there being a range of sources freely available, particularly in the TalkBank repository[1]. Additionally, the prominence of the basic level when addressing children has been acknowledged to be a good source of learning the basic level, yet, studies that use the CHILDES corpus have failed to exploit all the potential uses of it to learn the basic level, as they simply look at the binary occurrence of the basic level appearing or not [20]. It is also unknown if this trait of the basic level when addressing children applies to both spoken and written discourse.

3 RESEARCH DESIGN

Due to prior research, particularly from Hollink et al. [15], illustrating the prediction power of structural, lexical and frequency features in identifying the basic level, it is evident these three types of features are accurate and appropriate to use in machine learning with the basic level. With the ability for specified corpora to represent different types of communication, language, people and cultures, there is a surprising lack of research into using different corpora to learn the basic level in a data-driven manner. Thus, this thesis investigates different corpora properties and characteristics that enable learning of the basic level, considering type of discourse (spoken or written), intended audience (general or children) and size of corpus by addressing the following research questions:

RQ: *What corpora properties are useful in predicting the basic level?*

- SQ1: *Can we learn the basic level from both written and spoken discourse, such as the written and spoken sections in the British National Corpus (BNC)?*
- SQ2: *Can we learn the basic level from discourse that is intended for children to be the audience, such as children's books in the BNC, and child directed speech in the CHILDES corpus?*
- SQ3:*Can we learn the basic level from corpora samples smaller than Google ngrams?*

These research questions will be addressed using the methods and data outlined in Section 4.1. The exploration of these research questions using different corpora samples is driven by the following hypotheses, which will be tested using the experimental setup discussed in Section 4.2.

[1] TalkBank fosters research into human communication, particularly spoken. Can be accessed here: https://talkbank.org/

- H1: *Basic level concepts will appear more times than those more specific or more general, in all corpora, as they are more commonly used in natural language.*
- H2: *Spoken language corpora will have a larger proportion of basic level concepts than written language corpora.*
- H3: *Basic level concepts will appear more often in discourse directed at, or by children, than those targeted at a general adult audience.*
- H4: *Smaller, specified corpora such as spoken discourse and those directed at children, will perform better than small samples of general written discourse.*
- H5: *The frequencies of the basic level concepts in these corpora will aid existing training data to accurately identify and predict the basic level.*

4 METHODS

The classifier developed by Hollink et al. [15] will be used as a benchmark for measuring performance of these corpora frequency features to learn the basic level. The best performing settings will then be used to predict all synsets under the top noun in WordNet (*entity.n.01*), to provide a dataset of basic level synsets made publicly available for reuse[2]. The original gold standard from Hollink's [15] research will be reused, and extended to another two domains in WordNet that Rosch confirmed to have basic level effects - *furniture.n.01* and *garments.n.01*. They will be manually labelled by three annotators with good knowledge and understanding of the basic level to extend the gold standard (full annotation protocol is outlined in Appendix F). This new gold standard, in addition to the structural and lexical features from WordNet outlined in Section 4.1 below, will be tested in the experiments outlined in Section 4.2 by using different word frequency features, from a range of different corpora sources listed in Table 1 in place of ngrams, which is used as the benchmark.

4.1 Input Data Extraction

4.1.1 WordNet Features. The input features include a range of structural and lexical features extracted from WordNet to represent characteristics of the basic level. For instance, the basic level is the level that carries the most information according to Rosch et al. [25], hence, the extent of information about a concept in a KOS, such as WordNet, can be used as an indicator for the basic level, represented by the number of relations it has within the KOS, or the length of the concepts description. Other characteristics of the basic level such as being denoted by shorter and more polysemous words can be represented by the length of a term in characters, the number of senses of a word and the number of synonyms it has. To represent these indicators within the input to the classifier, the Natural Language ToolKit (NLTK) corpus reader on Python[3] was utilised to extract the following features from WordNet for each synset:

- The number of direct/indirect and total hyponyms;
- The depth from top synset (the number of steps from the top synset);
- The number of characters in the gloss;
- The number of lemmas in the synset;
- The sum of the frequencies of the lemmas in the synset;
- The poly score - the number of synsets in which the most polysemous word (lemma name) in the synset occurs;
- The total number of synsets in which the words (lemma names) in the synset occur, as a measure of polysemy;
- The mean number of characters in each word (lemma name) in the sysnet;
- The number of characters of the shortest word (lemma name) in the synset;
- The number of part-whole relations the synset participates in.

4.1.2 Frequency Features. Hollink et al. [15] incorporated frequency data from the most recent Google ngrams data (2008 corpus)[11], using the log mean score of the words in each synset. Thus, this will be used as the only frequency feature in the initial model to reproduce the results from Hollink's research to use as a benchmark, for evaluating performance of new frequency features from a range of corpora and subcorpora. The corpora outlined in Table 1 were extracted to compare performance of samples with different discourse type, intended audience and sizes. The unique word frequency counts were extracted from each corpora using the NLTK corpus reader [22] (for the BNC xml text files) and the python module *pylangacq*[12] (for CHAT transcript files). A range of metrics were taken from these frequency counts including binary appears/does not appear, the sum, mean and maximum of all frequencies of words in each synset, in addition to relative frequencies to the size of corpus for direct comparison between them. The different properties of the selected samples are outlined below:

- **Different Discourse Types**: To investigate differences between written and spoken discourse in learning the basic level, the written and spoken parts of the British National Corpus (BNC) [7] are compared. The written section of the BNC is freely available as full-text xml files with over 100 million words from an array of sources, providing 74k unique word counts. The conversationalist section of the BNC, the CABNC was additionally extracted to provide 30k unique word counts.
- **Different Audience Types**: To further investigate the prominence of the basic level when addressing children, all written texts in the BNC with the intended audience of children were extracted, to provide 30k unique word counts. Additionally, child, and child-directed, spoken discourse was extracted from the CHILDES [19] corpus, providing 33k unique word counts.
- **Different Sample Sizes**: To investigate what size of sample is required to learn the basic level, smaller written sample sizes were extracted from the BNC; a 10 percent, 2 percent and 1 percent sample, to gain insights into the requirements of the volume of text data for prediction.

[2] All data and materials to reproduce these experiments can be found here: https://github.com/niamhhenry/Thesis_Basic_Level/tree/main/Experiment%201

[3] Documentation available from https://www.nltk.org/howto/wordnet.html

[11] Google books ngram corpus: https://books.google.com/ngrams

[12] Documentation can be found at https://pylangacq.org/

Table 1: Corpora samples extracted for word frequency data

Corpus	Discourse Type	Intended Audience	Size (words)	Source
BNC Full Text	Written	General	100 million+	Text from wide range of sources[4].
CABNC	Spoken	General	2.4 million	Transcripts of naturalistic conversations from the BNC[5]
KBNC	Written	Children	1 million	All written works intended for children in the BNC[6]
CHILDES	Spoken	Children	5.7 million	Transcripts of child speech and child directed speech[7].
BNC General	Written	General	10 million	10 percent sample of general written works in the BNC[8]
BNC Sampler	Written and Spoken	General	2 million	2 percent sample of the BNC half written, and half spoken[9].
BNC Baby	Written	General	1 million	1 percent sample from the BNC (from the written fiction section)[10]

4.2 Experimental Setup

The three experiments outlined below (EXA, B and C) will be conducted to test the new gold standard domains using the features from Hollink et al.'s research (Experiment 1) to provide a benchmark to compare the performance of new frequencies (Experiment 2). The results from these experiments are used to gain insights into how well these model respond to different frequency features, and to find the best performance settings to predict the entirety of WordNet (Experiment 3). The best performing algorithm from Hollink's research was the Random Forest classifier, faced with the binary classification task of synsets as basic or not basic, run with the SMOTE algorithm to deal with class imbalances (the number of concepts in the basic level is much smaller than those not). Hence, the classifer used in these experiments will be a Random Forest. Hollink's was developed in R however, this research will reproduce, and extend the classifier in Python.[13]

These experiments will use a 10-fold cross validation set up, using StratifiedKFold and Cross_Val_Score in Python[14]. This cross validation performance will be judged primarily on cohen-kappa scores (like Hollink's), which indicates how well the inter-rater agreement is (between the gold standard and the predicted labels), while considering the probability of predicting the correct label randomly by chance. This is useful for interpreting results from the random forest, as it may produce random results.

4.2.1 EXA: Training and Testing Local Models. As the gold standard consists of five different domains (edible fruit, hand tools, musical instruments, furniture and garments), each will be individually used to train and test the random forest classifier within single domains. This will result in five separate models that indicates individual domain performance and feature importance of the training data within specific domains. Experiment 1 uses ngrams as the frequency feature to gain benchmark results, and Experiment 2 tests new frequency features in place of ngrams to compare performance.

4.2.2 EXB: Training and Testing a Global Model. The entirety of the gold standard (all five domains) will be used to train and test the classifier to see how compatible the training data is for performance between different domains. Experiment 1 will be used as the benchmark result, and each frequency feature will be tested in place of ngrams, and results compared.

4.2.3 EXC: Testing on a new, unseen domain. To better simulate the final task for this classifier, predicting WordNet under its top noun, *entity.n.01*, models will be trained on four domains, and tested on one single, unseen domain. For example, train on tools, musical instruments, furniture and garments and test on fruit. This will be repeated to test each domain as the unseen one, using the different frequency features. This experiment will provide insights into how useful the training data is to generalise to new domains, and what features are important for the individual domains. Again, this will be performed in both Experiment 1 and 2 to compare results against a benchmark.

4.2.4 EXD: Training on Gold Standard for prediction of WordNet. Using the model with the best performance settings, the entirety of WordNet under the branch *entity.n.01* will be predicted to develop a dataset of basic level synsets within WordNet, to enable research into the use of basic level concepts in applications. This will be compared with Hollink et al.'s [15] dataset for evaluation and validation.

5 RESULTS

5.1 EX1: Extending the Gold Standard

To increase the size and scope of the current gold standard, three annotators labelled the synsets under both the furniture and garments branches in the WordNet hierarchy as the basic level, or higher or lower than the basic level providing an additional 468 synsets to the existing gold standard[15], resulting in 986 labelled synsets from WordNet across five domains - all confirmed to have basic level effects by Rosch and her colleagues [25]. The majority vote labels between all annotators can be seen in Table 2, for which there was good inter-rater agreement between all annotators with cohen kappa scores of 0.85 (furniture) and 0.79 (garments), indicating the gold standard is reliable.

Table 2 shows that the majority of labels are not basic level (as expected), however, the proportions of the basic level is considerably lower compared with the older gold standard; in furniture and garments only 12 and 15 percent are the basic level, compared with fruit, musical instruments and tools (39, 41, 25 percent respectively). This makes classification of the minority (basic) class difficult, however it is addressed with the SMOTE algorithm to over-sample basic

[13] The same random seed (7) is used and the hyperparameters of the model are adjusted to closer resemble the settings in R. This includes increasing the number of trees from Python's default 100, to R's 500.

[14] Documentation found at: https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.StratifiedKFold.html and https://scikit-learn.org/stable/modules/generated/sklearn.model_selection.cross_val_score.html

[15] The term *trouser* appeared as two synsets, thus was dropped from the gold standard.

level data points to have more data to train the model. This is implemented using the imbalanced learning pipeline[16] to ensure no data leakage from the test set, to the training set when cross validating over 10 loops. The results of the three experiments outlined in Section 4.2 using the new gold standard are outlined below.

5.1.1 EX1A. Local Models. The cross validation cohen kappa (benchmark) results of fruit, music and tools using the same input features as Hollink and majority vote labels are displayed in Figure B.1. From this plot it is evident that the results from the classifier in R, can be reproduced in Python. The reasoning for the slight variance in scores is unknown, as Python performances seem to have a much larger range despite using the exact same input features. However, it may be due to the nature of the random forest on different programmes as it is a stochastic classifier, thus the randomness that determines the results may provide varying results.

Results of the two new domains, furniture and garments, using majority vote labels scored median kappas=0.44 and 0.38 respectively. They are the lowest performing local models, thus the toughest domains to classify out of all five. This may be due to the considerably lower number of basic level concepts within these domains. Hollink et al. found some improvement using synsets where all annotators agree, as some concepts may not show basic level effects, and those with good agreement the basic level effects are more evident. This implementation improved all domains as seen in Figure B.2, particularly garments and furniture. Thus, only labels where all annotators agree are used in following experiments, as they perform better in predicting the basic level.

5.1.2 EX1B. Global Model. The benchmark results of the three domain global model for fruit, music and tools was achieved, as seen in Figure B.1. The five domain model performed well with a median cross validation kappa score=0.68. However, not as well as the three domain model. This may be due to the addition of two difficult domains, or a change in the patterns seen by the model across the new domains. As per Hollink et al.'s findings, predicting across or to new domains improves with a normalisation step of some of the structural features by dividing by the domain mean (the depth of hierarchy, the gloss length and the number of part relations). This is supported in Figure B.3 thus, all models in the following inter-domain experiments will use these normalised features. [17]

5.1.3 EX1C. Unseen Domains. As per the results from the previous sections, the synsets where all annotators agree and normalisation of some structural features are the settings applied to test on new, unseen domains. Results are outlined in Table 3, which reinforces the good performance of the classifier on the fruit and tools domains, while highlighting the lower performance of the evidently tougher domains to predict when unseen; musical instruments, furniture and garments. Albeit, they still achieve good balanced accuracy scores, which provides to be a good indicator of recall of the minority class for imbalanced datasets, in this case - the recall of the basic level.

Table 2: Majority vote between all 3 annotators in furniture and garment domains.

Label	Furniture	Garments
Higher	13	20
Basic	24	40
Lower	160	211
Total	197	271

Table 3: Unseen domain results from Experiment 1 (using structural features and ngrams)

Trained on	Tested on	Cohen Kappa	Balanced Accuracy	Accuracy
Music, tools furniture and garments	Fruit	0.862	0.9344	0.852
Fruit, music, furniture and garments	Tools	0.761	0.892	0.859
Fruit, tools, furniture and garments	Music	0.562	0.7728	0.721
Fruit, music, tools and garments	Furniture	0.5695	0.778	0.685
Fruit, music, tools and furniture	Garments	0.4001	0.737	0.682

5.2 Exploring New Frequency Features

A range of metrics were taken from the frequency word counts of the synsets within the gold standard from each corpora outlined in Table 1. To compare the instances of the basic level, and those above and below, the sum of each lemma per synsets' occurrence was calculated in raw, and relative occurrences (per million words of each corpus) and plotted in Figures A.1 and A.2. From these, we can see H1 is confirmed, as the basic level concepts have a higher proportion of appearances and they are used more frequently. H2, that spoken discourse will have a higher proportion of basic level concepts to appear than written discourse, is reinforced through these diagrams as the CABNC has a slightly higher proportion of basic level concepts to appear than the BNC as seen in Figure A.1. When comparing discourse type when children are the intended audience, the proportion of basic level terms appearing in the corpora is the same (53 percent). Yet, basic level concepts from the gold standard occur more frequently per million words in the children's spoken corpus (CHILDES - the red bar in Figure A.2), than in any other, highlighting the prominence of use in the basic level when verbally addressing children, confirming H3. To enquire if this pattern will translate into the ability for the classifier to learn the basic level as hypothesised in H5, the three experiments outlined in 4.2 were conducted to investigate different discourse, intended audience, and sizes of corpora in learning the basic level. The metrics used for each corpora are:[18]

- **BNC Sum**: The sum of all instances of each lemma per synset in the BNC full text.
- **CABNC per 100k**: The frequency occurence of all lemmas per synset, per 100,000 words of the CABNC.

[16] Documentation found here: https://imbalanced-learn.org/stable/references/generated/imblearn.pipeline.Pipeline.html

[17] Normalising the remaining structural features, number of hypos and direct hypers showed to be harmful to results when predicting an unseen domain, therefore the raw values of these two features were used in the global and unseen models, and the other 3 normalised.

[18] As there were a range of metrics extracted from each corpora, they were all trialled on the global model to find the best performing from each corpora and the metric with the best median and smallest range of cross validated cohen kappa scores for each feature was selected.

- **KBNC Sum**: The sum of all instances of each lemma per synset in the KBNC texts.
- **CHILDES relative sum**: The sum of all instances of each lemma per synset in the CHILDES corpus, divided by the total number of words in the corpus.

5.3 EX2: Comparing Performance of New Features

5.3.1 EX2A: Results in local models. In the local domains, there were few changes in the cross validation results when the different individual frequency features were used, particularly within the fruit and tools models where the benchmark results did not budge, as seen in Figures C.3 and C.4. This emphasises the high utility of structural and lexical features within the fruit and tools domain. Therefore, any insights are only gained from the musical instrument, furniture and garments models that are plotted in Figure C.1. No results from this experiment were statistically significant from each other, however, from this plot, some slight differences are visible in the individual domains.

In music, we can see the only features to achieve the benchmark median (0.69) are the BNC and CHILDES features, with the BNC having the highest median kappa in music (0.74). In furniture, there is an interesting theme as the only features to improve the benchmark performance on furniture (median kappa=0.53) are both the spoken corpora, and the children's written discourse as seen in Figures C.1 and C.3 as individually, these three features each achieve a median kappa=0.62. When combining features based on discourse type, the two spoken features in combination increase this median score (0.64). Comparing audience type in Figure C.4, shows the children's discourse features reach the benchmark scores whereas the general audience features decrease performance. Evidently, the more specified corpora of spoken and children's discourse (CHILDES, CABNC and KBNC) perform better in the furniture domain, than the larger BNC and ngrams. In the garments models, all features achieve similar to the benchmark median result. However, none of the new features have results as reliable as ngrams (all 10 scores distributed across a small range as seen in Figures C.1, C.3 and C.4). This is perhaps due to each word in the garments domain appearing in the ngrams corpus as seen in Table C.8, showing that having data on all synsets provides more reliable results. This indicates the usefulness of larger corpora sources, as there is a higher chance of terms appearing in them.

5.3.2 EX2B: Results in global model. Evidently, there are no considerable changes in the results of the global model from the benchmark (ngrams, median k=0.68) as seen in Figure C.2. The highest median cohen-kappa score was the BNC with median kappa=0.72. The range of scores vary slightly over the 10 loops, however no results were statistically significant. Thus, we cannot reject the null hypothesis that all these features can perform equally as well as each other. Hence, the results of the KBNC and CABNC are impressive given the considerably smaller size of them (1 and 2 percent of size of BNC), to perform as well as resources 100x the size. This is seen in testing the smaller BNC samples on the global model in Figure C.7, as they achieve the benchmark median results in the global domain. However, we can see from Figures C.2 and C.7, that the smaller corpora features have a larger range, emphasising earlier findings that larger corpora are more reliable. This phenomena is clearly seen in Figure C.7 as the corpus size gets smaller from left to right, the distributed range of results gets larger.

5.3.3 EX2C: Results in unseen domain models. The results of all new features to train a model to predict on an unseen domain are outlined in Table D.1. The highest performing feature in each domain is highlighted in yellow. Ngrams remains the highest performer when testing the furniture and music domains. BNC performs best in fruit, and has the slight advantage over ngrams in the garments domain, purely on accuracy and balanced accuracy scores. In the tools domain, the CHILDES corpus is the best performer. Interestingly, when using new word frequency features in place of ngrams, the accuracy and balanced accuracy predicted scores increase, in some cases, even when there is a lower cohen kappa score than the benchmark. For instance, when testing on the music domain ngrams has the highest kappa score, but the rest have higher accuracy and balanced accuracy scores.

5.4 EX3: Predicting the Basic Level

5.4.1 EX3ABC: Finding the best performance settings. As no results were statistically significant to accept the hypothesis that different corpora may perform better (or worse) than ngrams, there is no clear indication of what the best performing settings are using new corpora sources. Additionally, all the word frequency features are highly correlated with each other which poses a risk of a correlation bias that may produce unstable results, as the algorithm would prioritise one of these features as the key predictor [27]. This was evident in a trial combination of CHILDES and ngrams, as prediction of the tools domain when unseen decreased from 0.82 to 0.77 with the additional feature. Thus, only one frequency feature should be used to ensure performance is not hindered. Combining them all into one aggregate metric was considered, but dismissed as each represent different real life interactions and thus, hold differing meanings - one single metric to represent them all would be arbitrary and diminish the value they hold individually. Hence, only one frequency feature from a single source should be considered.

Comparing the individual frequency features across all experiments outlined in Section 5.3 shows we can learn the basic level from all of them. However, some are considerably more reliable than others. The most reliable results seen from cross validation loops, were those from large corpora sources. Additionally, to generalise to all the new domains in WordNet, it is best to use a corpus that is likely to contain the concepts that are being classed. Google Ngrams is the largest corpus of all sources, and remained to be the strongest individual feature over all experiments. Therefore, it was selected as the most reliable feature for generalising across new domains.

5.4.2 EX3D: Predicting the Basic Level in WordNet. The features outlined in Table E.1 in addition to the gold standard of synsets where all annotators agree, are used as input to the final model that is trained to predict all synsets under entity.n.01 as basic or not basic. There are 74,374 synsets under the entity root, therefore a dataset was created of all input features for each synset. The synsets from the top of each domain, to the root entity were additionally added to the gold standard for training the final classifier (28 in total). These were labelled as 'not basic' as by definition they are all above the basic level. Hollink et al.'s [15] ad-hoc algorithm to define a WordNet "domain" was reused to normalise structural features per their subject domain. Additionally, the only hyperparmeter settings in the Random Forest to change from Python's default settings is the number of trees, which was increased to 1400 to build a larger forest, to address the potential of the model over-fitting new data [3].

This experiment resulted in a dataset of 15k synsets under entity.n.01 labelled as the basic level by the model with the best performing settings. This equates to approximately 20 percent of all synsets under entity belonging to the basic level class - somewhat in line with the gold standard, which showed an average of 26 percent of a domain to contain basic level concepts. Albeit, these are in domains confirmed to have basic level effects. A similar dataset of predicted basic level concepts under entity.n.01 is available from Hollink et al.'s classifier which consists of only 10k basic level synsets. With the additional synsets in the training data, an increase in the number of basic level synsets predicted is not a surprise. However, when directly comparing the two datasets for validation they do not have the majority of synsets in common, agreeing on only 4,229 basic level synsets. This is surprising as the same input features were used in both classifiers. Hence, predicting the basic level remains to be a difficult task as there is no standardised manner to confirm what synsets truly are the basic level. Nevertheless, we seen in previous experiments, the basic level effects are seen most when there is universal agreement, therefore, the synsets that are agreed between the two datasets are highly likely to be basic level concepts. The 15k predicted synsets, and the 4k agreed synsets are publicly available for future reuse[19] in applying the basic level with WordNet, and beyond it, in an array of applications in the semantic web.

6 DISCUSSION

6.1 Comparing discourse, audience and sample sizes

6.1.1 SQ1: Can we learn the basic level from both written and spoken discourse, such as from the written or spoken section in the British National Corpus (BNC)?. From the results of the experiments conducted in Section 5.3, it is evident that we can learn the basic level from both spoken and written discourse, albeit, one is not necessarily better to learn from than the other. From Figure A.2, we can see that in both spoken corpora sources, there is a higher level of frequency of basic level terms from the gold standard, than their written audience counterparts. Yet, this prominence of the basic level in spoken discourse does not translate to the classifiers performance in local domain models, as there were no statistically significant changes when using different discourse types on local models, as hypothesised in H5. However, interesting fluctuations can be seen in Figure C.1 as we can see the only features to improve the furniture model (one of the worst performing models), are spoken discourse features when used individually, and in combination as seen in Figures C.1 and C.3. Thus, spoken discourse is useful within the furniture domain.

When the discourse features were tested in combination on the global model the benchmark median kappa increases from 0.67 to 0.69, with a small, and thus, reliable range of results as seen in Figure C.3. These results were not statistically significant, nevertheless, still provides useful insights, as it is confirmed we can learn the basic level from both spoken and written discourse, just as well as each other. These findings are useful, particularly due to the differences in the sample sizes of the features used to compare - the spoken discourse data is extracted from approximately 6 percent of the total size of the written features data sources. Hence, the ability of the spoken discourse features to perform the same as, or better than, these features is remarkable. The differences sample sizes make on the performance of features is discussed below in Section 6.1.3.

6.1.2 SQ2: Can we learn the basic level from discourse that is intended for children to be the audience, such as children's books in the BNC, and child directed speech in the CHILDES corpus? Prior research highlighting the prominence of the basic level when addressing children [4] [18] is supported using the gold standard to compare word frequencies as seen in Section 5.2 Figures A.1 and A.2. In particular, Figure A.2 shows the CHILDES corpus has the highest frequency occurrences of the basic level concepts from the gold standard occur, per million words of the corpus. This is considerably higher than all other sources, with the next highest being written discourse aimed at children. It was hypothesised in H5 that this prominence in corpora would translate patterns to the algorithm, to allow for more accurate identification of the basic level. However, no results showed to be significantly different from the benchmark in learning the basic level, yet we can extract some insights from these results; as individual features, the KBNC and CHILDES features perform well on local models, particularly in furniture. In the global models, the individual features aimed at children perform slightly lower than those aimed at a general audience as seen in Figure C.2.

Testing combinations of features based on audience type, Figure C.4 shows no considerable changes but some common themes are identified to other experiments; the fruit and tools models do not change in performance, and no features reach the benchmark results in the garments domains. Music, however, is improved by sources for an intended general audience. Albeit, this may be due to the high number of appearances of the musical instrument domain found in the BNC and ngrams (as seen in Figure C.8), rather than the change in audience type. Consequently, the BNC dominates all results in music when used as a feature. Yet, when comparing differing sample sizes, further discussed in Section 6.1.3, the BNC General (10 percent of the BNC) and CHILDES (5 percent of the BNC) achieve similar results. This indicates the strong performance of childrens discourse, as it achieves the same as sources double its size - if there were larger sources of children's discourse available, the results of this experiment could be greatly improved.

To further compare the differences of audience type on similar sized sources, the KBNC performance is directly compared to the BNC Baby (1 million words of written text for a general audience) in Figure C.5. From this, it is evident there are no major differences

[19] All materials for all input data, models and results are available here: https://github.com/niamhhenry/Thesis_Basic_Level

between performance when changing audience type. In music, general audience (BNC Baby) has slightly higher median kappa scores, however, in furniture and garments (the tougher domains) KBNC performs better. In the global model, the KBNC seems to be more reliable with a small range of high results (0.56-0.78). In unseen domains, the KBNC outperforms the BNC Baby in all models, apart from when garments was unseen, with this difference being small (0.01). Thus, written discourse aimed at children, rather than a general audience, is more accurate at generalising to new, unseen data when the sample sizes are the same.

6.1.3 SQ3: Can we learn the basic level from corpora samples smaller than Google ngrams? From the results discussed in the previous section, it is evident that all frequency features introduced that are smaller in size than Google ngrams, have the ability to learn the basic level. However, this is impressive, as the frequency features were taken from a range of sample sizes (from 100 million words, to 1 million words). Yet, no results showed to be statistically significant from one another, so we cannot reject the null hypothesis of these samples performing the same. Thus, an impressive result for the predictive power behind small subcorpora such as the KBNC.

To further investigate differences in sample sizes, a range of samples from the BNC were tested, as seen in Figure C.7. Evidently from this plot, we can learn the basic level from a range of sample sizes, including small samples of 1 or 2 million words. However, reliability may be a concern when using these, as we can see that as the sample sizes get smaller (from left to right), the range increases and thus, the results get less reliable.

As previously discussed, similar sized samples were compared (both one million word sources, and both two million word sources, seen in Figure C.5 and C.6). Comparing audience type, showed the KBNC to be more reliable and a better ability to generalise to new domains, than its general audience counterpart (BNC Baby) in the global and unseen domains. Comparing the two million word sources is interesting, as we can directly compare pure spoken discourse, to a mix of discourse from the BNC Sampler. This showed the purely spoken corpus, the CABNC achieving higher median cross validated kappa score in the cross validation, with a smaller range of results. This indicates the reliability of the spoken language resource in comparison with the mixed discourse - when the sample sizes are small, spoken or children's discourse seems to be more reliable than written discourse for a general audience.

To investigate this notion further, CHILDES (5 million words) is directly compared with the BNC General sample (10 million words from general written texts). Despite being half the size of a data source, CHILDES evidently provides more useful information as it performs stronger in the global model, and when predicting new domains as seen in Table D.1 (apart from garments, for which BNC and ngrams always dominate). Notably, when tools is left as the unseen domain, the CHILDES feature is the best scoring out of all features tested in this research (kappa=0.82). These results signal the usefulness of more specified corpora, particularly when using smaller word samples.

From the results of this thesis, we can see that you can learn the basic level from all different types of corpora including different discourse, audiences and sample sizes. Evidently, it seems the trends from these results illustrate that the smaller the sample size, the less reliable the results. Comparing small sample sizes of written discourse for a general audience to the smaller, more specific corpora (spoken language, and discourse aimed at children) showed slightly better scores and more reliable results when a small specific corpora is used, rather than a general written discourse sample. This is in line with H4 outlined in Section 3.

6.2 Limitations of Gold Standard and Predicting WordNet

The results of the two new domains as local models, and their addition to the global model seem to hurt performance when compared to the smaller three domain model (seen in Figure B.2). Additionally, the resulting predicted basic level synsets from WordNet using this new gold standard did not have high agreement with the predicted synsets from Hollink's [15] research, despite the same input features being used. However, with good inter-rater agreement between all annotators, and with an external source of crowdsourced labels [20], the gold standard seems to be reliable. There are a number of considerations as to why this is not evident in the results. For instance, there is a lower proportion of basic level terms in both these domains compared with the previous three, some synsets were difficult to label as basic level or not - future work may be improved by excluding synsets that may not show any basic level effects. There were changes in two of the annotators from Hollink's research, and the annotation protocol itself was altered to allow for more than one basic level concept per branch[20]. Additionally, the gold standard consists only of non-biological domains, which Rosch [25] showed to have differing structural features to biological domains. Hence, the generalisability to these new, biological domains when predicting WordNet may be unreliable.

As there is no strict standardised classification of what is the basic level or not, or how many basic level concepts there are, it is hard to determine the correct and valid predictions amongst the two datasets of predicted basic concepts in WordNet. Prior research has pointed to a range of possible number of basic level synsets in WordNet's nouns, with Mills [20] extrapolating to there being <2,000 synsets as the basic level, Green [12] suggested there were around 7k and Hollink's [15] classifier identified 10k. Therefore, the resulting 15k synsets of the final model run in this research may be an over-estimation of the number of basic level concepts to be present in the hierarchy. Albeit, this research has found that when agreement is universal, the basic level effects are seen more (exemplified in Figure B.2). Thus, the agreed synsets from both classification tasks are highly likely to be basic level concepts. These are available as a dataset of synsets[21], which will provide to be a useful start for confirmed basic level synsets in future research. However, the vastly differing results of the number of basic level concepts within the WordNet reinforce the need for a shared, general approach to identifying, and predicting the basic level.

[20] This was to allow terms such as *bar* to be included as a basic level concept as they are by definition basic level, despite appearing below the basic level concept *table* in the hierarchy.

[21] Agreed synsets between this model prediction and Hollink's are available for reuse here: https://github.com/niamhhenry/Thesis_Basic_Level/blob/main/Experiment%203/BasicLevelSynsets_Agreed.csv

7 CONCLUSION

This thesis investigates the frequency feature of the model developed by Hollink et al. [15] to predict the basic level concepts within WordNet, to gain insights into the corpus properties that enable learning of the basic level. This research illustrates the findings from the original classifier can be reproduced in another programming language, and extended successfully to two new domains through an extended gold standard[22]. Different corpora resources were tested in place of ngrams to allow for representation of characteristics of the basic level theory within the system, such as our tendency to use these concepts most in communication, and particularly when addressing children as these traits can be seen when investigating occurrence of the gold standard in a range of corpora sources.

Experiments using specified corpora outlined in Table 1, highlight the potential value of using them to learn the basic level, as they achieved high and reliable results predicting within and between domains. However, their use in generalising to new data reliably is limited as results indicate larger corpus sizes have more reliable results. Therefore, the performance of the new features do not have sufficient amounts of data to perform as well as ngrams (huge resource), when generalising to new domains. Nevertheless, when comparing smaller samples of the same size, those containing spoken discourse, and discourse directed at children provided more reliable results than written text aimed at a general audience, signalling their value as resources to find, identify and learn the basic level. However, for more reliable results we need larger data samples of these kinds of corpora. If these were available in sizes comparable to ngrams or the BNC, it is hypothesised based on the research in this thesis that they would perform better, and more reliably at predicting the basic level.

8 FUTURE WORK

Although this research provides interesting findings to consider when developing a system to automatically identify the basic level in a hierarchy of concepts there is much more research to be done in this field, as it offers vast benefits for the semantic web. Future research should consider utilising these more specified corpora resources as they remain to be good and reliable sources of the basic level. In particular, the CHILDES corpus, which is also available in a range of languages, may provide useful in investigating the cultural basic level effects across languages. For further research, an extended gold standard beyond the domains Rosch has confirmed to have basic level effects is required, in addition to further investigation in biological domains. A larger label set could be gained through crowd sourcing techniques. Additionally, the confirmed basic level synsets from this research could be linked with ImageNet, to further investigate perceptual qualities of the basic level.

REFERENCES

[1] Marwan Al-Tawil, Vania Dimitrova, and Dhavalkumar Thakker. 2015. Using Basic Level Concepts in a Linked Data Graph to Detect User's Domain Familiarity.. In *UMAP Workshops*.

[2] Marwan Al-Tawil, Vania Dimitrova, Dhavalkumar Thakker, and Brandon Bennett. 2016. Identifying knowledge anchors in a data graph. In *Proceedings of the 27th ACM Conference on Hypertext and Social Media*. 189–194.

[3] Leo Breiman. 2001. Random forests. *Machine learning* 45, 1 (2001), 5–32.

[4] Roger Brown. 1958. Words and things. (1958).

[5] Yi Cai, Wen-Hao Chen, Ho-Fung Leung, Qing Li, Haoran Xie, Raymond YK Lau, Huaqing Min, and Fu Lee Wang. 2016. Context-aware ontologies generation with basic level concepts from collaborative tags. *Neurocomputing* 208 (2016), 25–38.

[6] JE Corter, MA Gluck, and GH Bower. 1988. Basic levels in hierarchically structured categories. In *Proceedings of the 10th Annual Conference of the Cognitive Science Society*. Erlbaum Hillsdale, NJ, 118–124.

[7] Mark. Davies. 2019. The British National Corpus (BNC). https://www.english-corpora.org/bnc/. (Accessed on 03/10/2021).

[8] Sara Feijoo, Carmen Muñoz, Anna Amadó, and Elisabet Serrat. 2017. When meaning is not enough: Distributional and semantic cues to word categorization in child directed speech. *Frontiers in psychology* 8 (2017), 1242.

[9] Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. Bradford Books.

[10] Caroline Graf, Judith Degen, Robert XD Hawkins, and Noah D Goodman. 2016. Animal, dog, or dalmatian? Level of abstraction in nominal referring expressions.. In *CogSci*.

[11] Rebecca Green. 2003. Vocabulary alignment via basic level concepts. *Final Report* (2003).

[12] Rebecca Green, Carol A Bean, and Michèle Hudon. 2002. Universality and basic level concepts. *ADVANCES IN KNOWLEDGE ORGANIZATION* 8 (2002), 311–317.

[13] Lala Hajibayova. 2013. Basic-level categories: A review. *Journal of Information Science* 39, 5 (2013), 676–687.

[14] L Hajibayova and EK Jacob. 2012. A Theoretical Framework for Operationalizing Basic Level Categories in Knowledge Organization Research. In *Categories, Contexts and Relations in Knowledge Organization: Proceedings of the Twelfth International ISKO Conference, Mysore*. 159–165.

[15] Laura Hollink, Aysenur Bilgin, Jacco van Ossenbruggen, and Human Centered Data Analytics. 2020. Predicting the basic level in a hierarchy of concepts. *In proceedings of the Metadata and Semantics Research Conference* (December 2020).

[16] Pierre Jolicoeur, Mark A Gluck, and Stephen M Kosslyn. 1984. Pictures and names: Making the connection. *Cognitive psychology* 16, 2 (1984), 243–275.

[17] Mary E Lassaline, Edward J Wisniewski, and Douglas L Medin. 1992. 9 Basic Levels in Artificial and Natural Categories: Are All Basic Levels Created Equal? In *Advances in psychology*. Vol. 93. Elsevier, 327–378.

[18] Joan Lucariello and Katherine Nelson. 1986. Context effects on lexical specificity in maternal and child discourse. *Journal of Child Language* 13, 3 (1986), 507–522.

[19] Brian Macwhinney. 1992. The CHILDES project: tools for analyzing talk. *Child Language Teaching and Therapy* 8, 2 (1992), 217–218. https://doi.org/10.1177/026565909200800211

[20] Chad Mills. 2018. *Labeling and Automatically Identifying Basic-Level Categories*. Ph.D. Dissertation.

[21] Gregory L Murphy and Hiram H Brownell. 1985. Category differentiation in object recognition: typicality constraints on the basic category advantage. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 11, 1 (1985), 70.

[22] Natural Language ToolKit (NLTK). 2013. Corpus Readers. https://www.nltk.org/howto/corpus.html. (Accessed on 04/24/2021).

[23] Vicente Ordonez, Wei Liu, Jia Deng, Yejin Choi, Alexander C Berg, and Tamara L Berg. 2015. Predicting entry-level categories. *International Journal of Computer Vision* 115, 1 (2015), 29–43.

[24] Abebe Rorissa and Hemalata Iyer. 2008. Theories of cognition and image categorization: What category labels reveal about basic level theory. *Journal of the American Society for Information Science and Technology* 59, 9 (2008), 1383–1392.

[25] Eleanor Rosch, Carolyn B Mervis, Wayne D Gray, David M Johnson, and Penny Boyes-Braem. 1976. Basic objects in natural categories. *Cognitive psychology* 8, 3 (1976), 382–439.

[26] Fei Song and Qingqing Lan. 2019. Corpus-based Research to Verify the Hypothesis of Preference for Basic-level Category Vocabulary (BLCV) Acquisition. *Theory and Practice in Language Studies* 9, 4 (2019), 405–410.

[27] Laura Toloşi and Thomas Lengauer. 2011. Classification with correlated features: unreliability of feature ranking and solutions. *Bioinformatics* 27, 14 (05 2011), 1986–1994. https://doi.org/10.1093/bioinformatics/btr300 arXiv:https://academic.oup.com/bioinformatics/article-pdf/27/14/1986/18530216/btr300.pdf

[28] Omer Topaloglu, Piyush Kumar, and Mayukh Dass. 2020. On the Extendibility of Brands with Subordinate versus Basic Category Concepts. *Journal of Retailing* (2020).

[29] Edward J Wisniewski and Gregory L Murphy. 1989. Superordinate and basic category names in discourse: A textual analysis. *Discourse Processes* 12, 2 (1989), 245–261.

[22] This process resulted in a gold standard of 986 synsets labelled as the basic level, higher, or lower over five domains in WordNet -Gold standard available here: https://github.com/niamhhenry/Thesis_Basic_Level/tree/main/Gold%20Standard%20Labels

APPENDIX

A EXPLORING THE GOLD STANDARD IN CORPORA

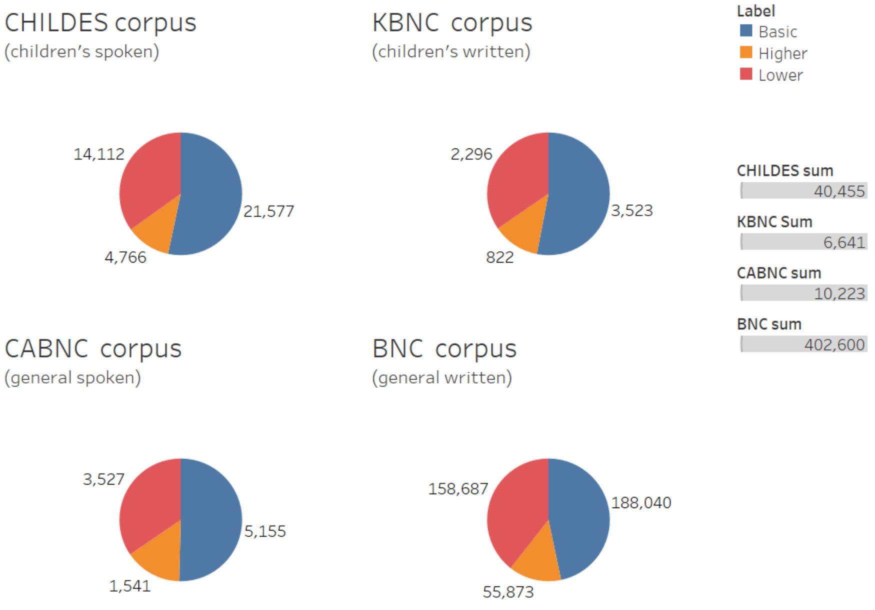


Figure A.1: Pie charts illustrating the levels of the hierarchy for the number of times each lemma per synset in the gold standard appears in each corpus. Higher indicates terms superordinate to the basic level, and lower indicates terms subordinate to the basic level.

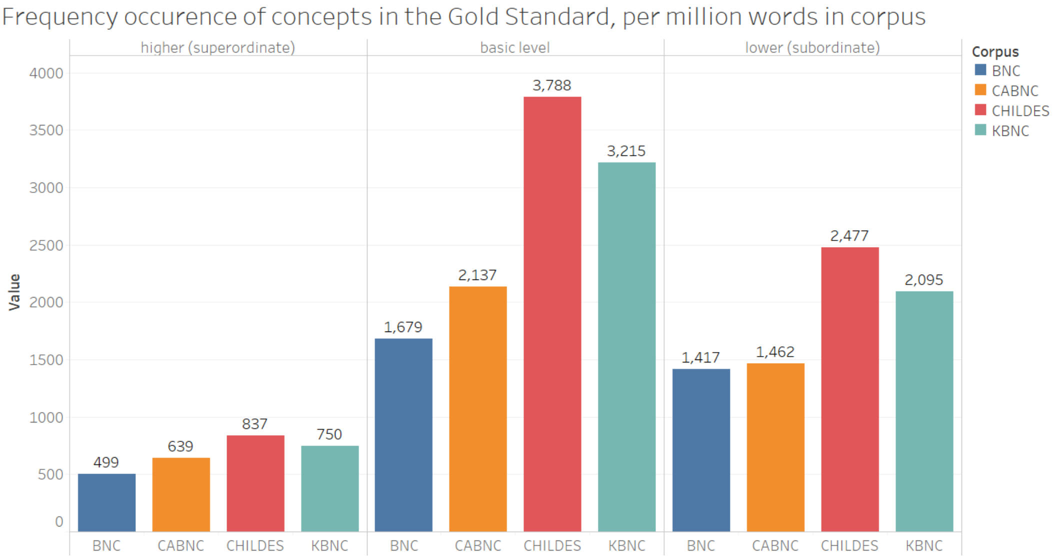


Figure A.2: Bar charts illustrating the levels of the hierarchy for the number of times each lemma per synset in the gold standard appears per 1 million words in each corpus. Higher indicates terms superordinate to the basic level, and lower indicates terms subordinate to the basic level.

B EXPERIMENT 1 RESULTS

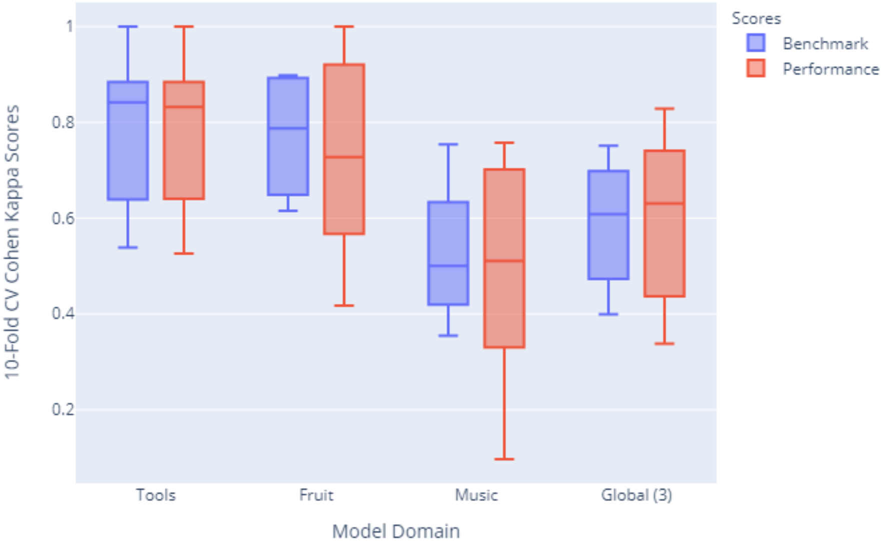


Figure B.1: Box plot of 10 fold cross validated cohen kappa scores for each local model and a global model of all 3 domains using majority vote (performance results in red), compared to the benchmark results from Hollink et al. [15] (in blue).

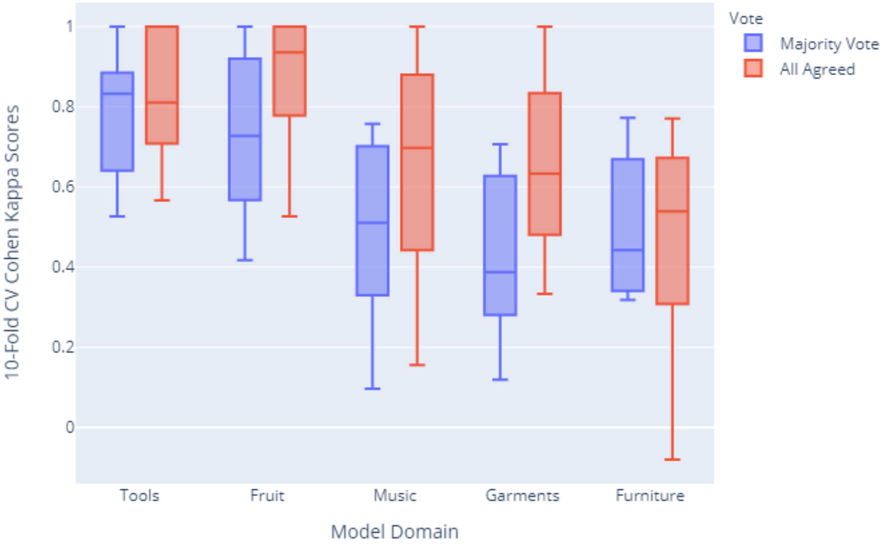


Figure B.2: Box plot of 10 fold cross validated cohen kappa scores for each local model using majority vote and labels where all annotators agree.

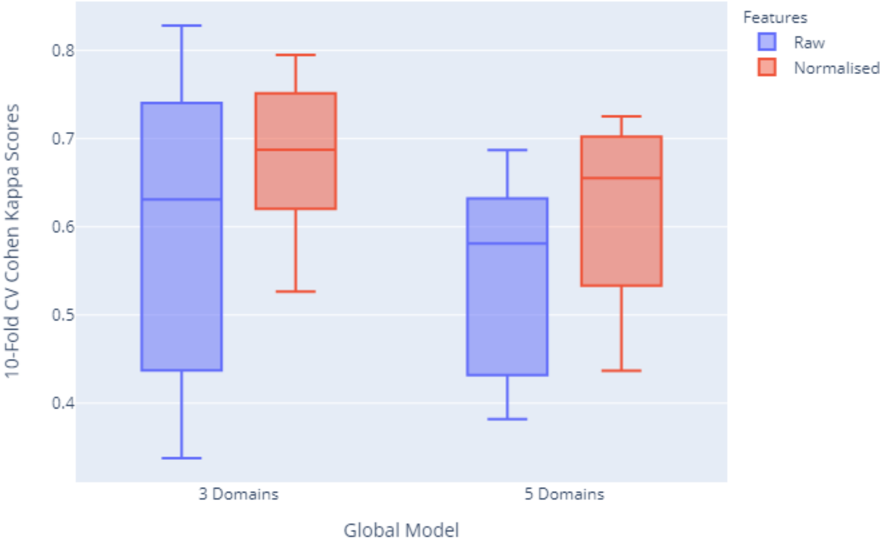


Figure B.3: Box plot of 10 fold cross validated cohen kappa scores for both global models of all 3 domains, and 5 domains using majority vote, comparing when structural features are raw or normalised per domain. [15]

C EXPERIMENT 2 RESULTS

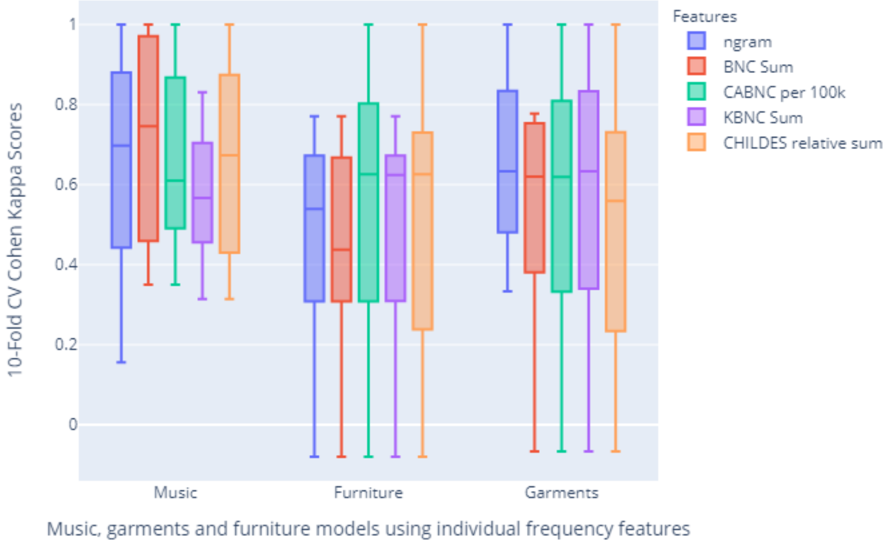


Figure C.1: Box plot of 10 fold cross validated cohen kappa scores for each new frequency input feature used in place of ngrams in the musical instruments, furniture and garments domains.

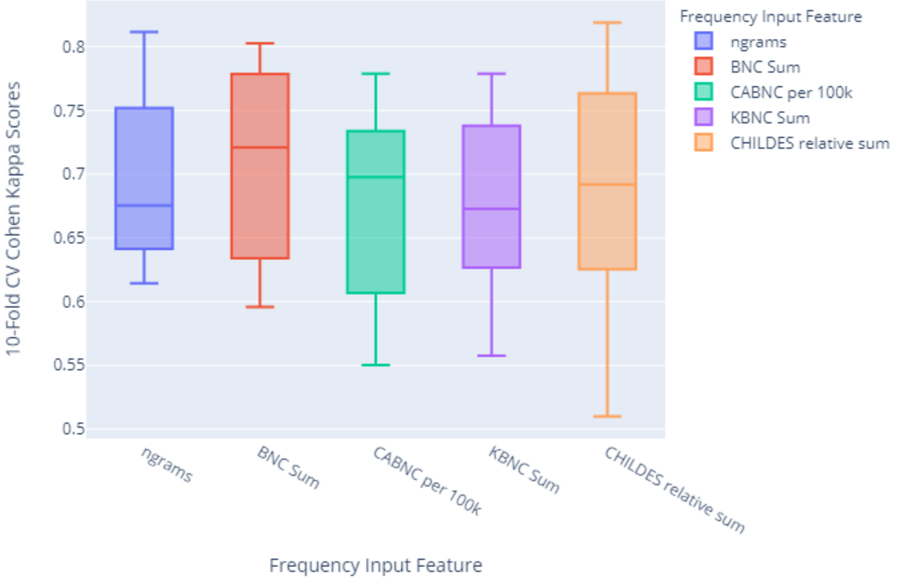


Figure C.2: Box plot of 10 fold cross validated cohen kappa scores for each new frequency input feature used in place of ngrams on the global model, trained and tested on all five domains.

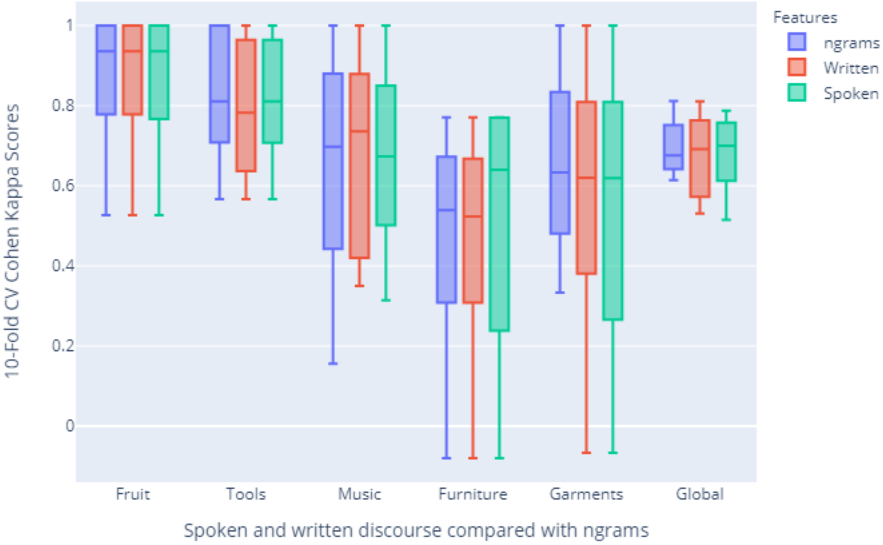


Figure C.3: Box plot of 10 fold cross validated cohen kappa scores for the local and global models of all spoken features compared with all written features with ngrams as the benchmark to compare against.

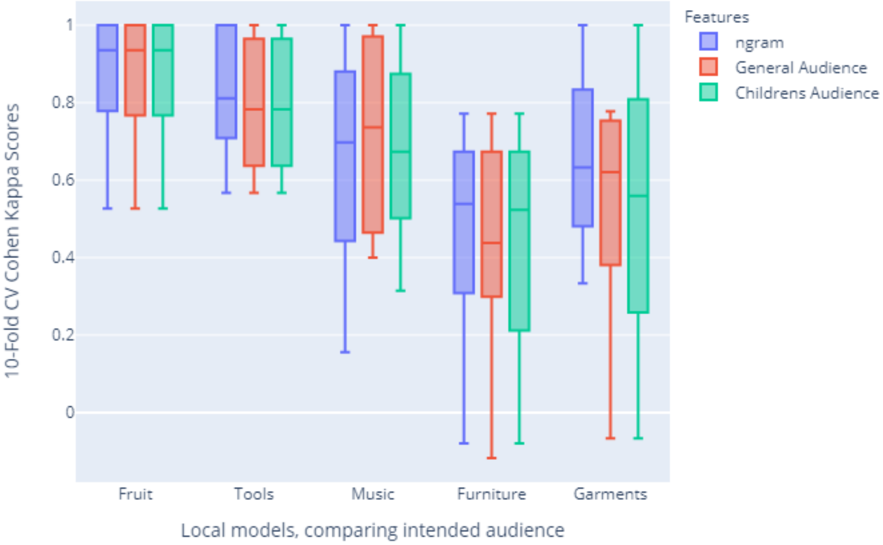


Figure C.4: Box plot of 10 fold cross validated cohen kappa scores for the local models of all corpora features intended for children compared with all features intended for a general audience.

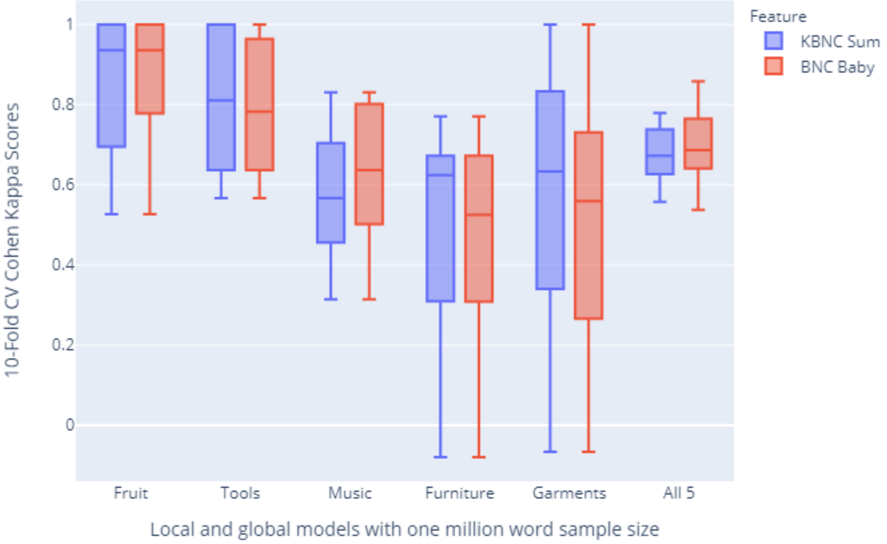


Figure C.5: Box plot of 10 fold cross validated cohen kappa scores for the local and global models of all both features from one million word sample sizes, with the KBNC intended for children compared with the BNC Baby intended for a general audience.

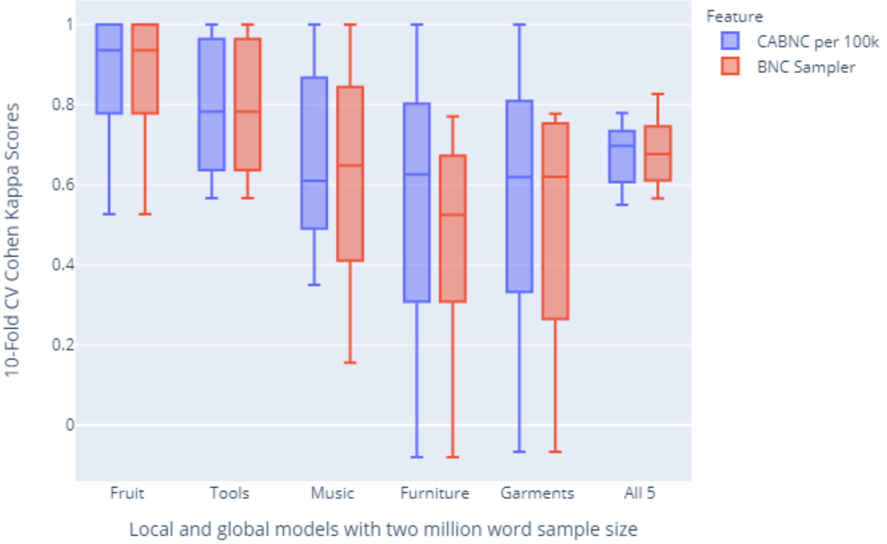


Figure C.6: Box plot of 10 fold cross validated cohen kappa scores for the local and global models of all both features from two million word sample sizes, with the CABNC of spoken discourse, compared with the BNC Sampler of half written, and half spoken discourse.

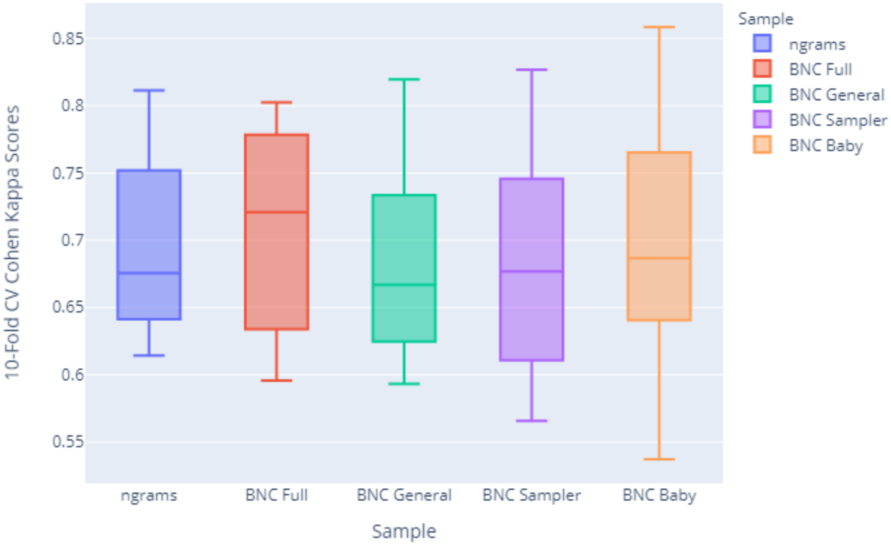
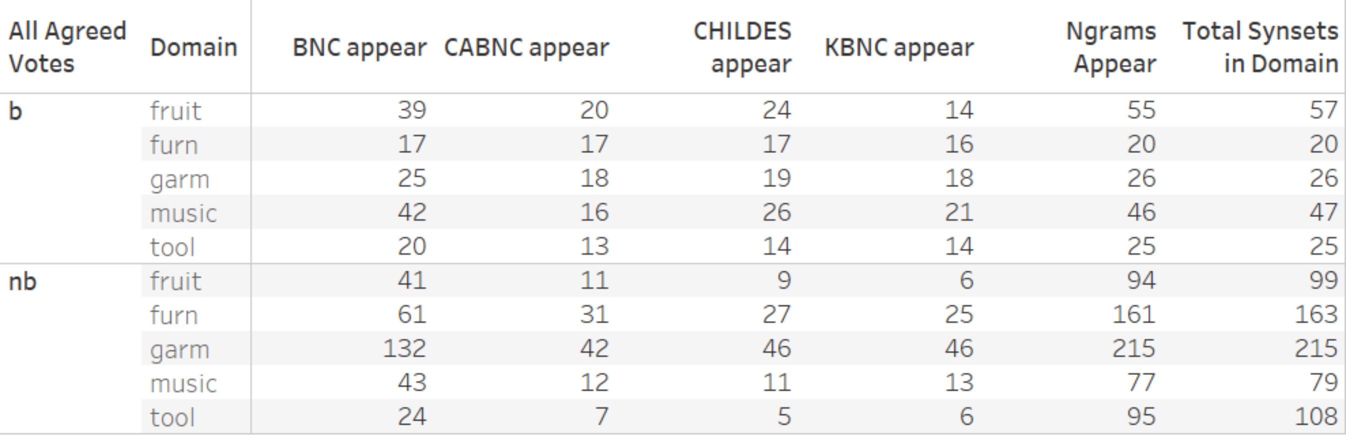


Figure C.7: Box plot of 10 fold cross validated cohen kappa scores for each BNC sample frequency input feature used in place of ngrams on the global model, trained and tested on all five domains.



All Agreed Votes	Domain	BNC appear	CABNC appear	CHILDES appear	KBNC appear	Ngrams Appear	Total Synsets in Domain
b	fruit	39	20	24	14	55	57
	furn	17	17	17	16	20	20
	garm	25	18	19	18	26	26
	music	42	16	26	21	46	47
	tool	20	13	14	14	25	25
nb	fruit	41	11	9	6	94	99
	furn	61	31	27	25	161	163
	garm	132	42	46	46	215	215
	music	43	12	11	13	77	79
	tool	24	7	5	6	95	108

Figure C.8: Table of binary appearances of all agreed gold standard synsets in each corpora, shown per domain (b for basic level, and nb for not basic level).

D EXPERIMENT 2C: UNSEEN DOMAIN RESULTS

Table D.1: Predicted results from models tested on unseen domains (trained on 4, tested on 1) using different frequency features, sorted by best performing feature in each domain, with the top results highlighted in yellow.

Feature	Trained on	Tested on	Cohen Kappa	Accuracy	Balanced Accuracy
BNC Sum	music tool furn garments	fruit	0.88	0.94	0.94
BNC gen	music tool furn garments	fruit	0.87	0.94	0.93
ngrams	tools music furn garments	fruit	0.86	0.85	0.93
CABNC per 100k	music tool furn garments	fruit	0.85	0.93	0.93
BNC Sampler	music tool furn garments	fruit	0.84	0.92	0.92
KBNC Sum	music tool furn garments	fruit	0.82	0.92	0.91
CHILDES rel sum	music tool furn garments	fruit	0.82	0.92	0.92
BNC Baby	music tool furn garments	fruit	0.76	0.88	0.89
ngrams	fruit tools music garments	furn	0.57	0.68	0.78
CABNC per 100k	tool fruit music garments	furn	0.43	0.81	0.84
CHILDES rel sum	tool fruit music garments	furn	0.41	0.81	0.83
BNC Sum	music tool fruit garments	furn	0.39	0.83	0.79
BNC Sampler	music tool fruit garments	furn	0.39	0.81	0.81
KBNC Sum	tool fruit music garments	furn	0.38	0.81	0.8
BNC gen	music tool fruit garments	furn	0.38	0.79	0.84
BNC Baby	music tool fruit garments	furn	0.3	0.78	0.75
ngrams	fruit tools music furn	garments	0.4	0.68	0.74
BNC Sum	music tool fruit furn	garments	0.4	0.83	0.79
BNC gen	music tool fruit furn	garments	0.4	0.83	0.79
BNC Baby	music tool fruit furn	garments	0.34	0.8	0.77
KBNC Sum	tool fruit music furn	garments	0.33	0.79	0.77
CHILDES rel sum	tool fruit music furn	garments	0.32	0.79	0.76
BNC Sampler	music tool fruit furn	garments	0.3	0.79	0.73
CABNC per 100k	tool fruit music furn	garments	0.27	0.79	0.72
ngrams	fruit tools furn garments	music	0.56	0.72	0.77
KBNC Sum	tool fruit furn garments	music	0.54	0.79	0.76
CABNC per 100k	tool fruit furn garments	music	0.54	0.79	0.76
BNC Sampler	tool fruit furn garments	music	0.54	0.79	0.76
BNC gen	tool fruit furn garments	music	0.52	0.79	0.75
BNC Baby	tool fruit furn garments	music	0.52	0.79	0.75
CHILDES rel sum	tool fruit furn garments	music	0.51	0.79	0.74
BNC Sum	tool fruit furn garments	music	0.5	0.78	0.74
CHILDES rel sum	music fruit furn garments	tool	0.82	0.95	0.91
BNC Sum	music fruit furn garments	tool	0.8	0.94	0.89
CABNC per 100k	music fruit furn garments	tool	0.8	0.94	0.89
BNC gen	music fruit furn garments	tool	0.8	0.94	0.89
KBNC Sum	music fruit furn garments	tool	0.77	0.93	0.88
ngrams	fruit music furn garments	tool	0.76	0.86	0.89
BNC Sampler	music fruit furn garments	tool	0.73	0.92	0.87
BNC Baby	music fruit furn garments	tool	0.71	0.91	0.87

E EXPERIMENT 3 RESULTS

Table E.1: All features used in the final model to predict all synsets under the top noun in WordNet, Synset('entity.n.01')

Type	Name	Data
Structural	Depth from top synset	Normalised per domain in WordNet
Structural	Gloss length of synset	Normalised per domain in WordNet
Structural	The number of part-of relations	Normalised per domain in WordNet
Structural	The number of direct hypernyms	Raw number of hypers in WordNet
Structural	The number of hyponyms	Raw number of hypos in WordNet
Lexical	Minimum word length	Smallest word in synset in WordNet (raw characters)
Lexical	Maximum polyscore	The number of synsets in which the most polysemous lemma of the synset appears
Frequency	ngrams mean log	The log of the mean frequency of all lemmas per synset in ngrams

F GOLD STANDARD ANNOTATION PROTOCOL

Task instructions:

For each subset of the hierarchy of synsets (synset.n.nn), provided from domains in WordNet (of garments, and furniture), please consider and note if it is at the basic level (b), above the basic level (h) or below the basic level (l). These levels are further explained and illustrated in the diagram below. This is followed by a range of definitions of the basic level to aid in understanding its characteristics. The last section of this document provides detailed steps to carry out this annotation task.

What is the basic level?

According to a cognitive psychology theory, the basic level theory stipulates that within a hierarchy of concepts, there is one level that we prefer to use when referring to things - the basic level. This level has a range of benefits as it allows us to be as specific and distinctive as we need to be, in as few utterances as possible. It is at this level we recognise things quickest and most accurately. The general shapes of the members of the basic level class usually are similar, thus the basic level is the level at which a concrete image for that category can be formed. We can picture a chair, a table or a lamp quickly and accurately. However, we can not form a single concrete image of furniture, an abstract general term above the basic level. When we picture the subordinate level, it does not provide us with much additional information - only what distinguishes it from other tables.

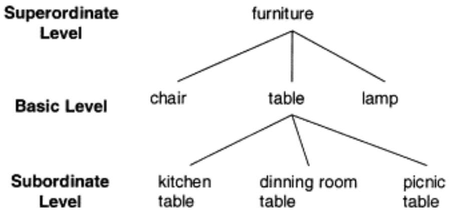


Figure F.1: An example by Zhang, J.[23] of the hierarchy of concepts within the furniture domain distinguished by the basic level and the superordinate higher and subordinate lower levels

Characteristics of the basic level:

Knowing what the basic level is can provide to be advantageous for how we organise information systems, as there are many benefits due to the characteristics of the basic level, further explained below;

- The BL is the level where things are most differentiated by each other by providing enough specificity to be differentiated to other things. Thus, they are distinctive, without providing too much information.
- Superordinate (above the BL) are distinctive, but less informative than the BL
- Subordinate (below the BL) are more informative than the BL, but less distinctive [Murphy and Brownwell, 1985]
- The BL is the level with the most alignable differences (difference is the degree, not the kind) [Markman et al., 1997]. Cars and motorcycles have differences that are alignable (comparable), such as number of wheels. Whereas, different types of cars have more similarities than differences (Sedan vs Audi).
- The BL is often denoted with short, simple words (rule of thumb; not always the case - television is BL)

- A young child is likely to know the basic level term, or parents are likely to teach their children the basic level before other terms [Brown, 1958]
- When faced with an object, or an image of an object, it is the level at which you would commonly refer to it as.
- When things look similar, or have similar sensory motor affordances they are likely to be the basic level.
- Variations of the item may be used in different contexts, but the defining properties are the same across those contexts. For example, ski gloves, motorcycle gloves, horse riding gloves and cleaning gloves are used in different contexts, made of different materials, but they have a similar shape and are put on with a similar movement. Glove is basic level; ski glove is not. Ski-glove lies below the basic level glove, as it is a more specific and defined type of glove [Hollink, 2020].

Finding the BL:

Task: label each synset in the hierarchy of WordNet for the subsets 'garments.n.01' and 'furniture.n.01' as the basic level (b), or higher (h) or lower (l) than the basic level.

- We call the path hyponym relations that connects the top synset to a leaf synset a 'branch.' Basic level synsets in different branches can be placed at different depths in the hierarchy.
- A useful test is to imagine seeing the object from afar - how would you identify it?
- Read the synonyms and glossary (description) of the item whilst considering your knowledge of the basic level, and the descriptions of it outlined above.
- If unsure, feel free to browse the internet to search for definitions, descriptions and examples and images of the items by their name.
- If there are similar shapes and motor actions with each, it may be an indication of the basic level.
- If you are unfamiliar of what the item is, please make a note in the appropriate column - only answer no if you have no idea what the item is.
- If you are unsure of the label you provide it with, and think it should be rechecked or excluded from the study please mark an X in the appropriate column 'unfamiliar and unsure'.
- To further develop our understanding of the basic level and our different interpretations of concepts, please note why the decision was made to label b, l or h. This will only be to provide a discussion for cases whereby annotators disagreed.
- Sometimes there may not be a basic level synset in a branch. It may be omitted from the WordNet hierarchy, or it may be too abstract to picture as a concrete thing such as terms like entitlement, population and magnetisation.
- Please define terms as the basic level, or higher or lower. However, if you believe it is a possible synset with no basic level, then leave an X in the column 'possible none' and note why you think this. This will only be used for discussion, and not as part of the study